Cerebral Cortex November 2014;24:2859–2872
doi:10.1093/cercor/bht138
Advance Access publication May 24, 2013

Aversive Learning Modulates Cortical Representations of Object Categories

Joseph E. Dunsmoor[1], Philip A. Kragel[1], Alex Martin[2] and Kevin S. LaBar[1]*

[1]Department of Psychology and Neuroscience, Center for Cognitive Neuroscience, Duke University, Durham, NC 27708, USA
[2]Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA

*Address correspondence to Kevin S. LaBar. Email: klabar@duke.edu

Experimental studies of conditioned learning reveal activity changes in the amygdala and unimodal sensory cortex underlying fear acquisition to simple stimuli. However, real-world fears typically involve complex stimuli represented at the category level. A consequence of category-level representations of threat is that aversive experiences with particular category members may lead one to infer that related exemplars likewise pose a threat, despite variations in physical form. Here, we examined the effect of category-level representations of threat on human brain activation using 2 superordinate categories (animals and tools) as conditioned stimuli. Hemodynamic activity in the amygdala and category-selective cortex was modulated by the reinforcement contingency, leading to widespread fear of different exemplars from the reinforced category. Multivariate representational similarity analyses revealed that activity patterns in the amygdala and object-selective cortex were more similar among exemplars from the threat versus safe category. Learning to fear animate objects was additionally characterized by enhanced functional coupling between the amygdala and fusiform gyrus. Finally, hippocampal activity co-varied with object typicality and amygdala activation early during training. These findings provide novel evidence that aversive learning can modulate category-level representations of object concepts, thereby enabling individuals to express fear to a range of related stimuli.

Keywords: anxiety, categories and concepts, fear conditioning, functional magnetic resonance imaging, generalization

Introduction

One way emotional experiences shape our lives is by changing how we represent and respond to objects we already know. For instance, an airline passenger surviving a harrowing flight may develop a widespread fear and avoidance of all aircraft. In this way, an emotional experience with an instance of a particular object category leads to generalizations about the properties of related objects through the induction of knowledge (Murphy 2002). If an aversive experience with a category exemplar draws on existing conceptual knowledge to modulate representations of related object concepts, this experience-dependent change in representational architecture could provide a potential mechanism for category-based fear generalization. Such a mechanism could contribute to an understanding for anxiety disorders like Specific Phobias, for which fears are typically expressed categorically.

Pavlovian fear conditioning provides an experimental procedure particularly suited for investigating how meaningful experiences with particular objects affect behavior to related objects. Fear-conditioning studies in rodents provide evidence for neural mechanisms involved in experience-dependent modulations in stimulus representations. Specifically, sensory information concerning the conditioned stimulus (CS) and aversive unconditioned stimulus (US) converges in the basolateral amygdala, leading to strengthened CS–US associations, production of conditioned fear responses (CR) (Maren 2001; Pape and Paré 2010), and plasticity in unimodal sensory cortices (LeDoux 2000; Weinberger 2004; Pape and Paré 2010; Letzkus et al. 2011). For instance, auditory fear conditioning leads to learning-induced retuning of neurons in auditory cortex toward the CS frequency, increased areas of representation for the CS frequency in auditory cortex, and behavioral fear generalization to similar tones (Weinberger 2004, 2007). Functional magnetic resonance imaging (fMRI) studies in humans confirm activity changes in the amygdala and sensory areas during conditioning with odors (Li et al. 2008), tones (Knight et al. 2005), and visual images (Lim et al. 2008). Furthermore, perceptually based fear generalization in humans occurs through changes in functional connectivity between the amygdala and visual cortical areas (Dunsmoor et al. 2011). Taken together, these investigations suggest that fear conditioning modulates activity in and connectivity between the amygdala and cortical areas coding for the sensory qualities of the CS.

Although prior evidence provides a mechanism to explain how perceptually based information in sensory cortices is modulated through fear conditioning, whether aversive experiences modulate neural representations and organization of conceptually based information is unknown. Candidate neural regions include those representing information about object concepts, such as the semantic categories of animals and tools (Tranel et al. 1997; Patterson et al. 2007). Neuroimaging investigations reveal category selectivity for these superordinate category domains along posterior occipitotemporal cortex (Chao and Martin 2000; Ishai et al. 2000; Beauchamp et al. 2002; Simmons and Martin 2011). Whereas images of animals reliably activate lateral fusiform gyrus (FFG), posterior superior temporal sulcus (pSTS), and inferior occipital regions, images of tools activate medial FFG, posterior middle temporal gyrus, and middle occipital cortex (Martin 2007b). The organization of category-selective regions may be related to intrinsic connectivity between unique brain networks (Mahon et al. 2007; Mahon and Caramazza 2011). For instance, amygdala-lateral FFG connectivity has been proposed as a key network involved in the representation of animate objects (Martin 2007a), whereas connectivity between motor cortex and medial FFG is important for representing graspable objects like tools (Mahon and Caramazza 2011). In support for this hypothesis, separated patterns of functional connectivity during rest have been identified between regions implicated in social- versus tool-related conceptual tasks (Simmons and Martin 2011). Finally, multivariate approaches including representational similarity analyses (RSA) provide an emerging tool for probing architecture of category representations, such as those distinguishing animate and inanimate objects in temporal cortex (Kriegeskorte, Mur, and Bandettini 2008; Kriegeskorte, Mur, Ruff, et al. 2008). Thus, well-established separations in the

© The Author 2013. Published by Oxford University Press. All rights reserved.
For Permissions, please e-mail: journals.permissions@oup.com

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

neural organization of category knowledge provide a framework to investigate whether instances of fear learning to category-specific exemplars modify categorical representations in occipitotemporal cortex and associated networks, thereby enabling individuals to express fear to a range of conceptually related objects.

The present study developed a novel trial-unique conditioning paradigm using event-related fMRI to investigate whether aversive learning modulates representations of object concepts in the human brain. Taking advantage of well-established organizational separations in representations for animate and inanimate objects, we utilized animals and tools as conditioned stimuli. One group of participants learned through experience that animals predicted an aversive electrical shock and images of tools were safe, while a separate group received the opposite contingencies. We predicted that aversive learning acquired at the category level would modify representations in category-selective areas along the occipitotemporal cortex, as well areas implicated in simpler forms of conditioning, such as the amygdala. Moreover, we predicted that multivariate patterns of activity in object-selective cortex would reveal enhanced representational similarity among different category members from the feared category, which may be integral to facilitate the transfer of affective learning between physically distinct objects. Consistent with the notion that the representations of animate object concepts is supported by a domain-specific functional network that encompasses the amygdala and lateral FFG (Martin 2007a; Mahon and Caramazza 2011), we also predicted dissociation in patterns of functional connectivity between these 2 regions across learning groups. Specifically, learning to fear a category of animate (as opposed to inanimate) objects was expected to selectively utilize amygdala-lateral FFG connectivity during learning, and may be further reflected in resting state connectivity after learning (Simmons and Martin 2011).

Finally, we sought to determine whether object typicality plays a role when aversive learning occurs at the category level. Object categories contain graded structures that can be organized around members that are regarded as more typical than other members (Rosch and Mervis 1975). Typicality effects play a well-known role in behavioral category-based induction (Murphy 2002). For instance, a categorical argument containing a premise about typical members (e.g., a premise involving robins vs. a premise involving penguins) is considered stronger and is more readily generalized to other category members (e.g., is true of other birds) (Rips 1975; Osherson et al. 1990). Although brain imaging investigations of typicality effects are scant, a candidate region involved in extracting conceptual information from a learning experience is the hippocampus. Neurons in the medial temporal lobe, for instance, evince selective, sparse, and invariant responses to specific object concepts (Quiroga et al. 2005). Quiroga (2012) has proposed that the explicit high-level representation of object concepts in the medial temporal lobe facilitates the creation of new associations and episodic memories. The hippocampus is also implicated broadly in the generalization of learning (Gluck and Myers 1993; Shohamy and Wagner 2008), and may thus be uniquely situated to generalize learning based on abstract properties (i.e., typicality) of an object concept. While it is unknown whether typicality effects apply to aversive learning, we predicted that the hippocampus would signal object typicality early in training, when representativeness may factor into category-level generalizations of threat.

Materials and Methods

Subjects

Thirty-four right-handed healthy adults provided written informed consent in accordance with the Duke University Institutional Review Board guidelines. Data from one subject were excluded due to technical problems with the MRI scanner. The final analysis included 33 participants (16 females; age range = 18–37 years; median age = 23 years). Subjects were assigned either to the group in which animals predicted shock and tools were safe (A+/T−, $n = 17$) or to the group in which tools predicted shock and animals were safe (T+/A−, $n = 16$).

Stimuli

Stimuli consisted of 80 images of tools ($N = 40$) and animals ($N = 40$) presented on a white background. Images were obtained from the website www.lifeonwhite.com and from publicly available resources on the internet. These stimuli were used in a previously published behavioral study on fear generalization (Dunsmoor et al. 2012) (see Supplementary Fig. 1 for complete list of stimuli and typicality ratings). We avoided the use of highly threat-relevant images such as knives and snakes so as to mitigate the potential arousal bias evoked by these objects (Öhman and Mineka 2001). The presentation of stimuli was controlled using Presentation Software (Neurobehavioral Systems, Albany, CA, USA).

A 6-ms electrical shock applied to the right wrist served as the US. Shock administration was controlled using the STM-100 and STM-200 modules connected to the MP-150 BIOPAC system (BIOPAC systems, Goleta, CA, USA). The level of the shock was calibrated for each subject individually prior to the start of the experiment using an ascending staircase procedure so as to reach a level deemed "annoying but not painful" (Dunsmoor et al. 2009). Following calibration, subjects rated intensity of the shock on a scale from 0 (not at all unpleasant) to 10 (extremely unpleasant). Shock ratings (mean = 5.56; SD = 1.04) were similar across groups.

