

# Human orbitofrontal cortex mediates extinction learning while accessing conditioned representations of value

Jay A Gottfried<sup>1,2</sup> & Raymond J Dolan<sup>1</sup>

**In extinction, an animal learns that a previously conditioned stimulus (CS+) no longer predicts delivery of a salient reinforcer (unconditioned stimulus, UCS). Rodent studies indicate that extinction relies on amygdala-prefrontal interactions and involves formation of memories that inhibit, without actually erasing, the original conditioning trace. Whether extinction learning in humans follows similar neurobiological principles is unknown. We used functional magnetic resonance imaging to measure human brain activity evoked during olfactory aversive conditioning and extinction learning. Neural responses in orbitofrontal cortex and amygdala were preferentially enhanced during extinction, suggesting potential cross-species preservation of learning mechanisms that oppose conditioning. Moreover, by manipulating UCS aversiveness via reinforcer inflation, we showed that a CS+ retains access to representations of UCS value in distinct regions of ventral prefrontal cortex, even as extinction proceeds.**

How an organism learns to predict danger (e.g., predators, poisons) is expressed in its most basic form in aversive classical (pavlovian) conditioning. In this type of emotional learning, a neutral item (the conditioned stimulus, CS+) acquires behavioral relevance through pairing with a salient reinforcer (the unconditioned stimulus, UCS). Typically the UCS consists of electric shocks<sup>1,2</sup>, but unpleasant smells and tastes are equally potent<sup>3,4</sup>. Although the amygdala is known to be centrally involved in conditioning<sup>1,2</sup>, recent animal<sup>5,6</sup> and human<sup>3,4</sup> studies indicate that orbital and ventromedial prefrontal cortices also participate in the establishment of aversive CS:UCS links.

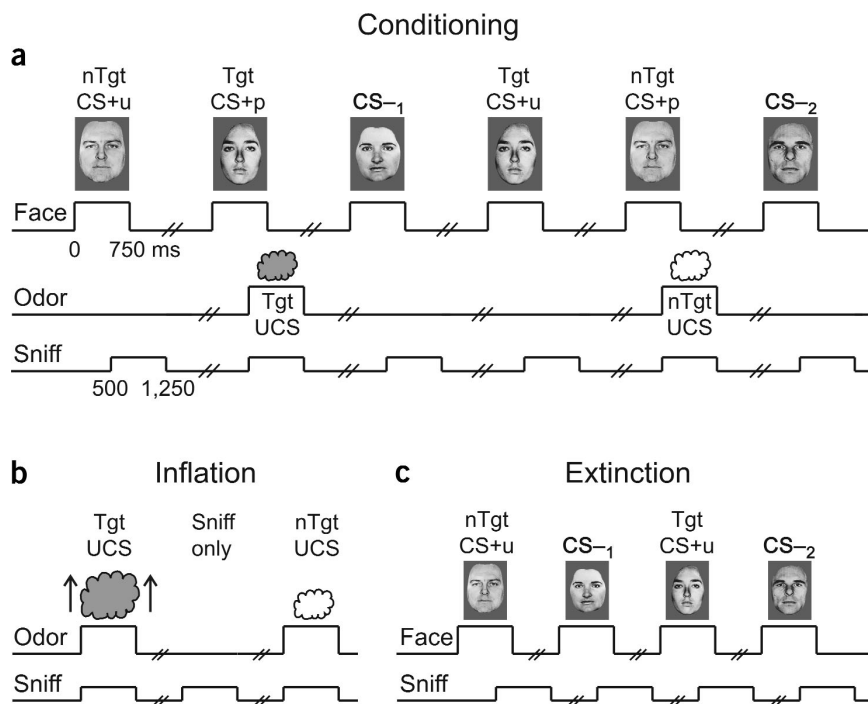
How an organism learns to disregard danger cues (when no longer a threat) can be studied using extinction protocols, whereby successive presentations of a nonreinforced CS+ (following conditioning) diminish conditioned responses (CRs). Animal research indicates that extinction is not simply unlearning of the original contingency, but reflects new learning that opposes, or inhibits, expression of conditioning<sup>7,8</sup>. Thus, the prevailing view is that conditioning and extinction lead to the formation of two distinct memory representations: CS:UCS (excitatory) and CS:no UCS (inhibitory). Which of these competing memories is activated by a given CS+ is influenced by sensory, environmental and temporal contexts<sup>9–11</sup>.

In animals, the neural mechanisms of extinction are only now beginning to be characterized<sup>8</sup>. Recent work suggests that interactions between medial prefrontal cortex (PFC) and discrete amygdala subnuclei<sup>12,13</sup> may regulate extinction processes, possibly through stimulation of NMDA and GABA receptors<sup>8</sup>, although the specific contributions of these systems may differ as a function of postextinc-

tion training intervals<sup>8</sup>. By comparison, the neural substrates of extinction in humans are poorly understood, despite the possibility that a failure of extinction learning may provide an explanatory basis for specific psychopathologies such as phobias and posttraumatic stress disorder<sup>14</sup>. One previous human imaging study of extinction reported amygdala activation, but characterization of responses in prefrontal regions was not possible because of technical limitations<sup>15</sup>. Important issues related to the neurobiology of extinction in humans therefore remain unresolved. For example, it is unclear whether conditioning and extinction are supported by common or distinct brain networks, and whether the systems mediating human extinction share structural correspondence to those identified in animal models. Furthermore, the proposal that a conditioned memory is accessible during extinction implies that a signature of the CS:UCS trace must persist, but no such representation has been documented in humans.

In the present study, we used event-related functional magnetic resonance imaging (fMRI) techniques to measure region-specific brain activity in human subjects during olfactory aversive conditioning (Fig. 1). This was followed in the same session by extinction, permitting a direct comparison between neural responses evoked during conditioning and extinction learning. Moreover, by manipulating the postconditioning, pre-extinction value of a target reinforcer via UCS inflation<sup>16,17</sup>, we were able to specifically 'tag' a neural representation of the UCS, in order to index its presence during the extinction procedure. Using these approaches we show that specific regions of orbitofrontal cortex (OFC) and lateral amygdala are preferentially activated in extinction. Our findings also indicate that during extinc-

<sup>1</sup>Wellcome Department of Imaging Neuroscience, Functional Imaging Laboratory, 12 Queen Square, London, WC1N 3BG, UK. <sup>2</sup>Present address: Cognitive Neurology and Alzheimer's Disease Center, 320 East Superior Street, Searle 11-453, Chicago, Illinois 60611, USA. Correspondence should be addressed to J.A.G. (j-gottfried@northwestern.edu).



**Figure 1** Experimental design. (a) In an olfactory version of aversive conditioning, two neutral faces (CS+) were repetitively paired with two different unpleasant odors (UCS). One odor, destined for UCS inflation, was designated the target (Tgt) UCS; the other served as the nontarget (nTgt) UCS. One-half of all CS+ presentations were coupled to the UCS, resulting in paired (CS+p) and unpaired (CS+u) trials. Two other faces were never paired with odor (CS<sub>-1</sub>, CS<sub>-2</sub>). Subjects were cued to sniff on every trial. (b) Immediately after conditioning, subjects underwent UCS inflation, whereby the Tgt UCS (at increased odor intensity) and nTgt UCS (at baseline intensity) were presented in the absence of the CS+, in order to enhance Tgt UCS aversiveness. (c) In a final extinction session, CS+ and CS- stimuli were delivered in the absence of the UCS.

tion, a predictive CS+ accesses dual representations of UCS value in dissociable regions of ventral PFC, implying these areas are critical in retaining associative links initially established during conditioning, even as extinction proceeds.

## RESULTS

### Aversive conditioning

We first characterized the behavioral correlates of aversive conditioning. Differential reaction times (RTs) provided an objective index of conditioning (Fig. 2a). Subjects responded significantly faster to the CS+ (target and nontarget) relative to the nonconditioned stimulus (CS-) in the first half-session (target CS+ versus CS-,  $P = 0.050$ ; nontarget CS+ versus CS-,  $P = 0.030$ ; one-tailed paired  $t$ -test). Subsequent response habituation in the second half-session was marked by an absence of significant differences between conditions (target CS+ versus CS-,  $P = 0.12$ ; nontarget CS+ versus CS-,  $P = 0.075$ ; one-tailed paired  $t$ -test), despite an increase in response latency to the target CS+. Our prior study on olfactory conditioning showed a similar RT profile for an aversive (but not an appetitive) CS+ (ref. 3). This suggests an initial response advantage to an aversive CS+ may be increasingly opposed by response interference due to greater attentional capture (and concomitant task distraction), as a subject learns that the CS+ predicts an aversive event. As a complementary measure, *post-hoc* debriefing conducted after the experiment indicated that 13/16 subjects became aware of CS:UCS contingencies. Finally, we note that postconditioning valence ratings of the CS+ faces were judged as more aversive than the CS- faces (significant trend,  $P = 0.050$ ; one-tailed paired  $t$ -test), in the absence of significant valence differences between target and nontarget CS+ types ( $P = 0.50$ ; two-tailed paired  $t$ -test).

The accompanying neural substrates of conditioning were identified in the conjunction of target (unpaired) CS+u and nontarget (unpaired) CS+u (each minus their respective CS-), which tested for brain regions commonly activated in response to both CS+ types, in the absence of odor UCS confounds. This contrast revealed significant conditioning-related responses in rostromedial

ing was proposed to support states of autonomic arousal and interoceptive awareness in the setting of external threat<sup>20</sup>.

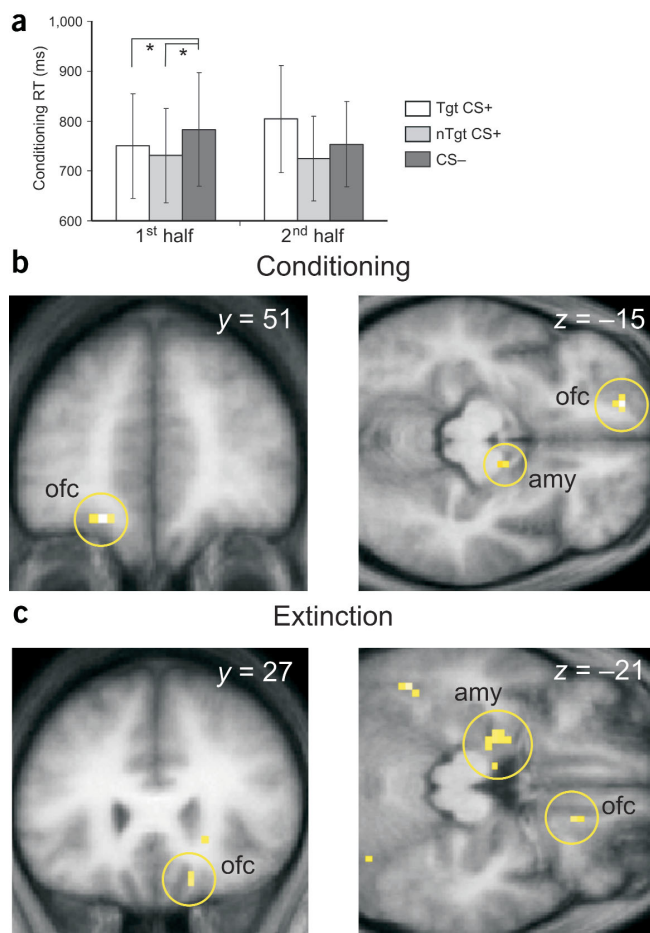
These results were not confounded by condition-specific respiratory differences, as no significant differences in sniff peak-amplitude ( $F_{2,3,34.9} = 1.48$ ;  $P > 0.2$ ; repeated-measures ANOVA; Greenhouse-Geisser corrected for nonsphericity) or sniff latency ( $F_{3,2,48.4} < 1$ ;  $P > 0.6$ ) were evident. Moreover, subjective ratings of target and nontarget UCS odor intensity (target,  $11.14 \pm 0.70$ ; nontarget,  $10.78 \pm 0.78$ ; mean  $\pm$  s.e.m.) and valence (target,  $13.27 \pm 0.77$ ; nontarget,  $13.51 \pm 0.54$ ) revealed no significant differences (both  $P$ 's  $> 0.7$ ; two-tailed paired  $t$ -test).

### Extinction learning

All 13 subjects reporting initial awareness of CS:UCS contingencies became aware that the CS+ items no longer predicted the UCS during the final scanning phase, providing an explicit measure of extinction learning (see below for analysis of extinction RTs showing a differential inflation effect). Thus, our next analysis examined brain responses involved during extinction learning. By performing a conjunction of target CS+u and nontarget CS+u (minus CS- baselines), we tested for commonalities in response at extinction, independently of UCS inflation. Significant conjoint activations in caudal OFC, ventromedial PFC and lateral amygdala (Fig. 2c and Table 1) were detected in this analysis. Notably, this network of activity encompasses homologous regions implicated in animal studies of extinction learning<sup>11,21-25</sup>. We again note an absence of significant respiratory differences between trial types, for either sniff peak amplitude ( $F_{2,5,37.2} = 1.07$ ;  $P > 0.3$ ) or sniff latency ( $F_{2,6,38.8} < 1$ ;  $P > 0.4$ ), during extinction.

We next considered whether the neural substrates of extinction learning overlap those involved in learning. The conjunction analysis of areas mutually activated in both conditioning and extinction sessions [conditioning + extinction] revealed significant common responses in medial amygdala (Fig. 3a-c), rostromedial OFC, PFC, insula and ventral striatum (Table 2). To identify regions exhibiting temporal changes in their response profile, we also tested

**Figure 2** Behavioral and neural substrates of conditioning and extinction. (a) Reaction times (RTs) provided a behavioral index of conditioning. Subjects responded more quickly to Tgt and nTgt CS+ stimuli, compared to CS-, during the first half of conditioning ( $*P \leq 0.05$ ; means  $\pm$  s.e.m. are shown). (b) Conditioning-related neural responses were detected in rostromedial OFC and amygdala (amy) and are overlaid on coronal (left) and horizontal (right) sections from the group-averaged T1-weighted image (threshold for display,  $P < 0.001$ ). (c) Extinction-related neural responses were identified in medial OFC and bilateral amygdala ( $P < 0.001$ ).



for a condition  $\times$  time interaction<sup>19</sup>. This analysis indicated that the activity in medial amygdala and rostromedial OFC significantly decreased over time (habituated) during both conditioning and extinction (Table 2).

In comparison, a direct contrast of [extinction – conditioning] tested for functional dissociations between these sessions. This indicated that neural responses in lateral amygdala, rostromedial OFC/gyrus rectus and caudal OFC were preferentially enhanced during extinction learning, over and above conditioning-evoked activity, for both CS+u types (Fig. 3 and Table 2). These peak activations occurred in the absence of significant interactions (at  $P < 0.05$  uncorrected) between phase (conditioning versus extinction) and CS+ type (target versus nontarget). No significant activations in *a priori* regions were detected in the reverse comparison of conditioning with extinction.