Task and Procedures

The scanning session contained 4 phases: animal-tool functional localizer, preconditioning resting state, fear conditioning, and postconditioning resting state. The animal-tool localizer consisted of 12 alternating blocks of animals, tools, and phase-scrambled images counterbalanced across subjects (Supplementary Methods). The fear conditioning session was a slow event-related fMRI design conducted over 4 runs of equal length (6 min). Each run contained 20 trials (10 animals and 10 tools) presented in a pseudorandomized order with no more than 2 objects from the same category occurring in a row. We used 8 different stimulus presentation orders to counterbalance presentation of animal and tool exemplars across participants. Each trial was 6 s in duration, during which time subjects rated expectancy for US delivery on a scale anchored between 0 (sure the shock will not occur) and 10 (sure the shock will occur) using an MRI-compatible joystick. Subjects were instructed that the final location of the rating bar at trial offset would be calculated as their response. A 10–12 s (average = 11 s) fixation cross-followed the offset of each trial. One category (e.g., animals) was designated the CS+, and 50% of exemplars from this category co-terminated with the US. The other object category (e.g., tools) served as the CS−, and none of its exemplars were reinforced with a US. Category assignment was counterbalanced (animals CS+: $n = 17$; tools CS+: $n = 16$). The choice of which CS+ items were paired with shock was random and counterbalanced between subjects. Every trial contained a distinct basic-level exemplar with a different name that was never repeated (e.g., there were not 2 different images of elephants). Subjects were not instructed on the CS–US contingencies and were not informed that images would be presented only once during learning. Twenty-four hours later, subjects returned to a different laboratory setting outside the MRI facility where they

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

conducted a follow-up recognition memory test for the CS+ and CS− items (not reported here) and rated each animal and tool picture for typicality to the superordinate category. Typicality ratings were made for each item on a 10-point scale (0 = highly atypical, 10 = highly typical).

SCR Methods and Analysis

Psychophysiological recording was controlled with the MP-150 BIOPAC system (BIOPAC systems) grounded through the resonance frequency filter panel and shielded from magnetic interference. MRI-compatible skin conductance responses (SCR) electrodes were placed on the hypothenar eminence of the palmar surface of the left hand. SCR analysis was carried out using AcqKnowledge software (BIOPAC systems) using procedures previously described (Dunsmoor et al. 2012) (see Supplementary Methods for details).

fMRI Parameters

Whole-brain functional imaging was conducted on a General Electric Signa EXCITE HD 3.0 Tesla MRI scanner. Blood oxygenation level-dependent functional images were acquired parallel to the AC-PC line using a SENSE™ spiral in sequence: acquisition matrix, 64 × 64; field of view, 256 × 256; flip-angle, 60°; 34 slices with interleaved acquisition; slice thickness, 3.8 mm with no gaps between slices; in-plane resolution = 3.75 × 3.75 mm; repetition time, 2 s; echo time, 27 ms.

fMRI Preprocessing

Preprocessing was conducted using SPM8 (Wellcome Trust Center, www.fil.ion.ucl.ac.uk) implemented in MATLAB (The Mathworks, Inc., Natick, MA, USA). Images were spatially normalized into Montreal Neurological Institute (MNI) space, voxel size resampled to 2 × 2 × 2 mm, and smoothed using an isotropic 8-mm^3 Gaussian full-width half-maximum kernel. Functional images were co-registered to each subject's high-resolution T_1-weighted structural scan. The first 4 functional images were removed from each scanning run to account for magnetic equilibration, and remaining images were corrected for head motion using center-of-mass movement threshold of 3 mm. Padding was inserted around the subject's head to minimize head motion.

fMRI Analysis

Statistical analysis of preprocessed data was conducted using the general linear model in SPM8 with trial regressors for onset and duration of each event convolved with the canonical hemodynamic response function. Head motion parameters and shock delivery were included as covariates of no interest. A high-pass filter of 128 s was applied to account for low-frequency drifts. Statistical thresholds for whole-brain analyses were set at $P < 0.001$, with a minimum extent threshold of 60 voxels. This spatial extent for multiple comparisons provides cluster correction of $P < 0.05$, derived using the REST AlphaSim utility (www.restfmri.net; toolkit V 1.3), which computes alpha level using 1000 Monte Carlo simulations to verify activations of this cluster size were unlikely to have occurred due to chance. Based on extensive literature on the role of the amygdala in fear conditioning (LaBar et al. 1998) and the hippocampus on learning and generalization (Shohamy and Wagner 2008), we included these as a priori ROIs for small volume correction analysis. Thresholds for small volume correction were conducted using bilateral anatomical masks from the PickAtlas toolbox (Maldjian et al. 2003) and a threshold of familywise error (FWE) corrected $P < 0.05$ with 10 contiguous voxels. Brain activations are displayed on the average anatomical image from 33 subjects. Labels for the anatomical regions provided by Talairach Client (Lancaster et al. 2000) based on peak coordinates in MNI space, converted to Talairach space using GingerALE 2.0 (Laird et al. 2010).

Aversive Learning

To examine effects of aversive learning in category-selective regions, mean beta-parameters were extracted from the 6 ROIs identified from the independent object localizer. Statistical tests conducted on extracted parameter estimates included a repeated-measures analysis of variance (ANOVA) with CS type (CS+, CS−) and group (animals CS+/tool CS−, tools CS+/animals CS−) as factors, considered significant at $P < 0.05$. Whole-brain group-level fMRI analysis of aversive learning was also analyzed using a 2 × 2 ANOVA with the factors CS type (CS+, CS−) and group (animals CS+/tools CS−, tools CS+/animals CS+). Areas exhibiting enhanced activations on CS+ versus CS− trials were assessed through a conjunction analysis for the contrast CS+ > CS− for both groups; a conjunction analysis for the reverse contrast (CS− > CS+) was conducted to reveal regions exhibiting enhanced activity on control versus threat trials in both groups.

Representational Similarity Analysis

The RSA followed general methods outlined by Kriegeskorte, Mur, and Bandettini (2008) and Kriegeskorte, Mur, Ruff, et al. (2008). Further details on this analysis are provided in the Supplementary Methods. We constructed a general linear model with each regressor containing a single trial. Preprocessing did not include spatial smoothing to retain information at a finer spatial scale. Activity was sampled separately from voxels within the object localizer mask (animals + tools > scrambled objects) and bilateral anatomical amygdala ROIs from the AAL atlas. To rule out potential artifactual differences in correlation estimates in the RSA across CS+ and CS conditions, voxels identified from the univariate analysis as exhibiting greater signal differences to the CS+ versus the CS− in either group were excluded (see also Larocque et al. 2013). This included voxels from the object-selective visual cortex identified at the whole-brain level using the global null hypothesis with an uncorrected threshold of $P < 0.001$, and from the amygdala RSA using the global null hypothesis and a small-volume correction (FWE $P < 0.05$) on the amygdala anatomical mask. Results yielded an 80 object by 3563 voxel (object localizer) or 396 voxel (amygdala) activity patterns for each subject. Representational dissimilarity matrices (RDM) were computed from these patterns by calculating 1-r (Pearson correlation coefficient) across all object selective voxels for every possible pairing of objects, producing a symmetric 80 × 80 RDM. To overcome distributional and independence assumptions of standard statistical approaches, bootstrap resampling (using the bootstrp function in MATLAB with 10 000 iterations) was used to estimate mean similarity patterns within and between categories for each subject to examine whether within-category dissimilarity for CS+ items was different from within-category dissimilarity for CS− items and between-category dissimilarities. Mean dissimilarities were assessed by repeated measures ANOVA with RDM quadrants as within-subjects factors [lower triangle of the first quadrant (animals-to-animals); lower triangle of the fourth quadrant (tools-to-tools); and third quadrant (animals-to-tools)] and group as a between-subjects factor. Additional analyses were conducted on the derived similarity measures to examine whether factors other than CS type (e.g., trial-by-trial univariate activity) influenced pattern similarity (see Supplementary Fig. 7 and Supplemental Results). Additional RDMs were constructed for 2 control regions, primary visual cortex (Brodmann area 17) and left motor cortex (see Supplemental Results).

Psychophysiological Interaction Analysis

We conducted a psychophysiological interaction (PPI) analysis (Friston et al. 1997) implemented in SPM 8 to examine patterns of functional connectivity during the experimental task. The purpose of the PPI analysis was to demonstrate functional coupling across the brain with a source region and an experimental parameter. We used this analysis to test the hypothesis that task-based connectivity (CS+ vs. CS−) between the lateral FFG and amygdala is selectively enhanced in the group fear conditioned to animals. The source region was provided by the amygdala ROI identified from the main effect of fear conditioning (CS+ > CS−). The representative time course was extracted from voxels in the amygdala mask—which included voxels in both left and right amygdala—using the first eigenvariate calculated from singular value decomposition. The time course from the source region, the psychological context (CS+ > CS−), and the interaction term between the time series and psychological context (PPI) were included in a general linear model (GLM). The PPI analysis reveals brain regions exhibiting stronger functional coupling with the amygdala during aversive

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

learning, while accounting for the main effect of conditions by including the psychological and physiological time course into the GLM. A supplementary whole-brain analysis (cluster correction $P < 0.05$) was conducted to examine amygdala connectivity across other brain regions in both learning groups.

Resting State Connectivity Analyses

Immediately before and immediately after fear conditioning, subjects rested with their eyes open during a 6-min resting state scan. One subject from the A+/T− group did not complete the postconditioning resting state due to time constraints and is thus not included in the resting state analysis. We used a targeted approach to examine correlations between the lateral FFG and amygdala before and after aversive learning. Time courses were extracted from the amygdala (identified from the fear-conditioning analysis contrast CS+ > CS−) and category-selective lateral FFG that dissociated animals from tools during the independent localizer using methods described above for the PPI analysis. Time courses were fit using a GLM incorporating head motion and the mean signal over the course of the run. A 128-s filter was applied to remove low-frequency drift. Pearson correlation coefficients were calculated for each subject during baseline and post conditioning rest for amygdala-category ROI pairs. Correlation coefficients were Fisher *Z*-transformed and baseline values were subtracted from the postconditioning values to gain a measure of the change in correlations between regions from before to after fear conditioning. One-sample *t*-tests were used to determine significant differences in the pre-to-post change in connectivity.