#### Persisting CS:UCS representations at extinction

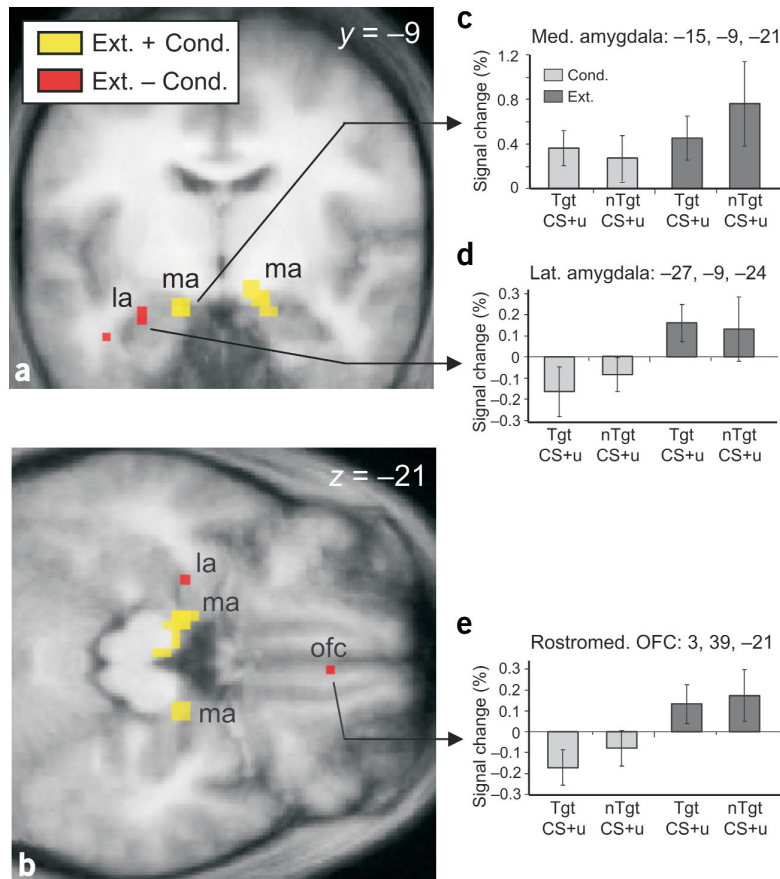
A central focus was to identify sites of persisting CS:UCS representations during extinction. However, the above findings do not enable a distinction to be made between CS+-evoked activation of UCS representations and those related more generally to extinction learning. Consequently, we used UCS inflation<sup>16</sup> to create an updated trace of UCS value that could be selectively indexed during extinction. This technique involves postconditioning presentations of the UCS alone at increased intensity, to heighten its aversiveness. As a result, conditioned responses subsequently elicited by the CS+ become accentuated, indicating that the predictive cue accesses an updated representation of UCS value. The versatility of this method has been demonstrated in both animal<sup>16,17</sup> and human<sup>26,27</sup> models of learning.

We assessed the behavioral impact of UCS inflation on CS+ processing during extinction in two ways. First, subjective ratings of CS+ aversiveness were significantly higher for the target CS+ relative to the nontarget CS+ ( $P = 0.038$ ; one-tailed paired *t*-test) (Fig. 4a). This was expected, since our analysis excluded those subjects ( $n = 3/16$ ) who rated the CS+ types in the opposite directions. Second, UCS inflation had a differential effect on RTs. Compared to conditioning (second-half), subjects responded significantly faster to the target CS+ during extinction than to the CS- ( $P = 0.040$ ; one-tailed paired *t*-test) (Fig. 4b), though we note that a bias toward finding a latency reduction to the postinflation target CS+ could have resulted from a longer preinflation latency (confer Fig. 2a). Taken together, these measures indicate that UCS inflation successfully altered the behavioral salience of the target CS+, in a sensory-specific manner. Critically, this effect was not contingent on explicit pairing between the CS+ and the UCS, because after initial conditioning, subjects no longer experienced these items in combination. Thus, the most parsimonious explanation of UCS inflation is that at extinction, the target CS+ accessed a current, and updated, representation of UCS aversive value.

**Table 1** Conditioning- and extinction-related neural activations

Brain region	MNI coordinates (mm)			Vol. (mm <sup>3</sup> )	Peak Z	P value
	x	y	z			
<b>Conditioning</b>						
Left rostromedial OFC	-18	51	-15	108	4.52	< 0.05 (SVC)
Left insula	-33	21	9	27	3.21	< 0.001
Right insula	36	15	9	27	3.45	< 0.001
Right dorsomedial amygdala	12	-9	-12	81	3.41	< 0.05 (SVC)
Left ventral midbrain	-9	-18	-21	81	3.43	< 0.05 (SVC)
<b>Extinction</b>						
Left ventromedial PFC	-9	30	0	54	3.34	< 0.001
Right caudomedial OFC	15	27	-21	108	3.62	< 0.05 (SVC)
Right caudomedial OFC	12	21	-12	27	3.15	< 0.001
Right insula	39	12	9	162	4.10	< 0.05 (SVC)
Left insula	-33	3	18	54	3.30	< 0.001
	-33	15	12	27	3.12	< 0.001
Right lateral amygdala	18	3	-27	27	3.40	< 0.05 (SVC)
Left lateral amygdala	-27	-9	-15	378	3.97	< 0.05 (SVC)
Left medial amygdala	-9	-9	-21	27	3.35	< 0.001

SVC, corrected for small volume of interest.



**Figure 3** Functional overlaps and dissociations in amygdala and OFC during extinction and conditioning. **(a,b)** Medial amygdala (ma) was commonly activated during extinction and conditioning (yellow), whereas lateral amygdala (la) and rostromedial OFC were selectively activated during extinction (red). Neural responses are superimposed on coronal **(a)** and horizontal **(b)** sections ( $P < 0.001$ ). **(c)** A plot of percent signal change (mean  $\pm$  s.e.m.) from medial amygdala shows that neural responses were evoked by target (Tgt) and nontarget (nTgt) CS+u during both extinction and conditioning (adjusted for CS- baselines). **(d,e)** In contrast, signal changes evoked by the CS+u in lateral amygdala and rostromedial OFC were preferentially enhanced during extinction.

Our next aim was to identify the neural substrates of this behavioral effect. We hypothesized that if representations of UCS value are accessible to the CS+ at extinction, then neural activity evoked by the target CS+u (compared to the nontarget CS+u) should be enhanced in brain regions that encode these 'inflated' representations. Consequently, the contrast [target CS+u - nontarget CS+u] at extinction (each minus their respective CS-) demonstrated a significant response in left lateral OFC ( $x = -39, y = 42, z = -18; Z = 3.24; P < 0.1$  SVC) (Fig. 5a,b). Group-averaged plots of OFC activity for each condition (adjusted for CS- baselines) showed that evoked activity was greater for the target CS+u than the nontarget CS+u (Fig. 5c), a response pattern that parallels behavioral inflation (CS+ aversiveness, Fig. 4a). These findings suggest that a predictive cue retains access to representations of reinforcer value in lateral OFC, even as extinction proceeds.

We further tested the sensitivity of CS+-evoked responses to UCS inflation within lateral OFC by examining the interaction between phase (extinction versus conditioning) and CS+ type (target versus nontarget). This revealed significant activity in an adjacent orbitofrontal region, albeit at a reduced threshold ( $-45, 42, -15; Z = 2.56; P = 0.005$  uncorrected). Condition-specific plots from this area indicate this interaction was largely driven by postinflation signal increases to target CS+u expressed during extinction

CS+. Consequently, we compared nontarget and target CS+u activity at extinction (minus CS- baselines) to index areas sensitive to relative decreases in aversive value. Compared to the response profile in lateral OFC, this contrast elicited an inverted pattern in ventromedial PFC ( $12, 54, -3; Z = 3.75; P < 0.05$  SVC) (Fig. 6a,b), such that neural activity was relatively enhanced to the nontarget CS+u (Fig. 6c). In the phase  $\times$  CS+ type interaction, significant ventromedial PFC activity ( $18, 51, 3; Z = 2.58; P < 0.005$  uncorrected) was principally driven

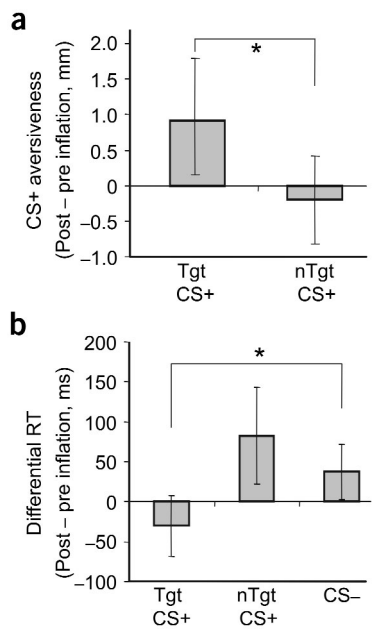
(Fig. 5d). To specify the content of these reinforcer representations more explicitly, we performed a correlation analysis of the data. Subject-specific estimates of neural activity, derived from the phase  $\times$  CS+ type interaction (as in Fig. 5d) were regressed upon ratings of relative CS+ aversiveness. This yielded a significant positive correlation in an adjacent portion of lateral OFC ( $-45, 45, -12; R = 0.60; P < 0.05$ ) (Fig. 5e), implying that relative magnitude of predictive (aversive) value is encoded, and updated, within this structure.

As UCS inflation enhanced target CS+ aversiveness, the nontarget CS+ concurrently became less aversive, relative to the target

**Table 2** Common and distinct responses evoked during extinction and conditioning

Brain region	MNI coordinates (mm)			Vol. (mm <sup>3</sup> )	Peak Z	P value
	x	y	z			
<b>Extinction + Conditioning</b>						
Left rostromedial OFC *	-18	48	-15	81	3.61	< 0.05 (SVC)
Left ventromedial PFC	-3	30	3	27	4.03	< 0.05 (SVC)
Right insula	39	15	6	243	3.79	< 0.05 (SVC)
Right ventral striatum	12	3	-6	27	3.30	< 0.05 (SVC)
Left ventral striatum	-15	0	-9	27	3.27	< 0.001
Right medial amygdala *	12	-6	-15	189	4.60	< 0.05 (SVC)
Left medial amygdala	-15	-9	-21	135	4.07	< 0.05 (SVC)
<b>Extinction - Conditioning</b>						
Right medial OFC/gyrus rectus	3	39	-21	54	3.28	< 0.1 (SVC)
Right caudal OFC	27	12	-18	27	3.24	< 0.001
Right lateral amygdala	33	-3	-27	108	3.66	< 0.05 (SVC)
Left lateral amygdala	-27	-9	-24	54	3.17	< 0.1 (SVC)
<b>Conditioning - Extinction</b>						
No significant activations						

SVC, corrected for small volume of interest. \*, also exhibited response habituation (condition  $\times$  time interaction).



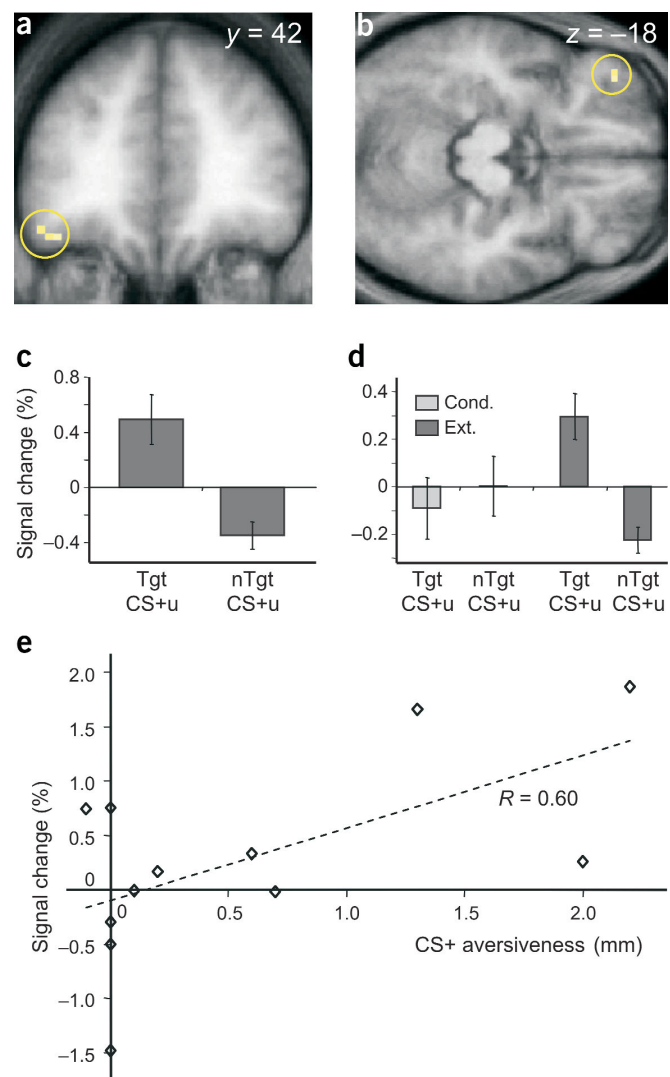
**Figure 4** Behavioral impact of UCS inflation on CS+ processing. (a) Postinflation (versus preinflation) ratings of the Tgt CS+ were significantly higher (more aversive) than ratings of the nTgt CS+ (mean  $\pm$  s.e.m., adjusted for CS- baselines). (b) Postinflation (versus preinflation) RTs were significantly faster for the Tgt CS+ compared to the CS-. The difference between nTgt CS+ and CS- RTs was not significant. \* $P < 0.05$ .

by the nontarget CS+u response at extinction (Fig. 6d). Finally, regression analysis demonstrated that neural responses in ventromedial PFC were significantly and negatively correlated with differential CS+ aversiveness (12, 57, -3;  $R = 0.63$ ;  $P < 0.05$ ) (Fig. 6e).

## DISCUSSION

Our initial objective was to characterize the neural substrates of extinction learning in the human brain. The results indicate that discrete regions of rostral and caudal OFC and lateral amygdala are preferentially activated during extinction. The findings cannot be attributed to general mechanisms of CS+ processing, as extinction-related activity was selectively enhanced in these areas over and above that evoked during conditioning. Furthermore, the effects were not due to nonspecific changes in sensory responsiveness, since these regions exhibited differential activation to the CS+ (versus CS-). On these grounds we propose that during extinction, CS+-evoked recruitment of an OFC-amygdala network provides the basis for memory-based processes that regulate expression of conditioning. We acknowledge that since our study focused on extinction learning, rather than recall<sup>8</sup>, it remains possible that long-term maintenance of extinction memories are supported elsewhere.