Parametric Modulation Analysis of Object Typicality

To examine whether any brain regions tracked object typicality during fear conditioning, we used typicality ratings (obtained for each subject 24 h after the imaging session) for each CS+ image as a trial-by-trial parametric regressor. Typicality ratings were Z-scored and used in the parametric modulation analysis. Three subjects were not included in this analysis due to a lack of variability in typicality ratings (SD of raw typicality scores <2). First-level model estimation was conducted using the GLM, with the following events: CS+ and its parametric modulation, CS− and its parametric modulation, US, and head motion parameters. Group-level analysis included a factorial design with the parametric modulation of the CS+ across the 4 scanning runs as factors, assuming unequal variances and nonindependence of samples. The analysis focused on regions showing a linear decrease in CS+ modulation over the 4 runs of fear conditioning [contrast weights; 3 1 −1 −3] to broadly characterize changes across the conditioning session.

Lastly, we interrogated the role of the hippocampus further by incorporating the hippocampal region identified from the parametric modulation analysis (see Results section) as a seed region in a PPI analysis using PPI procedures described above. The purpose of this analysis was to examine regions exhibiting task-based connectivity with the hippocampus during early training (when the hippocampus exhibited typicality effects) but not late training (when the typicality effect subsided). The interaction term between the hippocampus time series and experimental parameter (CS+ vs. CS−) was entered into a group-level model that included run (1–4) as a factor. We examined whether any regions exhibited a linear decrease in connectivity across the conditioning session.

Results

Participants learned that different objects from a category domain were predictive of shock while a set of objects from another category domain were used as an unpaired control condition (see Fig. 1). For half of the participants, a subset of images of animals was paired with the delivery of an electrical shock US whereas all tools were presented without the US (i.e., animal CS+, tool CS−). The other half of participants received the reverse contingencies (i.e., tool CS+, animal CS−). Thus, each participant viewed the same images, but individualized learning histories determined which category attained threat-relevance.

Another important feature of this task is that individual category exemplars were only presented once. This approach differs markedly from the classic literature on conditioning and stimulus generalization, which traditionally present a single repeated CS prior to testing with unpaired values that parametrically vary from the CS along a basic sensory dimension (for review of classic stimulus generalization experiments, see Honig and Urcuioli 1981). In this case, fear acquisition requires subjects to successfully generalize learning beyond each specific trial-unique instance, rather than through repeated exposure to the same CS. Thus, subjects were expected to acquire a representation of threat at the superordinate category level based on conceptual knowledge of the relationship between basic-level members.

Behavioral Results

Aversive learning was evaluated throughout the scanning session using online subjective ratings of shock expectancy and SCR. Shock expectancy (Fig. 1*b* and Supplementary Figs 2 and 3) was greater to the CS+ than the CS−, $F_{1,\ 31} = 334.54$, $P < 0.001$. There was no effect of group ($P = 0.24$) and no interaction between CS type and group, $P = 0.39$. Analysis of SCRs (Fig. 1*c*) likewise revealed a main effect of CS type, $F_{1,\ 28} = 33.66$, $P < 0.001$, but no effect of group ($P = 0.56$) and no interaction with group, $P = 0.89$. No correlation was found between SCRs and shock expectancy, or between typicality ratings and other behavioral indices. Finally, tonic skin conductance levels (SCL) were acquired during 6-min resting state scans immediately before and after fear conditioning. SCLs significantly increased from before to after conditioning ($P < 0.001$) with no difference between groups ($P > 0.5$), providing further evidence that the aversive learning episode had a lasting impact on emotional arousal. Together, these results demonstrate that subjective and autonomic contingency learning differentiated as a function of CS type, but were unaffected by which category (animals or tools) predicted shock, replicating our behavioral findings using this paradigm (Dunsmoor et al. 2012).

Animal-Tool Functional Localizer

Prior to fear conditioning, object-selective regions along occipitotemporal cortex were localized in a separate task. Consistent with prior fMRI studies examining category-specificity for animate versus manmade objects (Martin 2007b), whole-brain analysis revealed clusters selective to animals in the lateral FFG and inferior occipital gyrus extending into the pSTS, and clusters selective to tools in medial FFG/parahippocampal gyrus and middle occipital gyrus (Fig. 2*a* and Supplementary Table 1). Based on these localized activations from a priori occipital-temporal regions, we created regions of interest (ROI) for use in further analyses (see Experimental Procedures and Supplementary Materials and Methods).

Aversive Learning Modulates Activity in Posterior Category-Selective Cortical Regions

A primary question in this investigation was whether aversive learning enhances activity in category-selective cortex as a function of whether animals or tools attained threat value. To investigate this question, we probed neural activity during

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

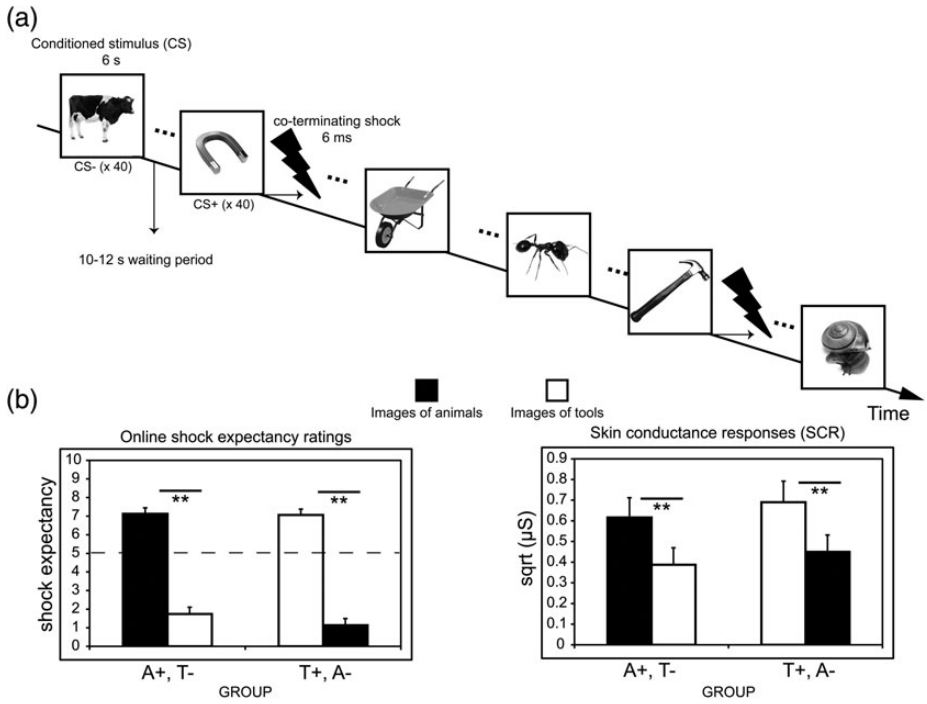

Figure 1. Fear conditioning procedure and behavioral results. (*a*) Subjects were presented with 80 unique exemplars of animals (40) and tools (40) for 6 s each. Exemplars were never repeated during the conditioning session. Half the objects from one category (CS+) co-terminated with a 6-ms electrical shock unconditioned stimulus (US), whereas objects from the other category were never reinforced (CS−). In this example, tools served as the CS+ and animals served as the CS− (category assignment was counterbalanced across subjects). Subjects learned through experience that an object category predicted shock. (*b*) Expectancy ratings showed that subjects learned the CS-US contingencies. Skin conductance responses (SCRs) revealed differentially greater responses to the category exemplars predictive of shock versus the safe category exemplars. Neither shock expectancy ratings nor SCRs were influenced by which object category (animals or tools) predicted shock. Dashed line depicts chance level of US expectancy. A+, images of animals predicted shock; T+, images of tools predicted shock; A−, images of animals were safe; T−, images of tools were safe; μS, microSiemens; Error bars represent ± SEM; $^{**}P < 0.01$, 2-tailed *t*-tests.

learning using the functional ROIs defined by the independent object localizer. As expected, analysis of mean parameter estimates extracted from these ROIs revealed interactions between CS type and group, with the sole exception of the tool-selective right middle occipital gyrus (Fig. 2*b*,*c* and Table 1). Post hoc *t*-tests indicated that these interactions were driven by greater activity to images from the CS+ (vs. the CS−) category within the ROI preferential to the CS+ category. In the tool-selective ROIs, (Fig. 2*c*), greater activity to the CS+ versus the CS− was only observed in the group fear conditioned to tools. Within animal-selective ROIs, greater activity to CS+ than CS− was observed in the group fear conditioned to animals. Moreover, activity to animal CS+'s in this group was greater than activity evoked by both animal CS−'s and tool CS+'s in the opposite group (Fig. 2*b*). Altogether, these findings provide evidence that fear learning enhanced activity within category-selective cortex as a function of the reinforcement contingency. These results held when the analysis was restricted to CS+ unpaired trials only, despite the loss of power in this analysis (Supplementary Fig. 4).

Effects of Aversive Learning on Representational Similarity in Object-Selective Cortex

In addition to the univariate fMRI analysis reviewed above, we examined activity in object selective cortex using a multivariate RSA (Kriegeskorte, Mur, and Bandettini 2008; Kriegeskorte, Mur, Ruff, et al. 2008). In this approach, the response patterns in a set of voxels are extracted and their similarity structure is compared across stimulus categories or experimental manipulations (see Supplementary Materials and Methods for further details on this analysis). For this analysis, we examined patterns of neural activity within bilateral occipitotemporal cortex identified as signaling intact objects during the object localizer task. Notably, we excluded voxels exhibiting greater mean activation to the CS+ (vs. the CS−) in either learning group to mitigate the possibility that increased signal-to-noise artifactually induced differences in correlation estimates due to attentional gain modulation for the CS+ category (see Supplementary Materials and Methods). This analysis showed a delineation between animate and inanimate objects in the occipitotemporal cortex (Fig. 3*a*). By quantifying similarity structures within each quadrant of the RDM, we elucidated alterations in representational structures as a function of which object category was reinforced. We used a bootstrap resampling approach to quantify mean dissimilarity of multivariate patterns for the lower triangle of the first quadrant (animals-to-animals), lower triangle of the fourth quadrant (tools-to-tools), and the third quadrant (animals-to-tools) (Fig 3*b*) (see Supplementary Materials and Methods) for each subject. For subjects in the animal CS+ group, representational structures for animals-to-animals were more similar than for tools-to-tools (CS−) or animals-to-tools. In contrast, the tool CS+ group exhibited more similarity among tool-to-tool exemplars (CS+) than animal-to-animal (CS−) and animal-to-tool exemplars (Ps < 0.01). A similar increase in within-category similarity for CS+ versus CS− items was observed in the amygdala (Supplementary Fig. 5), whereas category structures in control regions (V1 and left motor cortex)

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

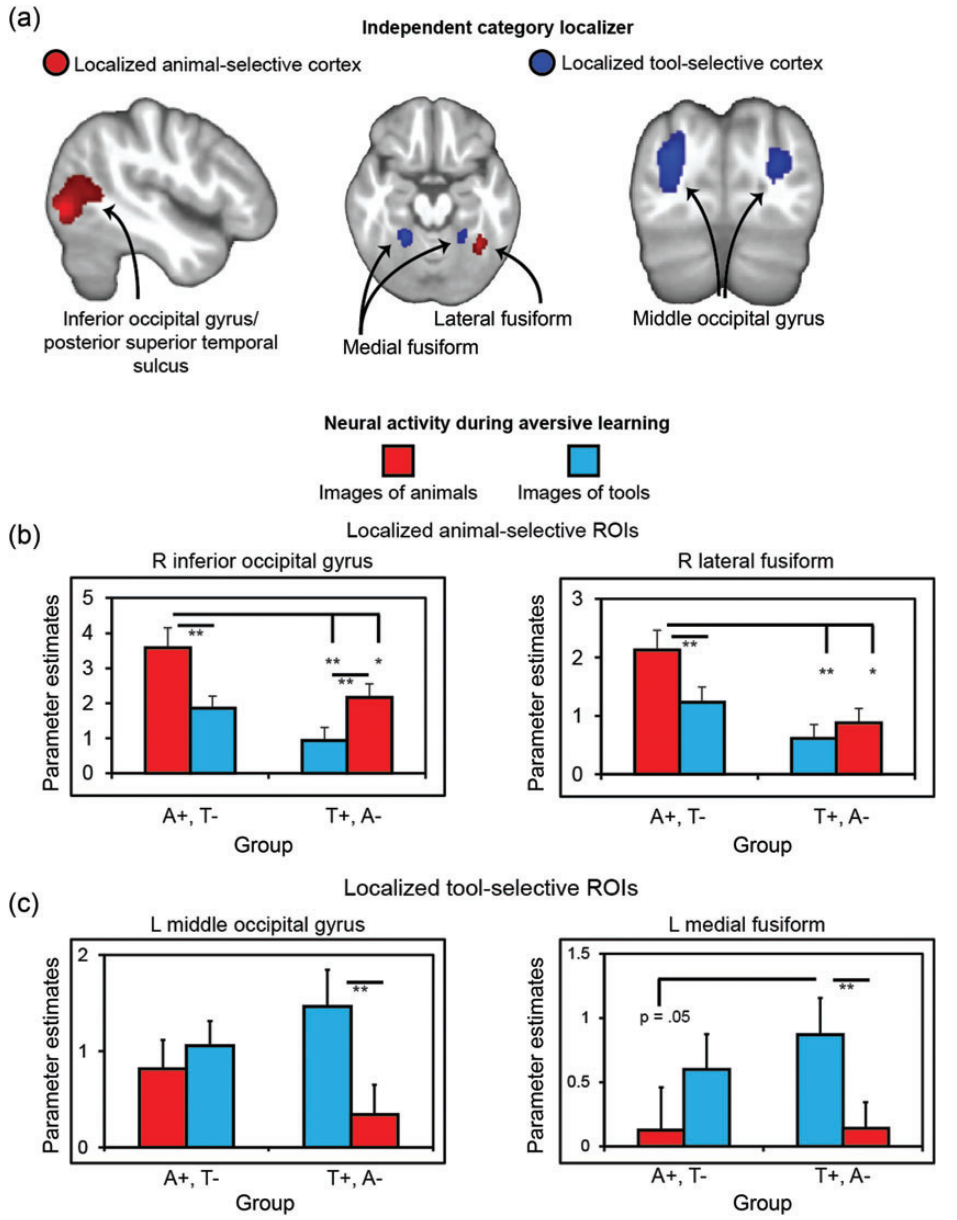

Figure 2. Activity in localized category-selective cortex during aversive learning. (*a*) Category-selective regions were independently identified prior to aversive learning through a functional localizer. The right lateral fusiform gyrus and right inferior occipital gyrus/posterior superior temporal sulcus showed selectively to images of animals, whereas bilateral medial fusiform and bilateral middle occipital gyrus showed selectively to images of tools. These category-selective regions of interest were then used to interrogate neural activity during aversive learning. (*b*) Within animal-selective regions, activity related to aversive learning was enhanced for subjects who learned to fear animals versus subjects who learned to fear tools. (*c*) Within tool-selective regions, activity related to aversive learning was enhanced for subjects who learned to fear tools versus subjects who learned to fear animals. pSTS, posterior superior temporal sulcus; ROIs, regions of interest; A+, images of animals predicted shock; T+, images of tools-predicted shock; A−, images of animals were safe; T−, images of tools were safe; Error bars represent ± SEM; $*P < 0.05$ and $**P < 0.01$, 2-tailed *t*-tests. Activations displayed on subject averaged anatomical image at $P < 0.001$, cluster corrected $P < 0.05$.

showed weak structure as a function of object category (Supplementary Fig. 6). These multivariate results indicate that aversive learning selectively enhances representational similarity among categorically related exemplars, which may be integral to facilitate the spread of learning across physically distinct objects.

Role of the Medial Temporal Lobe and Other Brain Regions

In the amygdala, learning-related effects (CS+ > CS−) were observed bilaterally in both groups, and there was no effect of group and no interaction between CS type and group (*P*s > 0.1) (Fig. 4*a*). As the amygdala's response profile typically habituates over the course of traditional fear-conditioning procedures (Buchel et al. 1998), we further interrogated its activity from voxels identified from the ROI analysis across the 4 scanning runs. Learning-related amygdala activity showed a steady decrease over time as responses to the CS+ diminished (Fig. 4*b*), implicating a time-sensitive role in initial acquisition of conditioned fear. A main effect of CS type (CS+ > CS−) was also observed in bilateral hippocampus, and there was no effect of group and no interaction between CS type and group (*P*s > 0.1)

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

Table 1.
Statistical analysis of aversive-learning related activity extracted from category-selective ROIs

Brain region	ANOVA			Post hoc *t*-tests		
	CS by category interaction	Main effect CS	Main effect category	Animals CS+ vs. tools CS−	Tools CS+ vs. animals CS−	Animals CS+ vs. Tools CS+
ROIs identified as selective to animals						
R inferior occipital gyrus	$^*P < 0.001$	$P = 0.3$	$^*P = 0.04$	$^*P < 0.001$	$^*P < 0.001$	$^*P < 0.001$
R lateral FFG	$^*P < 0.001$	$^*P = 0.04$	$^*P = 0.01$	$^*P < 0.001$	$P = 0.25$	$^*P = 0.007$
ROIs identified as selective to tools						
L middle occipital gyrus	$^*P < 0.001$	$^*P = 0.01$	$P = 0.93$	$P = 0.21$	$^*P < 0.001$	$P = 0.16$
L medial FFG	$^*P < 0.001$	$P = 0.45$	$P = 0.69$	$P = 0.06$	$^*P = 0.006$	$P = 0.05$
R medial FFG	$^*P < 0.001$	$^*P = 0.005$	$P = 0.57$	$P = 0.14$	$^*P < 0.001$	$P = 0.13$
R middle occipital gyrus	$P = 0.10$	$^*P < 0.001$	$P = 0.49$	$P = 0.09$	$^*P < 0.001$	$P = 0.97$

Note: Statistical tests were conducted on the mean parameter estimates from category-selective regions identified using an independent localizer prior to fear conditioning. See Figure 2 for brain imaging results. FFG, fusiform gyrus; L, left; R, right. *denotes significance above $P < 0.05$.

(Supplementary Table 2). Similar to the amygdala, learning-related activity in the hippocampus showed a decrease across conditioning runs due to diminished responses to the CS+. These findings indicate medial temporal lobe regions were engaged in learning to fear a category of objects but that differences in CS+ versus CS− activations were strongest in early learning, consistent with existing literature using more traditional fear-conditioning procedures (Buchel et al. 1998; LaBar et al. 1998; but see also Bach et al. 2011).

Whole-brain univariate analyses complemented ROI approaches to reveal other regions exhibiting effects of generalized aversive learning. Enhanced activity to CS+ versus CS− was observed in bilateral insula, inferior frontal gyrus, dorsolateral prefrontal cortex (PFC), thalamus, and anterior cingulate cortex (ACC) (Fig. 4*c* and Supplementary Table 2). The reverse contrast (CS− > CS+), highlighting areas preferentially sensitive to safety signals, revealed activations in ventromedial PFC and posterior cingulate cortex. In all, the collection of areas identified as showing a main effect of aversive learning have been commonly identified in traditional human neuroimaging studies employing single, repeated CS+ and CS− exemplars (LaBar and Cabeza 2006; Sehlmeyer et al. 2009; Etkin et al. 2011). These results thus extend the roles of these structures to trial-unique fear learning involving exemplars spanning a superordinate category.

Learning to Fear Animals is Characterized by Enhanced Functional Connectivity Between the Amygdala and Lateral FFG

It has been proposed that patterns of intrinsic amygdala-lateral FFG connectivity provide a key network for processing animate objects (Martin 2007a; Mahon and Caramazza 2011). We therefore examined functional connectivity between these 2 ROIs using task-based connectivity measures and during rest both before and following aversive learning, with the prediction that learning to fear animals would involve enhanced functional coupling. Task-based connectivity was examined using a PPI analysis (Friston et al. 1997), with the amygdala as the source region and the lateral FFG as the a priori target. This analysis revealed significant amygdala-lateral FFG connectivity in subjects fear conditioned to animals ($t_{16} = 2.52$, $P = 0.02$) that was enhanced relative to subjects fear conditioned to tools ($t_{15} = 2.27$, $P = 0.03$) (Fig. 5*a*). A whole-brain analysis revealed additional regions exhibiting task-based connectivity in the group fear conditioned to animals, including the insula and cingulate gyrus (see Supplementary Table 3), whereas no regions exhibited task-based amygdala connectivity in the group fear conditioned to tools. A more liberal threshold of $P < 0.005$ did not reveal any activity near category-selective extrastriate visual cortex in the group fear conditioned to tools. The PPI results in the group fear conditioned to tools suggest that learning to fear inanimate objects might rely on an alternative functional pathway that does not center on connectivity between the amygdala and tool-selective cortical ROIs.