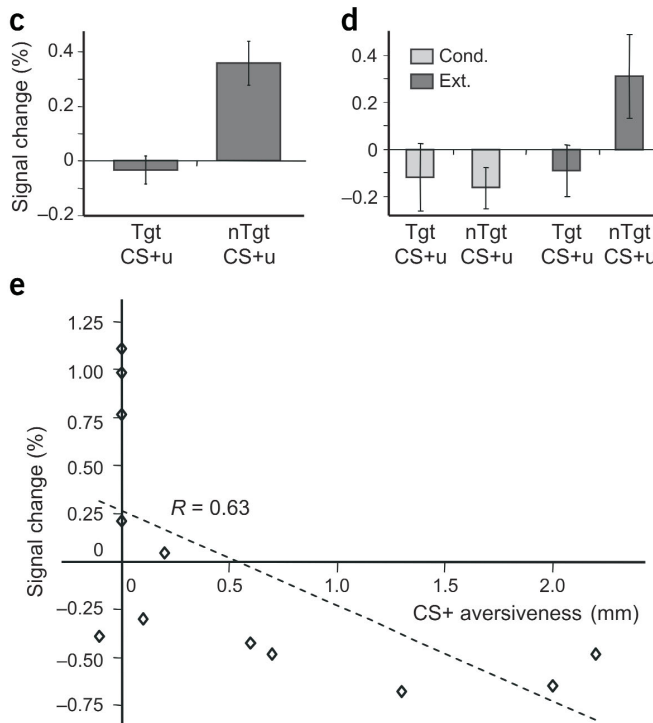
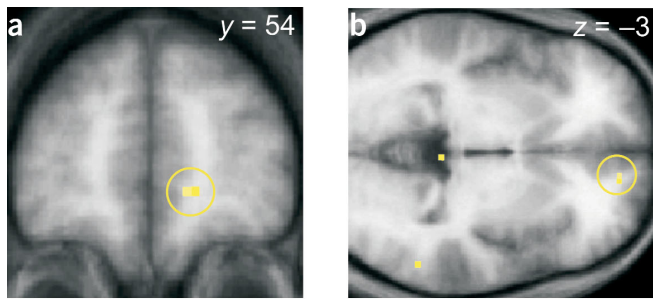
In rodent models of extinction, both ventromedial PFC<sup>22,24,25</sup> and amygdala<sup>11,21,23</sup> are involved in extinction-related processes. Recent studies propose that excitatory projections from medial PFC to interneurons in the lateral amygdaloid nucleus<sup>13,28</sup>, or to neighboring intercalated cell masses<sup>12,29</sup>, gate the flow of excitatory (conditioning-related) impulses into the central nucleus of the amygdala, diminishing the expression of conditioned responses. While fMRI obviously lacks the spatial refinement of cellular electrophysiology, it is worth speculating (with reference to the atlas of Mai *et al.*<sup>30</sup>) that the extinction-evoked activation in lateral amygdala spans the amygdaloid sub-



**Figure 5** A representation of (aversive) UCS value in lateral OFC during extinction. (a,b) Significant activity in left lateral OFC was detected in the comparison between Tgt CS+u and nTgt CS+u at extinction. Responses are superimposed on coronal (a) and horizontal (b) sections from the group-averaged T1 scan ( $P < 0.001$ ). (c) Activity plots in lateral OFC indicate that responses increased to the Tgt CS+u and decreased to the nTgt CS+u (mean  $\pm$  s.e.m.). (d) In the interaction between phase (extinction versus conditioning) and CS+ type (Tgt versus nTgt), signal variation in lateral OFC was mostly confined to increases elicited during the extinction period following inflation. (e) Subject-specific neural activity derived from the phase  $\times$  CS+ type interaction (as in d) was regressed upon subject ratings of differential CS+ aversiveness, highlighting a significant positive correlation in lateral OFC.

divisions implicated as potential PFC targets in animals, implying a potential cross-species preservation of function. Nonetheless, it is important to emphasize that even with optimized data acquisition and preprocessing, the spatial limitations of fMRI make such observations tentative at best.

In contrast to the selective effects of extinction, neural responses in distinct regions of rostral OFC and medial amygdala were common to conditioning and extinction. Notably, activity in these areas significantly decreased over time, indicating preferential responding early during acquisition and extinction, periods when attention, arousal



**Figure 6** A representation of (less aversive) UCS value in ventromedial PFC during extinction. (a–c) Responses in ventromedial PFC were identified in the opposite contrast of nTgt CS+u – Tgt CS+u at extinction. Neural activity is overlaid on coronal (a) and horizontal (b) sections ( $P < 0.001$ ) and plotted as percent signal change for each condition (c), showing that evoked responses were chiefly driven by increases to the nTgt CS+u. (d) Condition-specific plots of activity derived from the phase (extinction versus conditioning)  $\times$  CS+ type (nTgt versus Tgt) interaction demonstrate that the response in ventromedial PFC was preferentially evoked by the postinflation nTgt CS+u. (e) Regression analysis revealed a significant negative correlation between ventromedial PFC activity and CS+ aversiveness.

plausible that CS+ access to UCS representations is steadily maintained over the course of extinction, our findings do not rule out the possibility that CS+ activation of UCS representations is being progressively lost during extinction. Moreover, in the absence of a demonstrable overlap between the neural substrates of UCS value and aversive conditioning, we cannot conclude whether postinflation UCS representations in orbital prefrontal cortex are functionally related to regions that encoded the original UCS value.

The content of these UCS representations differed with respect to relative value and exhibited a double dissociation: increasing aversive value (target – nontarget) was selectively encoded within ventrolateral OFC, whereas decreasing aversive value (nontarget – target) was encoded within ventromedial PFC. Regression analyses indicated that both regions were sensitive to the relative magnitude of predictive value, but in opposite directions. While these correlations explain 35–40% of the between-subject variability (based on  $R^2$  values), it is evident that a considerable proportion of signal variance remains. One possible explanation may be that ratings of the CS+ faces were not collected during fMRI data acquisition. As such, these *post-hoc* regressors may not have adequately reflected the within-scanner experience, leading to imprecisions in capturing variance in the correlation analyses.

Our findings suggest that ventral PFC supports dual mnemonic representations of UCS value, both of which are accessible to a predictive cue. The presence of a dual representational system that responds as a function of the degree of preference (or nonpreference) could provide a basis for fine-tuned regulation over conditioned behavior and other learned responses. Indeed, an organism that needs to optimize its choices from among a set of different predictive cues (e.g., which to approach? which to avoid?) would be well served by a system that integrates information about their relative values in such a parallel and differentiated manner. The idea that orbitofrontal cortex synthesizes sensory, affective and motivational cues in the service of goal-directed behavior accords with animal models of associative learning<sup>34–36</sup> and recent human imaging studies of incentive states<sup>37</sup>, reinforcer devaluation<sup>38</sup> and valence-outcome contingencies<sup>39</sup>.

Previous work in our laboratory suggested that predictive reward value was collectively encoded in OFC and amygdala<sup>38</sup>, in keeping with animal lesion studies of reinforcer devaluation<sup>40–43</sup>. This conflicts with the present experiment, in which the CS+ gained access to UCS value representations in OFC, but not amygdala (even when examined at a liberal threshold of  $P < 0.01$ ). Differences in the conditioning mode (appetitive versus aversive), or in the underlying paradigm (satiety versus inflation), could account for the disparity. Alternatively, in the earlier study, explicit pairing between the CS+ and the newly devalued UCS (postconditioning) could have led to new associative learning between the CS+ and UCS, resulting in amygdala activation. The same criticism applies to a recent fMRI

and stimulus-reinforcer unpredictability are maximal. Any, or all, of these mechanisms promote associative learning<sup>31–33</sup> and may underpin the observed responses. We suggest the putative role of rostral OFC and medial amygdala in attention-based learning mechanisms is distinct from mechanisms of reinforcer representation and other aspects of extinction learning, as these latter processes had separate anatomical substrates. However, whether the role of these structures is confined to attentional processing, or supplements other facets of associative learning, remains unclear. Intriguingly, the behavioral impact of attention and associability on CS+ processing relies on the integrity of the central amygdaloid nucleus<sup>33</sup>, an area possibly homologous to the dorsomedial amygdala activation described here, though again, we caution against any strong claim relating fMRI activations to amygdala subregions.