Resting state scans acquired immediately before and after aversive learning allowed us to examine whether aversive learning modulates resting-state connectivity in a domain-specific network linked to the representation of animate objects (Simmons and Martin 2011). We focused this analysis on differences in resting state correlations between the amygdala and lateral FFG from pre- to postlearning. Correlations between the amygdala and lateral FFG increased from before to after fear conditioning in the animal CS+ group ($t_{15} = 3.87$, $P = 0.001$) but not for the tool CS+ group ($t_{15} = 0.67$, $P = 0.51$) (Fig. 5*b*). For the animal CS+ group, this significant increase in amygdala-lateral FFG connectivity was positively correlated with the change in tonic SCLs ($r_{12} = 0.54$, $P = 0.04$) (Fig. 5*c*), suggesting a relationship between increases in domain-specific resting state networks and peripheral measures of arousal following an emotional episode. Amygdala-lateral FFG connectivity and SCLs were not correlated in the tool CS+ group ($r_{13} = -0.27$, $P = 0.34$), and the difference between correlations across groups was significant, $P = 0.03$. In all, both task-based and resting-state functional connectivity analyses indicate functional enhancement of the amygdala's connectivity with animal-selective processing regions, which may provide a mechanism to support acquisition of fear to animate objects.

Hippocampal Signaling of Object Typicality Contributes to Category-Based Fear Learning Early in Training Through Interactions With the Amygdala

To test our prediction that typicality effects apply to early training trials, we used individual subjects' typicality ratings (see Experimental Procedures and Supplementary Fig. 1) in a trial-by-trial parametric modulation analysis to investigate whether brain activity was modulated by the typicality of feared category exemplars. In line with our prediction, a linear decrease in modulation by object typicality over the course of the 4 learning runs revealed selective activation in the left

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

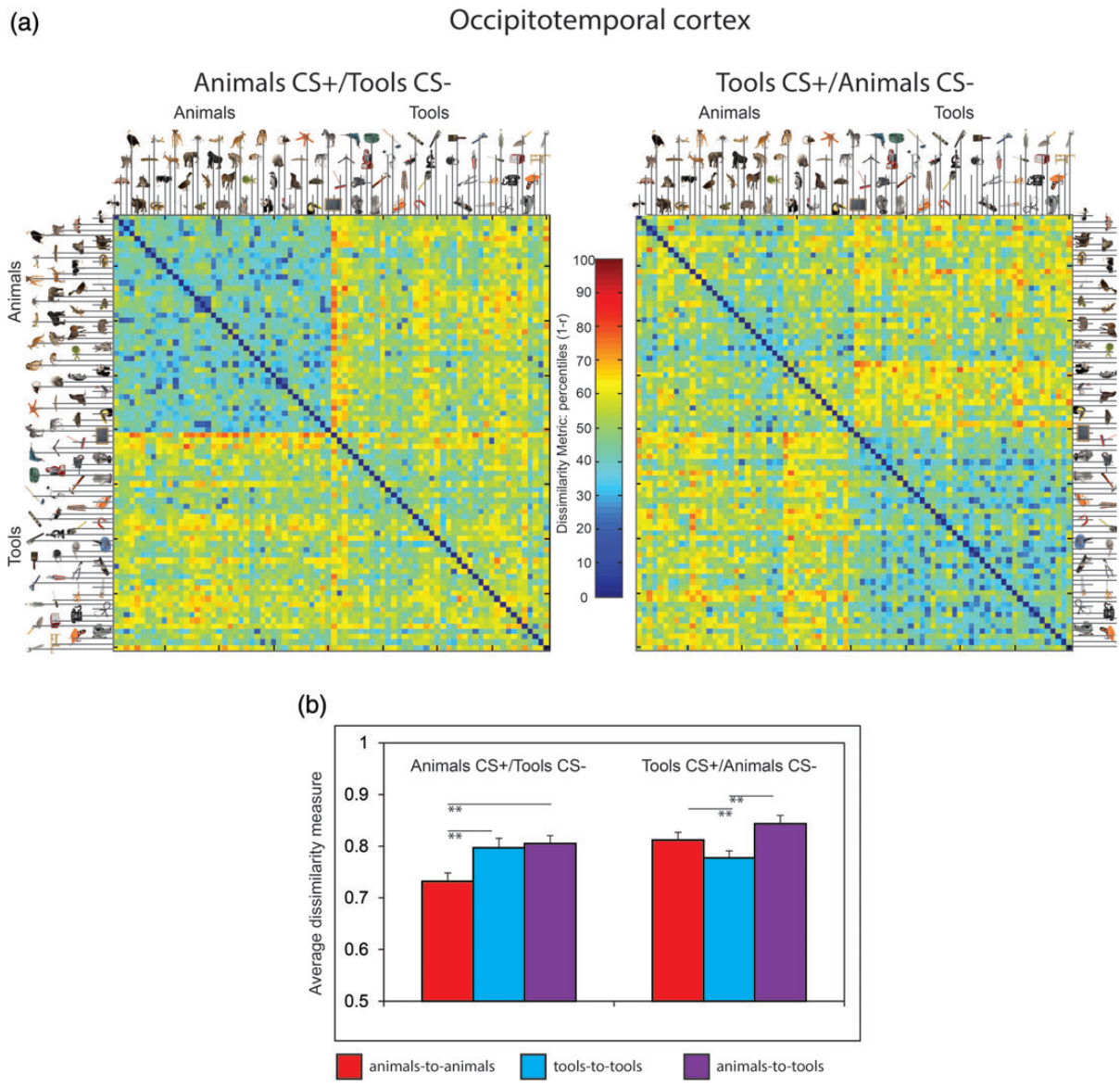

Figure 3. Representational similarity analysis. (*a*) The similarity of activity patterns across voxels in object-selective occipitotemporal cortex reflects a category boundary between animals or tools during aversive learning in groups fear conditioned to animals or tools, respectively. Results are presented in a representational dissimilarity matrix (RDM), which illustrates the correlation between activity patterns evoked by 2 different stimuli as "1-r" (Pearson correlation coefficient) (Kriegeskorte, Mur, Ruff, et al. 2008). The images used in this analysis are arranged by alphabetical order within category, with each cell depicting the similarity in response patterns between 2 stimuli. For display, the dissimilarity metric is presented in percentiles. (*b*) The average dissimilarity among images of animals, tools, and between animals and tools was quantified within each subject group using bootstrap resampling. The dissimilarity among objects from the CS+ category was reduced for each group relative to objects from within the CS− category. Error bars represent ± SEM; $^{**}P < 0.01$, 2-tailed *t*-tests. See Supplementary Methods for further details on this analysis.

hippocampus during early but not late training trials (peak-level activation at $x = -30$, $y = -20$, $z = 18$; $T = 3.75$) (Fig. 6*a*).

Based on prior research showing generalization of learning is supported by hippocampal-midbrain (Shohamy and Wagner 2008) and hippocampal-ventromedial PFC (Kumaran et al. 2009) interactions, we incorporated the hippocampus identified from the typicality analysis as a seed region in a PPI analysis to interrogate connectivity with other conditioning-related regions during early training. Results revealed a temporally graded pattern of connectivity with the left amygdala (peak-level activation at $x = -30$, $y = -2$, $z = -16$; $T = 3.46$) (Fig. 6*b*) such that task-based connectivity between the hippocampus and amygdala was strong during early training trials but then dissipated. This spatiotemporal pattern implicates one neurobehavioral mechanism through which category-level inductive inferences regarding threat value are initially generalized toward exemplars that are more representative of the feared category. These effects diminished over time, however, as participants learned through experience to broaden the category-level representation to include either typical or atypical category members—as the likelihood of receiving the US was not in fact determined by typicality.

Discussion

Although fear conditioning is traditionally considered an evolutionarily conserved system mediating only simple forms of learned behaviors, the present results reveal how fear can

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

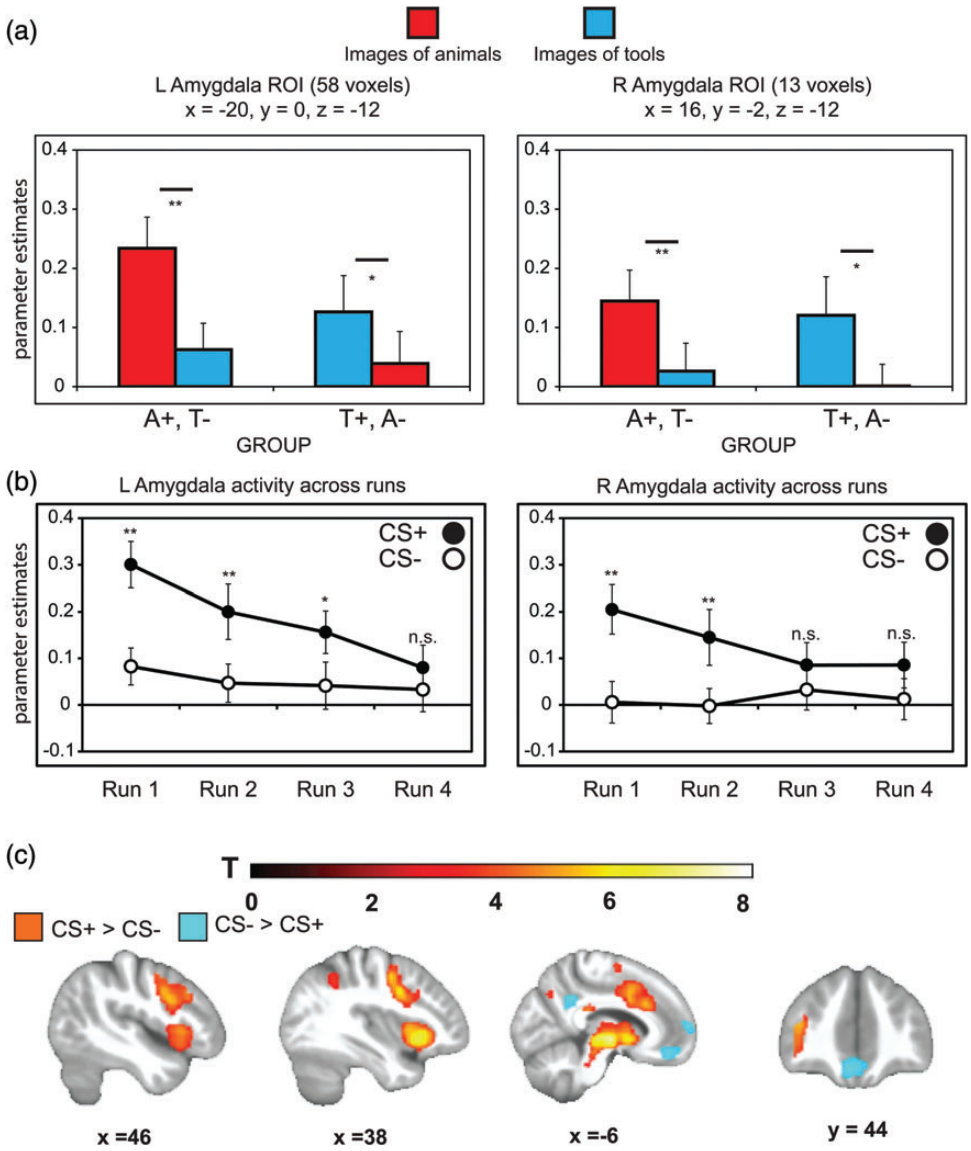

Figure 4. Amygdala and whole-brain activity during aversive learning. (*a*) Activity in the amygdala was identified from a small volume corrected region of interest (familywise error corrected $P < 0.05$). Differential fear-conditioned activity as a function of stimulus type (CS+, CS−) was observed in bilateral amygdala in subjects who learned to fear animals and in subjects who learned to fear tools. (*b*) In both left and right amygdala, differential fear-conditioned activity diminished over time as responses to the CS+ habituated. (*c*) Whole-brain analysis (cluster corrected <0.05) revealed greater activity to the CS+ versus CS− in bilateral insula, anterior cingulate cortex, and dorsolateral PFC, whereas the CS− elicited greater activity in the ventromedial PFC (see Supplementary Table 1 for a full list of regions).

generalize to more abstract, complex representations of object categories. Utilizing object concepts as stimuli during aversive learning yielded 3 primary findings. First, conjoint activation in brain areas associated with emotional processing and representation of object concepts supports aversive learning based on sparse, trial-unique reinforcement of select members of a superordinate category. Representational category structures were modulated by aversive learning, such that activity patterns in object-selective cortex were more similar for different members from the threat versus safe category. Second, we found evidence for domain-specific neural pathways in learning to fear animate (as opposed to inanimate) objects. Finally, the hippocampus was modulated by the typicality of threat exemplars and exhibited a time-delimited coupling with the amygdala that was strongest early in acquisition training. Taken together, these results reveal advanced conceptual abilities contribute to the pursuit of fear learning, and, importantly, emulate characteristics of real-world fear acquisition that may serve as a more ecologically valid model of clinical anxiety disorders characterized by widespread fear.

Prior human fMRI and nonhuman electrophysiological experiments have shown modulation in subcortical and unimodal sensory cortex as a result of aversive learning (Pape and Paré 2010). This modulatory effect helps ensure that, due to association with an aversive event, the CS acquires emotional significance and is treated appropriately the next time it is encountered. Yet, prior neurophysiological studies almost universally incorporate a single CS instance into the fear conditioning design, and thus, it is unclear whether neural modulation of a CS leads to widespread fear of related stimuli (but see Weinberger 2004). Moreover, it has remained unclear whether aversive learning modulates activation in cortical

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

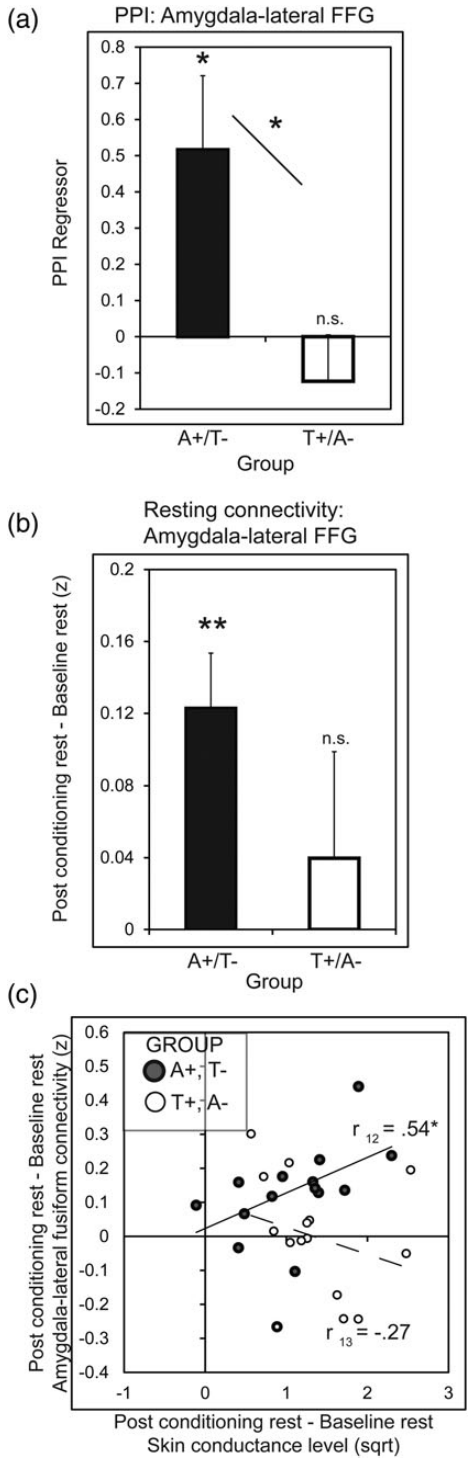

Figure 5. Amygdala-fusiform functional connectivity during learning and at rest. (*a*) Results from the psychophysiological interaction (PPI) analysis show that enhanced amygdala-lateral fusiform gyrus (animal-selective ROI) connectivity during fear conditioning was selective to the group fear conditioned to animals. (*b*) During 6-min rest scans, connectivity between the amygdala and lateral FFG increased significantly from before to after fear conditioning in the group fear conditioned to animals. (*c*) In the group conditioned to animals, increases in amygdala-lateral FFG connectivity were positively correlated with resting-state increases in tonic skin conductance levels from pre- to postconditioning. Regions were defined on the basis of functional activity during learning (amygdala; CS+ > CS−) and from an independent category localizer (fusiform gyrus; see Materials and Methods section). A+, images of animals predicted shock; T+, images of tools predicted shock; A−, images of animals were safe; T−, images of tools were safe; Error bars represent ± SEM; $^{*}P < 0.05$ and $^{**}P < 0.01$, 2-tailed *t*-tests.

regions implicated in broadly coding object categories, despite the fact that these are the types of stimuli most often encountered in real-world-learning episodes and contribute to clinical presentation in anxiety disorders such as Specific Phobia. Our results show dissociable activity in category-selective ROIs along posterior occipitotemporal cortex as a function whether participants acquired fear of animals or tools. Moreover, activations observed in other regions traditionally implicated in conditioned learning (e.g., insula, ACC, and brainstem structures) extend the role of these structures beyond fear acquisition to a single, repeated CS. These different patterns of activity, identified from a univariate fMRI analysis, are especially noteworthy given that each conditioning trial contained a unique exemplar. Thus, while participants had no direct knowledge for whether a given exemplar predicted the US, the results suggest that knowledge of the affective significance of the category enhanced the representation of different basic-level members. As many animals and tools tend to share metric features that help determine category membership (e.g., many animals have eyes), it is possible that participants could rely to some extent on perceptual strategies during this task. We therefore incorporated a range of basic-level exemplars that ranged in shape and appearance (e.g., fish, birds, insects, mammals, etc.) to minimize this possibility. In sum, these findings complement a rich behavioral literature on stimulus generalization in classical conditioning (Pavlov 1927), and extend neuroimaging findings on fear generalization beyond stimulus sets that vary parametrically along a unimodal perceptual dimension, for example, faces (Dunsmoor et al. 2011), color (Dunsmoor and Labar 2013), or shapes (Greenberg et al. 2011). (It should be noted that stimulus generalization in the classical and operant conditioning literature is frequently concerned with responses evoked by novel stimuli that are not ever reinforced in order to plot gradients of conditioned responses as a function of similarity to the CS. Rather than conducting a generalization test with unreinforced stimuli after acquisition, the present study required subjects to generalize throughout acquisition between novel stimuli that were intermittently reinforced, thus distinguishing this design from traditional stimulus generalization experiments).

The conclusion that aversive learning modulates cortical representations of object concepts is further bolstered by the multivariate RSA. This analysis distinguished between patterns of activity in occipitotemporal cortex elicited by categorically related objects, as previously demonstrated (Kriegeskorte, Mur, and Bandettini 2008; Kriegeskorte, Mur, Ruff, et al. 2008). Importantly, we discovered that representational structure of these categories was functionally altered in an experience-dependent fashion. We specifically found enhanced representational similarity between images of animals in subjects fear conditioned to animals, and analogous effects between images of tools in subjects fear conditioned to tools. Put another way, aversive learning selectively enhanced similarity structure of an object category representation that acquired threat value. A strengthening of similarity structures as a result of learning may provide a means to support broad generalization between physically dissimilar items that portend the same significant outcome.