We also sought to identify persisting neural signatures of CS:UCS links during extinction. By differentially altering UCS value during an inflation phase, we disambiguated CS+-evoked responses relating to UCS representations from those relating to extinction *per se*. The behavioral enhancement of target CS+ aversiveness provides compelling evidence that predictive cues retain access to updated memories of UCS value during extinction. The complementary imaging data show that representations of predictive value are maintained in distinct areas of ventral PFC, even as extinction proceeds. While it is

study of reversal learning suggesting that conditioned memories are maintained in the amygdala<sup>44</sup>. Here, new learning between the new CS+ (reversed CS-) and the UCS could have similarly driven amygdala responses. Our current study was protected against this confounding factor, as the CS+ and the UCS were never re-paired after conditioning. It is therefore tempting to consider that the amygdala is necessary for the formation of representations in OFC, without itself being a repository of those traces. This interpretation is supported by increasing evidence from rodent models of associative learning<sup>34,35,45</sup> that emphasize subtle but distinct roles for amygdala (acquisition) and OFC (maintenance) in the encoding of links between predictive cues and outcome values.

The idea that extinction is a form of new learning, yet leaves intact associations originally established during conditioning, receives strong support from animal studies. Our current data show that extinction learning (compared to conditioning) in humans relies on partially independent neural systems with close structural homology to animal models. Areas in rostromedial OFC and medial amygdala exhibited overlapping activity during both conditioning and extinction sessions, whereas separate regions in OFC and lateral amygdala participated selectively in extinction. These latter sites mediate formation of 'CS:no UCS' memories that oppose the expression of conditioning. In turn, the identification of 'CS:UCS' representations in distinct regions of ventral PFC during extinction lends biological credence to the idea that associative links are not simply erased with behavioral extinction. An important focus of future investigations will be to determine mechanisms of selection by which either of these competing memories ultimately exerts control over behavior.

## METHODS

**Subjects.** Informed consent was obtained from 18 healthy subjects (10 women; mean age, 24 years; range, 18–37 years). The study was conducted with local ethics approval of the Institute of Neurology & National Hospital of Neurology and Neurosurgery. Two subjects (1 woman, 1 man) were eliminated from further analysis due to technical malfunctions during scanning ( $n = 1$ ) or difficulty detecting the odors ( $n = 1$ ), leaving 16 subjects whose data were analyzed.

**Stimuli.** Four neutrally valenced faces (2 male, 2 female) comprised the 2 CS+ and 2 CS- stimuli. These were modified from a standardized set<sup>46</sup> by cropping out facial hair to increase difficulty on a gender discrimination task. The images (308 × 225 pixels) were back-projected onto a headbox mirror inside the scanner. Two unpleasant odors ('rotten eggs': ammonium sulfide [AS], 0.016% v/v in distilled water; and 'sweaty socks': 4-methylpentanoic acid [4MP], 1% v/v in mineral oil; Sigma-Aldrich) comprised the aversive UCS stimuli. Odors were delivered using a computer-controlled olfactometer that presents discrete odor pulses with rapid on-off times and is suitable for the MRI environment<sup>47</sup>. Stimulus presentation was controlled using Cogent 2000 (Wellcome Department, London, UK), implemented in Matlab 6 (Mathworks Inc.).

**Procedure.** Subjects participated in an olfactory version of classical (Pavlovian) conditioning, whereby the CS+ faces were repetitively paired with the two different UCS odors (Fig. 1a). One odor, destined for UCS inflation, was designated the target UCS. The other odor underwent no incentive manipulation (nontarget UCS). One-half of all CS+ presentations was coupled to the UCS, resulting in paired (CS+p) and unpaired (CS+u) conditions. Importantly, the use of unpaired trial types allowed us to investigate CS+ processing in the absence of odor UCS confounds. Two additional faces, never paired with odor, served as nonconditioned controls (CS-<sub>1</sub> and CS-<sub>2</sub>). Immediately after conditioning, subjects underwent UCS inflation, whereby the target UCS (at increased odor intensity: either 0.2% AS or 20% 4MP) and nontarget UCS (at baseline intensity) were presented in the absence of the

CS+ (Fig. 1b). The aim of this session was to selectively enhance or 'inflate' target UCS aversiveness. The impact of this manipulation was assessed in a final extinction session, whereby the target CS+ and nontarget CS+ were repeatedly delivered in the absence of odor UCS (Fig. 1c). Assignments of odors (as target and nontarget UCS) and faces (as CS+ and CS-) were counterbalanced across subjects.

**Experiment.** Conditioning and extinction trials were identical in design (Fig. 1a,c). These began with presentation of a face ( $t = 0$ ), which appeared on-screen for 750 ms. A red crosshair ( $t = 500$  ms) then signaled subjects to sniff for 750 ms. In paired CS:UCS trials (conditioning only), odor delivery coincided with appearance of the crosshair, ensuring a 250-ms overlap between CS+ and UCS. Trials recurred every 7.5 s. Subjects sniffed on every trial, regardless of odor delivery, which balanced this factor across conditions. Importantly, subjects were never informed about CS:UCS contingencies (or their disruption), to minimize the influence of expectancy and other cognitive factors on the activation patterns. Instead, they were asked to indicate facial gender by pressing a button as quickly and accurately as possible. This also ensured task constancy across conditioning and extinction sessions. During conditioning (13.5 min), there were 14 presentations each of target CS+p, target CS+u, nontarget CS+p and nontarget CS+u, and 26 presentations each of CS-<sub>1</sub> and CS-<sub>2</sub>, so that roughly equal numbers of faces were viewed. The use of different CS- faces provided orthogonal baselines for conjunction analysis. Extinction (7 min) consisted of 14 presentations each of target CS+u, nontarget CS+u, CS-<sub>1</sub> and CS-<sub>2</sub>.

Interposed between conditioning and extinction was UCS inflation (3.5 min) (Fig. 1b). At the start, subjects were instructed that only sniff cues and odors would be delivered. There were seven trials each of the inflated (target) and non-inflated (nontarget) UCS. Inclusion of the nontarget odor controlled for mere effects of UCS exposure, and 14 sniff-only (odor-free) trials were included to minimize sensory habituation. On each trial, subjects pressed one of two buttons to indicate odor presence or absence. As above, crosshair cues and odors were delivered for 750 ms, and trials recurred every 7.5 s. Stimulus presentation was randomized across all phases.

**Behavioral measurements.** Subject-specific reaction times (RTs) were log-transformed to ensure a normal distribution. Median RTs were determined for each event type, then group-averaged. Online respirations were monitored using thoracic and abdominal breathing belts<sup>47</sup>. Condition-specific sniff amplitudes and latencies (times to peak) were calculated for each subject and then group-averaged. We also examined the effect of sniffing on head movement during scanning, by calculating event-related (sniff-related) movements from each subject's movement parameters (obtained from spatial realignment during image preprocessing). This indicated that the mean three-dimensional translation across subjects was 0.10 mm ( $\pm 0.018$  mm; s.e.m.), which represented a very small fraction of the actual in-plane scanner resolution (approximately 3%, based on a 3-mm voxel size).