Learning to fear animals was characterized by enhanced amygdala-lateral FFG coupling during fear acquisition, and during rest immediately following acquisition. This result

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

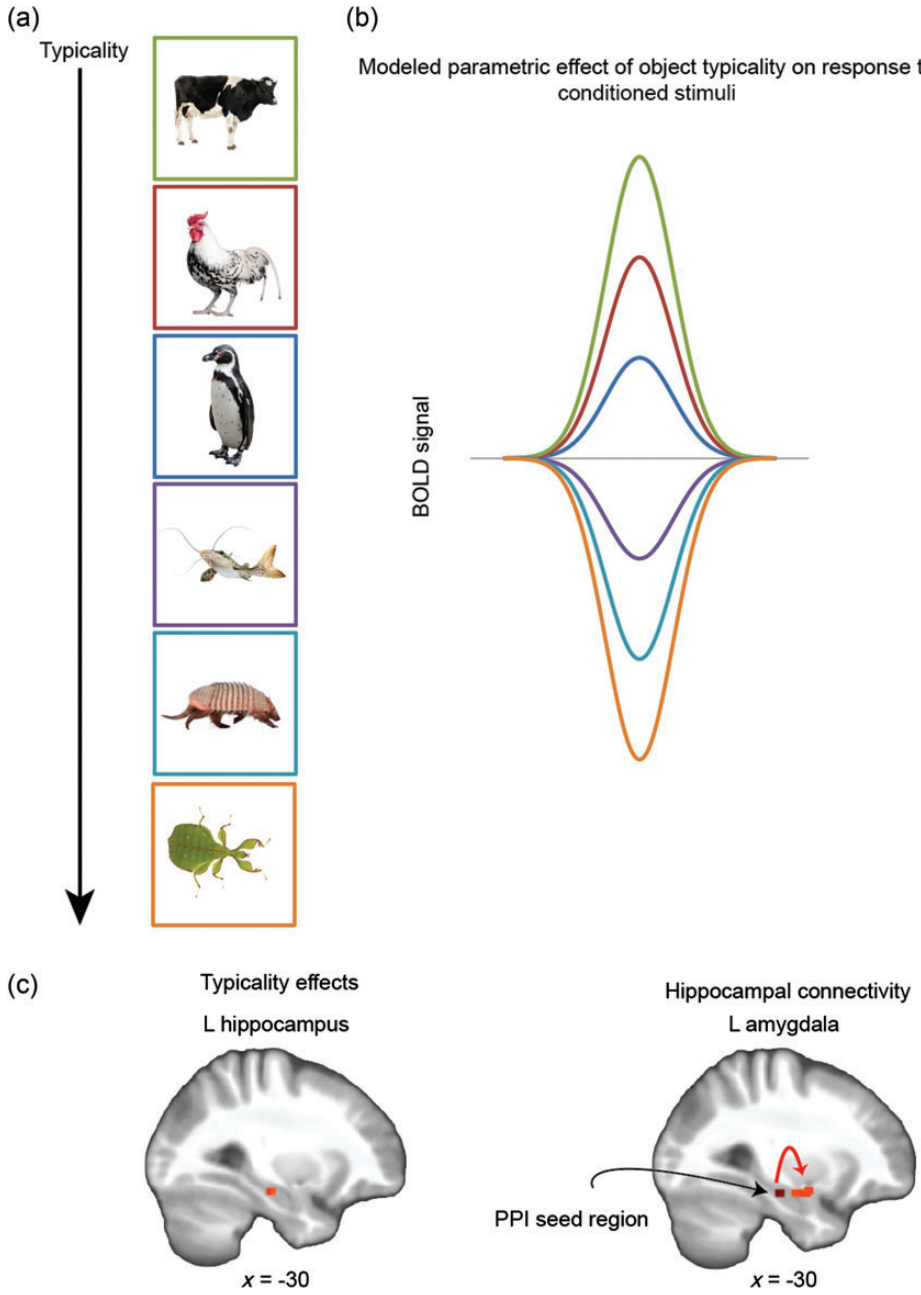

Figure 6. Activity in the hippocampus is modulated by object typicality during early aversive learning. (*a*) A schematic of an analysis using subjects' own typicality measures in a parametric modulation of fear conditioning-related brain activity. (*b*) During the early phase of aversive learning, neural activity in the left hippocampus was modulated by the typicality of objects from the feared category. (*c*) A psychophysiological interaction (PPI) analysis using the left hippocampus as a seed region showed enhanced connectivity with the left amygdala during early aversive learning. This typicality effect dissipated over time, as subjects learned to abstract fear learning to all members of the superordinate category.

complements the univariate analysis, which revealed enhanced activity across these ROIs, and the RSA analysis, which revealed strengthened representational structures as a function of learning group. Interestingly, functional connectivity was selectively enhanced for the group fear conditioned to animals. This finding fits the prediction that a domain-specific neural pathway supporting the representation of animate objects is utilized when individuals learn to fear animate (but not inanimate) objects. Intrinsic cortical connectivity with the amygdala may indicate that it was evolutionarily more important to integrate emotional information into the representation of animals and conspecifics than for artifacts (Martin 2007a; Mahon and Caramazza 2011). Thus, while both neutral animals and artifacts can enter association with an aversive US through the process of Pavlovian conditioning, separate mechanisms may be involved in the acquisition and expression of fear to objects that possess animacy and, hence, are more likely to be agents capable of delivering an aversive US. At a broader level, this finding is in line with theoretical and empirical work concerned with whether particular classes of CSs serve as prepared stimuli during aversive learning (Öhman and Mineka 2001), the importance of animacy on the neural representation of biological motion (Allison et al. 2000; Beauchamp et al. 2002) and face processing (Adolphs 2009), as well

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

as univariate fMRI (Yang et al. 2012) and electrophysiological studies showing a category-selective response to images of animals in the amygdala (Mormann et al. 2011).

It is important to note that behavioral results were similar across groups. The question is therefore raised regarding the differential patterns of amygdala connectivity identified for the animal and tool CS groups. Prior human fear-conditioning research employing classes of fear-relevant and fear-irrelevant stimuli shows that subjects have similar US expectancy (Öhman and Mineka 2001) and conditioned SCRs (Olsson et al. 2005) across stimulus classes; the differences exist primarily in the rate of extinction (Hugdahl and Ohman 1977; Olsson et al. 2005) and conditioning outside awareness (Soares and Ohman 1993). Although we did not incorporate highly threat-relevant items in this task, one possibility is that domain-specific pathways between the amygdala and extrastriate visual cortex promote acquisition and expression of fear to animal exemplars under conditions that are more ambiguous than those employed here (e.g., presentation outside awareness) as well as maintained responses to novel animal exemplars during extinction.

Finally, we discovered a typicality effect during early aversive learning. The hippocampus, a region commonly implicated in generalization processes (Gluck and Myers 1993; Shohamy and Wagner 2008; Kumaran et al. 2009), exhibited a time-delimited modulation by object typicality of feared exemplars and a time-sensitive functional coupling with the amygdala. Hippocampal-amygdala coupling may provide a mechanism through which category-level inductive inferences are generalized preferentially toward typical category members. Typicality effects play a well-known role in category-based induction (Rips 1975; Osherson et al. 1990; Murphy 2002). For instance, a premise that incorporates a typical category member tends to lead to stronger arguments and broader generalizations (Rips 1975; Osherson et al. 1990). The role of typicality in fear learning is unknown but may be relevant to processing certain threats, since some exemplars are regarded as more prototypically dangerous than others. For example, if an individual develops a fear of dogs following an aversive experience with a particular dog, they may be more likely to generalize from this experience to dogs that are more prototypically threatening (e.g., a Pit bull vs. a Chihuahua). In the context of the present study, experience with different CS+ trials over the course of the session likely weakened the status of typicality, resulting in a broadening of the category-level representation of threat in line with the notion of premise diversity and premise monotonicity described by Osherson et al. (1990).

These findings raise a number of intriguing questions for future research. First, as we show evidence that conceptual systems can be utilized during fear acquisition, it is worth considering whether these systems can be recruited during fear "extinction," when the CS no longer predicts danger (Milad and Quirk 2012). For instance, a prediction based on the category induction literature would propose that extinction to more typical (Rips 1975) or diverse (Osherson et al. 1990) category members would lead to better generalization of extinction. Another question for future research is whether these fMRI findings are selective to aversive learning per se. The present study was based on fear conditioning models, which are routinely used in basic neuroscience research on learning and memory and have informed models of clinical anxiety (Foa et al. 1989; Milad and Quirk 2012). However, fear-independent learning processes undoubtedly contribute to category-based generalizations as well, and so fMRI results in category-selective regions may extend to other domains, such as appetitive learning. Finally, it is important to consider the role of attention when learning about stimuli at the category level. For instance, attentional modulation can influence activity in category-selective extrastriate cortex (Johnson and Johnson 2009). However, attention is also a core component of associative learning models and, according to the influential Pearce and Hall (1980) model, attention is expected to increase if the outcome is uncertain (i.e., in the case of intermittent reinforcement or when a cue is novel). It is therefore possible that aversive learning helps increase the relevance of items detected as belonging to a feared category and therefore judged to be potential threats.

An essential question raised by these findings is by what mechanism are categorical representations modulated by fear conditioning. A low-level account for these results is provided by research in the nonhuman animal domain showing that fear conditioning induced changes in auditory cortex rely on auditory input from the amygdala and auditory thalamus (Pape and Paré 2010). Visual stimuli are far less utilized in rodent neurophysiological fear conditioning (Shi and Davis 2001), but investigations in monkeys have traced connections from the amygdala to targets along the occipitotemporal cortex (Kravitz et al. 2013). Human neuroimaging investigations have provided a broad role for this pathway in modulating visual stimuli that have both intrinsic and acquired affective properties. For example, viewing faces predictive of an aversive shock is associated with conjoint activations and functional connectivity between the amygdala and FFG (Petrovic et al. 2008; Dunsmoor et al. 2011). A low-level conditioning account may support a categorical fear of animate objects, as evidenced by the connectivity results. Alternatively, a mechanism that involves conscious deployment of conceptual processing abilities (e.g., explicit categorization) may provide another pathway to acquire category-level representations of threat. Of note, the univariate modulation observed in the present study in category-selective regions could reflect intrinsic generalization processes, or may be a consequence of generalization processes occurring elsewhere in the brain. Although future investigations are needed to resolve whether there is a precise source of category-based fear generalizations, we propose that mechanisms involved in low-level fear conditioning, higher order category formation, and conceptual representations can operate in tandem to acquire fear to known objects in novel situations.