*Post-hoc* odor ratings were obtained using a 15-mm analog scale, with anchors labeled 'undetectable' and 'very strong' (intensity), or 'very pleasant' and 'very unpleasant' (valence), with midpoints marked 'medium' (intensity) or 'neutral' (valence). Preinflation and postinflation valence ratings of the faces were also collected using a 15-mm scale (anchors very pleasant and very unpleasant). Relative CS+ aversiveness, as a behavioral index of inflation, was calculated as follows: postinflation [(target CS+ - CS-<sub>1</sub>) - (nontarget CS+ - CS-<sub>2</sub>)] minus preinflation [(target CS+ - CS-<sub>1</sub>) - (nontarget CS+ - CS-<sub>2</sub>)]. In this way, positive values reflect greater aversiveness to the target CS+ face at extinction (versus conditioning), over and above nonspecific effects on the nontarget CS+ or CS- faces. Those subjects ( $n = 3/16$ ) who showed poor behavioral evidence for inflation (*i.e.*, values below -0.1) were eliminated from the analysis of UCS inflation (though all 16 subjects were still included in the conditioning and extinction analyses).

**Imaging analysis.** Gradient-echo T2\*-weighted echoplanar images (EPI) were collected on a Siemens Vision 2T scanner, using a specialized sequence to reduce signal dropout in orbitofrontal lobes<sup>48</sup>. Imaging parameters were as follows: TE, 35 ms; TR, 2.31 s; in-plane resolution, 3.0 mm; field-of-view, 192 mm; slice thickness, 1.8 mm; gap, 1.2 mm. We collected 644 volumes (33

slices/volume, providing 80% whole-brain coverage) per subject, plus 6 volumes to permit T1 relaxation. Image preprocessing, including spatial realignment, slice-time correction, normalization and smoothing (6-mm kernel) was conducted in SPM2 (Wellcome Department of Imaging Neuroscience, London, UK). High-resolution T1-weighted anatomical images were normalized to each subject's mean EPI and then group-averaged ( $n = 14$ ).

The event-related fMRI data were analyzed in SPM2 using the general linear model<sup>49</sup>. Subject-specific regressors of interest were assembled by convolving  $\delta$  functions (corresponding to the onset times for each condition) with a canonical hemodynamic response function (HRF) and its temporal and dispersion derivatives. Condition  $\times$  time interactions were modeled by multiplying the regressors with exponential decay functions (time-constant, one quarter of session length) to estimate response plasticity over time<sup>19</sup>. Regressors of no interest included a high-pass filter (1/128 Hz) and six movement parameters. Auto-correlation was adjusted using an AR(1) model. Parameter estimates pertaining to the height of the HRF for each regressor were calculated for each voxel. Subject-specific contrasts were computed, and then entered into a second-level (random-effects) model in SPM2 via one-sample  $t$ -tests or ANOVAs (with sphericity correction) for conjunction analysis.

We tested five principal contrasts: (1) conditioning conjunction of (target CS+u – CS–<sub>1</sub>) + (nontarget CS+u – CS–<sub>2</sub>); (2) extinction conjunction of (target CS+u – CS–<sub>1</sub>) + (nontarget CS+u – CS–<sub>2</sub>); (3) conditioning and extinction conjunction of (1) and (2) above; (4) extinction versus conditioning: (2) minus (1) above; and (5) inflation effect: (target CS+u – CS–<sub>1</sub>) – (nontarget CS+u – CS–<sub>2</sub>) at extinction. The use of conjunctions guards against the possibility that interactions between the target CS+, the nontarget CS+ and the two CS– types might otherwise underlie the observed activations, and ensures both the target and nontarget CS+ are significantly activated above their respective CS– baselines<sup>50</sup>. Moreover, at extinction, as the two CS+ stimuli were not equivalent (target inflated, nontarget noninflated), conjunction analysis provided a means of identifying shared effects (namely, extinction).

We report significant activations at  $P < 0.05$  (trend,  $P < 0.1$ ) in *a priori* regions surviving correction for multiple comparisons across small volumes of interest<sup>51</sup>. These regions comprise structures implicated in conditioning studies from our laboratory, including amygdala, OFC, ventromedial PFC, insula, ventral striatum and ventral midbrain<sup>3,4,18,38,44</sup>. Small-volume corrections were based on peak activation coordinates derived from these studies, to limit the effective search space. For descriptive purposes, we also report activations in *a priori* areas surviving an uncorrected threshold of  $P < 0.001$ . Extent (volume) of activation is based on the number of within-cluster voxels observed at  $P < 0.001$  (uncorrected) and voxel volume of 3 mm<sup>3</sup>.

Finally, correlation analysis was performed in SPM2 by regressing subject-specific neural activity, derived from a phase (extinction versus conditioning)  $\times$  CS+ type (target versus nontarget) interaction, upon ratings of differential CS+ aversiveness (excluding one outlying subject whose rating exceeded 3 standard deviations from the group mean). We report significant correlations at  $P < 0.05$  in PFC regions identified in the principal inflation contrasts. Reported voxel locations conform to Montreal Neurological Institute (MNI) coordinate space. For display, the right side of the image corresponds to the right side of the brain.

#### ACKNOWLEDGMENTS

This work was supported by a Howard Hughes Physician-Scientist Fellowship Grant (J.A.G.) and a Wellcome Trust Programme Grant (R.J.D.). We thank J.S. Winston and J.M. Kilner for helpful discussions.

#### COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Received 4 May; accepted 3 August 2004

Published online at <http://www.nature.com/natureneuroscience/>

- LeDoux, J.E. Emotion circuits in the brain. *Annu. Rev. Neurosci.* **23**, 155–184 (2000).
- Maren, S. Neurobiology of Pavlovian fear conditioning. *Annu. Rev. Neurosci.* **24**, 897–931 (2001).
- Gottfried, J.A., O'Doherty, J. & Dolan, R.J. Appetitive and aversive olfactory learning in humans studied using event-related functional magnetic resonance imaging. *J. Neurosci.* **22**, 10829–10837 (2002).