In conclusion, our findings show that learning to fear an object category modulates activity and representational architecture of category-selective cortex and the amygdala. For subjects fear conditioned with animate objects, this modulation may be provided by a special mechanism through connectivity between the amygdala and lateral occipitotemporal cortex. Finally, typicality effects during early training impact category-based fear learning through hippocampal signaling and functional interactions with the amygdala. Once the relevant superordinate category representation is activated, this role for the hippocampus decreases in importance as participants begin generalizing fear to nonreinforced category exemplars, resulting in enhanced local representations in category-selective cortices. In sum, these findings demonstrate myriad ways humans learn to fear stimuli from the environment, and provide potential avenues for understanding how conceptual systems are recruited in the overgeneralization of fear characteristic of clinical anxiety disorders.

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

Supplementary Material

Supplementary material can be found at: http://www.cercor.oxfordjournals.org/.

Funding

This work was supported by NSF grant 0745919 and NIH grants R01 DA027802 and F31 MH090682.

Notes

Conflict of Interest: None declared.

References

Adolphs R. 2009. The social brain: neural basis of social knowledge. Annu Rev Psychol. 60:693–716.

Allison T, Puce A, McCarthy G. 2000. Social perception from visual cues: role of the STS region. Trends Cogn Sci. 4:267–278.

Bach DR, Weiskopf N, Dolan RJ. 2011. A stable sparse fear memory trace in human amygdala. J Neurosci. 31:9383–9389.

Beauchamp MS, Lee KE, Haxby JV, Martin A. 2002. Parallel visual motion processing streams for manipulable objects and human movements. Neuron. 34:149–159.

Buchel C, Morris J, Dolan RJ, Friston KJ. 1998. Brain systems mediating aversive conditioning: an event-related fMRI study. Neuron. 20:947–957.

Chao LL, Martin A. 2000. Representation of manipulable man-made objects in the dorsal stream. Neuroimage. 12:478–484.

Dunsmoor JE, LaBar KS. 2013. Effects of discrimination training on fear generalization gradients and perceptual classification in humans. Behav Neurosci. doi:10.1037/a0031933.

Dunsmoor JE, Martin A, LaBar KS. 2012. Role of conceptual knowledge in learning and retention of conditioned fear. Biol Psychol. 89:300–305.

Dunsmoor JE, Mitroff SR, LaBar KS. 2009. Generalization of conditioned fear along a dimension of increasing fear intensity. Learn Mem. 16:460–469.

Dunsmoor JE, Prince SE, Murty VP, Kragel PA, LaBar KS. 2011. Neurobehavioral mechanisms of human fear generalization. Neuroimage. 55:1878–1888.

Etkin A, Egner T, Kalisch R. 2011. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends Cogn Sci. 15:85–93.

Foa EB, Steketee G, Rothbaum BO. 1989. Behavioral cognitive conceptualizations of post-traumatic stress disorder. Behav Ther. 20:155–176.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ. 1997. Psychophysiological and modulatory interactions in neuroimaging. Neuroimage. 6:218–229.

Gluck MA, Myers CE. 1993. Hippocampal mediation of stimulus representations—a computational theory. Hippocampus. 3:491–516.

Greenberg T, Carlson JM, Cha J, Hajcak G, Mujica-Parodi LR. 2011. Neural reactivity tracks fear generalization gradients. Biol Psychol. 92:2–8.

Honig WK, Urcuioli PJ. 1981. The legacy of Guttman and Kalish (1956)—25 years of research on stimulus-generalization. J Exp Anal Behav. 36:405–445.

Hugdahl K, Ohman A. 1977. Effects of instruction on acquisition and extinction of electrodermal responses to fear-relevant stimuli. J Exp Psychol Hum Learn. 3:608–618.

Ishai A, Ungerleider LG, Haxby JV. 2000. Distributed neural systems for the generation of visual images. Neuron. 28:979–990.

Johnson MR, Johnson MK. 2009. Top-down enhancement and suppression of activity in category-selective extrastriate cortex from an act of reflective attention. J Cogn Neurosci. 21:2320–2327.

Knight DC, Nguyen HT, Bandettini PA. 2005. The role of the human amygdala in the production of conditioned fear responses. Neuroimage. 26:1193–1200.

Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M. 2013. The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends Cogn Sci. 17:26–49.

Kriegeskorte N, Mur M, Bandettini P. 2008. Representational similarity analysis—connecting the branches of systems neuroscience. Front Syst Neurosci. 2:4.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA. 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron. 60:1126–1141.

Kumaran D, Summerfield JJ, Hassabis D, Maguire EA. 2009. Tracking the emergence of conceptual knowledge during human decision making. Neuron. 63:889–901.

LaBar KS, Cabeza R. 2006. Cognitive neuroscience of emotional memory. Nat Rev Neurosci. 7:54–64.

LaBar KS, Gatenby JC, Gore JC, LeDoux JE, Phelps EA. 1998. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. Neuron. 20:937–945.

Laird AR, Robinson JL, McMillan KM, Tordesillas-Gutierrez D, Moran ST, Gonzales SM, Ray KL, Franklin C, Glahn DC, Fox PT et al. 2010. Comparison of the disparity between Talairach and MNI coordinates in functional neuroimaging data: validation of the Lancaster transform. Neuroimage. 51:677–683.

Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT. 2000. Automated Talairach atlas labels for functional brain mapping. Hum Brain Mapp. 10:120–131.

Larocque KF, Smith ME, Carr VA, Witthoft N, Grill-Spector K, Wagner AD. 2013. Global similarity and pattern separation in the human medial temporal lobe predict subsequent memory. J Neurosci. 33:5466–5474.

LeDoux JE. 2000. Emotion circuits in the brain. Annu Rev Neurosci. 23:155–184.

Letzkus JJ, Wolff SB, Meyer EM, Tovote P, Courtin J, Herry C, Luthi A. 2011. A disinhibitory microcircuit for associative fear learning in the auditory cortex. Nature. 480:331–335.

Li W, Howard JD, Parrish TB, Gottfried JA. 2008. Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. Science. 319:1842–1845.

Lim SL, Padmala S, Pessoa L. 2008. Affective learning modulates spatial competition during low-load attentional conditions. Neuropsychologia. 46:1267–1278.

Mahon BZ, Caramazza A. 2011. What drives the organization of object knowledge in the brain? Trends Cogn Sci. 15:97–103.

Mahon BZ, Milleville SC, Negri GA, Rumiati RI, Caramazza A, Martin A. 2007. Action-related properties shape object representations in the ventral stream. Neuron. 55:507–520.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH. 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage. 19:1233–1239.

Maren S. 2001. Neurobiology of Pavlovian fear conditioning. Annu Rev Neurosci. 24:897–931.

Martin A. 2007a. Neural foundations for conceptual representations: evidence from functional brain imaging. In: Hart J, Kraut MA, editors. Neural basis of semantic memory. New York: Cambridge University Press. p. 302–330.

Martin A. 2007b. The representation of object concepts in the brain. Annu Rev Psychol. 58:25–45.

Milad MR, Quirk GJ. 2012. Fear extinction as a model for translational neuroscience: ten years of progress. Annu Rev Psychol. 63:129–151.

Mormann F, Dubois J, Kornblith S, Milosavljevic M, Cerf M, Ison M, Tsuchiya N, Kraskov A, Quiroga RQ, Adolphs R et al. 2011. A category-specific response to animals in the right human amygdala. Nat Neurosci. 14:1247–1249.

Murphy GL. 2002. The big book of concepts. Cambridge, MA: MIT Press.

Öhman A, Mineka S. 2001. Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. Psychol Rev. 108:483–522.

Olsson A, Ebert JP, Banaji MR, Phelps EA. 2005. The role of social groups in the persistence of learned fear. Science. 309:785–787.

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016

Osherson DN, Wilkie O, Smith EE, Lopez A, Shafir E. 1990. Category-based induction. Psychol Rev. 97:185–200.
Pape HC, Paré D. 2010. Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiol Rev. 90:419–463.
Patterson K, Nestor PJ, Rogers TT. 2007. Where do you know what you know? The representation of semantic knowledge in the human brain. Nat Rev Neurosci. 8:976–987.
Pavlov IP. 1927. Conditioned reflexes. London: Oxford University Press.
Pearce JM, Hall G. 1980. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev. 87:532–552.
Petrovic P, Kalisch R, Pessiglione M, Singer T, Dolan RJ. 2008. Learning affective values for faces is expressed in amygdala and fusiform gyrus. Soc Cogn Affect Neurosci. 3:109–118.
Quiroga RQ. 2012. Concept cells: the building blocks of declarative memory functions. Nat Rev Neurosci. 13:587–597.
Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. 2005. Invariant visual representation by single neurons in the human brain. Nature. 435:1102–1107.
Rips LJ. 1975. Inductive judgments about natural categories. J Verbal Learn Verbal Behav. 14:665–681.
Rosch E, Mervis CB. 1975. Family resemblances—studies in internal structure of categories. Cogn Psychol. 7:573–605.
Sehlmeyer C, Schoning S, Zwitserlood P, Pfleiderer B, Kircher T, Arolt V, Konrad C. 2009. Human fear conditioning and extinction in neuroimaging: a systematic review. Plos One. 4:16.
Shi C, Davis M. 2001. Visual pathways involved in fear conditioning measured with fear-potentiated startle: behavioral and anatomic studies. J Neurosci. 21:9844–9855.
Shohamy D, Wagner AD. 2008. Integrating memories in the human brain: hippocampal-midbrain encoding of overlapping events. Neuron. 60:378–389.
Simmons WK, Martin A. 2011. Spontaneous resting-state BOLD fluctuations reveal persistent domain-specific neural networks. Soc Cogn Affect Neurosci. 7:467–475.
Soares JJ, Ohman A. 1993. Backward masking and skin conductance responses after conditioning to nonfeared but fear-relevant stimuli in fearful subjects. Psychophysiology. 30:460–466.
Tranel D, Damasio H, Damasio AR. 1997. A neural basis for the retrieval of conceptual knowledge. Neuropsychologia. 35:1319–1327.
Weinberger NM. 2007. Associative representational plasticity in the auditory cortex: a synthesis of two disciplines. Learn Mem. 14:1–16.
Weinberger NM. 2004. Specific long-term memory traces in primary auditory cortex. Nat Rev Neurosci. 5:279–290.
Yang J, Bellgowan PS, Martin A. 2012. Threat, domain-specificity and the human amygdala. Neuropsychologia. 50:2566–2572.

Downloaded from http://cercor.oxfordjournals.org/ at Duke University Law School on January 28, 2016