- O'Doherty, J., Deichmann, R., Critchley, H.D. & Dolan, R.J. Neural responses during anticipation of a primary taste reward. *Neuron* **33**, 815–826 (2002).
- Morrow, B.A., Elsworth, J.D., Rasmussen, A.M. & Roth, R.H. The role of mesoprefrontal dopamine neurons in the acquisition and expression of conditioned fear in the rat. *Neuroscience* **92**, 553–564 (1999).
- Baeg, E.H. *et al.* Fast spiking and regular spiking neural correlates of fear conditioning in the medial prefrontal cortex of the rat. *Cereb. Cortex* **11**, 441–451 (2001).
- Rescorla, R.A. Experimental extinction, in *Handbook of Contemporary Learning Theories* (eds. Mowrer, R.R. & Klein, S.) 119–154 (Lawrence Erlbaum, Mahwah, New Jersey, 2001).
- Myers, K.M. & Davis, M. Behavioral and neural analysis of extinction. *Neuron* **36**, 567–584 (2002).
- Bouton, M.E. Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychol. Bull.* **114**, 80–99 (1993).
- Garcia, R. Postextinction of conditioned fear: between two CS-related memories. *Learn. Mem.* **9**, 361–363 (2002).
- Hobin, J.A., Goossens, K.A. & Maren, S. Context-dependent neuronal activity in the lateral amygdala represents fear memories after extinction. *J. Neurosci.* **23**, 8410–8416 (2003).
- Quirk, G.J., Likhtik, E., Pelletier, J.G. & Pare, D. Stimulation of medial prefrontal cortex decreases the responsiveness of central amygdala output neurons. *J. Neurosci.* **23**, 8800–8807 (2003).
- Rosenkranz, J.A., Moore, H. & Grace, A.A. The prefrontal cortex regulates lateral amygdala neuronal plasticity and responses to previously conditioned stimuli. *J. Neurosci.* **23**, 11054–11064 (2003).
- Quirk, G.J. & Gehlert, D.R. Inhibition of the amygdala: key to pathological states? *Ann. NY Acad. Sci.* **985**, 263–272 (2003).
- LaBar, K.S., Gatenby, J.C., Gore, J.C., LeDoux, J.E. & Phelps, E.A. Human amygdala activation during conditioned fear acquisition and extinction: a mixed-trial fMRI study. *Neuron* **20**, 937–945 (1998).
- Rescorla, R.A. Effect of inflation of the unconditioned stimulus value following conditioning. *J. Comp. Physiol. Psychol.* **86**, 101–106 (1974).
- Bouton, M.E. Differential control by context in the inflation and reinstatement paradigms. *J. Exp. Psychol. Animal Behav. Proc.* **10**, 56–74 (1984).
- Morris, J.S., Ohman, A. & Dolan, R.J. Conscious and unconscious emotional learning in the human amygdala. *Nature* **393**, 467–470 (1998).
- Buchel, C., Morris, J., Dolan, R.J. & Friston, K.J. Brain systems mediating aversive conditioning: an event-related fMRI study. *Neuron* **20**, 947–957 (1998).
- Critchley, H.D., Mathias, C.J. & Dolan, R.J. Fear conditioning in humans: the influence of awareness and autonomic arousal on functional neuroanatomy. *Neuron* **33**, 653–663 (2002).
- Falls, W.A., Miserendino, M.J.D. & Davis, M. Extinction of fear-potentiated startle: blockade by infusion of an NMDA antagonist into the amygdala. *J. Neurosci.* **12**, 854–863 (1992).
- Morgan, M.A., Romanski, L.M. & LeDoux, J.E. Extinction of emotional learning: contribution of medial prefrontal cortex. *Neurosci. Lett.* **163**, 109–113 (1993).
- Repa, J.C. *et al.* Two different lateral amygdala cell populations contribute to the initiation and storage of memory. *Nat. Neurosci.* **4**, 724–731 (2001).
- Herry, C. & Garcia, R. Prefrontal cortex long-term potentiation, but not long-term depression, is associated with the maintenance of extinction of learned fear in mice. *J. Neurosci.* **22**, 577–583 (2002).
- Milad, M.R. & Quirk, G.J. Neurons in medial prefrontal cortex signal memory for fear extinction. *Nature* **420**, 70–74 (2002).
- White, K. & Davey, G.C.L. Sensory preconditioning and UCS inflation in human 'fear' conditioning. *Behav. Res. Ther.* **27**, 161–166 (1989).
- Hosoba, T., Iwanaga, M. & Seiwa, H. The effect of UCS inflation and deflation procedures on 'fear' conditioning. *Behav. Res. Ther.* **39**, 465–475 (2001).
- Rosenkranz, J.A. & Grace, A.A. Cellular mechanisms of infralimbic and prelimbic prefrontal cortical inhibition and dopaminergic modulation of basolateral amygdala neurons *in vivo*. *J. Neurosci.* **22**, 324–337 (2002).
- Royer, S. & Pare, D. Bidirectional synaptic plasticity in intercalated amygdala neurons and the extinction of conditioned fear responses. *Neuroscience* **115**, 455–462 (2002).
- Mai, J.K., Assheuer, J. & Paxinos, G. *Atlas of the Human Brain* (Academic Press, San Diego, California, 1997).
- Pearce, J.M. & Hall, G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552 (1980).
- Wilson, P.N., Boumphrey, P. & Pearce, J.M. Restoration of the orienting response to a light by a change in its predictive accuracy. *Q. J. Exp. Psychol.* **44B**, 17–36 (1992).
- Holland, P.C. & Gallagher, M. Amygdala circuitry in attentional and representational processes. *Trends Cogn. Sci.* **3**, 65–73 (1999).
- Schoenbaum, G., Setlow, B., Saddoris, M.P. & Gallagher, M. Encoding predicted outcome and acquired value in orbitofrontal cortex during cue sampling depends upon input from basolateral amygdala. *Neuron* **39**, 855–867 (2003).
- Pickens, C.L. *et al.* Different roles for orbitofrontal cortex and basolateral amygdala in a reinforcer devaluation task. *J. Neurosci.* **23**, 11078–11084 (2003).
- Tremblay, L. & Schultz, W. Relative reward preference in primate orbitofrontal cortex. *Nature* **398**, 704–708 (1999).
- Arana, F.S., Parkinson, J.A., Hinton, E., Holland, A.J. & Owen, A.M. Dissociable contributions of the human amygdala and orbitofrontal cortex to incentive motivation and goal selection. *J. Neurosci.* **23**, 9632–9638 (2003).

38. Gottfried, J.A., O'Doherty, J. & Dolan, R.J. Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science* **301**, 1104–1107 (2003).
39. O'Doherty, J., Critchley, H., Deichmann, R. & Dolan, R.J. Dissociating valence of outcome from behavioral control in human orbital and ventral prefrontal cortices. *J. Neurosci.* **23**, 7931–7939 (2003).
40. Hatfield, T., Han, J.S., Conley, M., Gallagher, M. & Holland, P. Neurotoxic lesions of basolateral, but not central, amygdala interfere with Pavlovian second-order conditioning and reinforcer devaluation effects. *J. Neurosci.* **16**, 5256–5265 (1996).
41. Malkova, L., Gaffan, D. & Murray, E.A. Excitotoxic lesions of the amygdala fail to produce impairment in visual learning for auditory secondary reinforcement but interfere with reinforcer devaluation effects in rhesus monkeys. *J. Neurosci.* **17**, 6011–6020 (1997).
42. Gallagher, M., McMahan, R.W. & Schoenbaum, G. Orbitofrontal cortex and representation of incentive value in associative learning. *J. Neurosci.* **19**, 6610–6614 (1999).
43. Baxter, M.G., Parker, A., Lindner, C.C., Izquierdo, A.D. & Murray, E.A. Control of response selection by reinforcer value requires interaction of amygdala and orbital prefrontal cortex. *J. Neurosci.* **20**, 4311–4319 (2000).
44. Morris, J.S. & Dolan, R.J. Dissociable amygdala and orbitofrontal responses during reversal fear conditioning. *Neuroimage* **22**, 372–380 (2004).
45. Schoenbaum, G., Chiba, A.A. & Gallagher, M. Changes in functional connectivity in orbitofrontal cortex and basolateral amygdala during olfactory discrimination learning. *J. Neurosci.* **20**, 5179–5189 (2000).
46. Ekman, P. & Friesen, W.V. *Pictures of Facial Affect* (Consulting Psychologists Press, Palo Alto, California, USA, 1976).
47. Gottfried, J.A., Deichmann, R., Winston, J.S. & Dolan, R.J. Functional heterogeneity in human olfactory cortex: an event-related functional magnetic resonance imaging study. *J. Neurosci.* **22**, 10819–10828 (2002).
48. Deichmann, R., Gottfried, J.A., Hutton, C. & Turner, R. Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage* **19**, 430–441 (2003).
49. Friston, K.J. *et al.* Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1995).
50. Price, C.J. & Friston, K.J. Cognitive conjunction: a new approach to brain activation experiments. *Neuroimage* **5**, 261–270 (1997).
51. Worsley, K.J. *et al.* A unified statistical approach for determining significant voxels in images of cerebral activation. *Hum. Brain Mapp.* **4**, 58–73 (1996).