OXFORD

Social Cognitive and Affective Neuroscience, 2016, 1363–1373

doi: 10.1093/scan/nsw053
Advance Access Publication Date: 22 April 2016
Original article

Racial stereotypes impair flexibility of emotional learning

Joseph E. Dunsmoor,[1] Jennifer T. Kubota,[2,3] Jian Li,[4,5] Cesar A.O. Coelho,[6] and Elizabeth A. Phelps[1,7]

[1]Department of Psychology and Center for Neural Sciences, New York University, New York, NY, 10003, USA, [2]Department of Psychology, University of Chicago, Chicago, IL, USA, 60637, [3]Center for the Study of Race, Politics and Culture, University of Chicago, [4]Department of Psychology and Beijing Key Laboratory of Behavior and Mental Health, [5]PKU-IDG/McGovern Institute for Brain Research, Peking University, [6]Departamento de Psicobiologia, Universidade Federal de São Paulo, São Paulo, SĐo Paulo 04023062, Brazil, and [7]Emotional Brain Institute, Nathan Kline Institute for Psychiatric Research, Orangeburg, New York 10962, USA

Correspondence should be addressed to Joseph Dunsmoor, Department of Psychology, 6 Washington Place Room 890, New York University, New York, NY 10003 USA. E-mail: joseph.dunsmoor@nyu.edu.

Abstract

Flexibility of associative learning can be revealed by establishing and then reversing cue-outcome discriminations. Here, we used functional MRI to examine whether neurobehavioral correlates of reversal-learning are impaired in White and Asian volunteers when initial learning involves fear-conditioning to a racial out-group. For one group, the picture of a Black male was initially paired with shock (threat) and a White male was unpaired (safe). For another group, the White male was a threat and the Black male was safe. These associations reversed midway through the task. Both groups initially discriminated threat from safety, as expressed through skin conductance responses (SCR) and activity in the insula, thalamus, midbrain and striatum. After reversal, the group initially conditioned to a Black male exhibited impaired reversal of SCRs to the new threat stimulus (White male), and impaired reversals in the striatum, anterior cingulate cortex, midbrain and thalamus. In contrast, the group initially conditioned to a White male showed successful reversal of SCRs and successful reversal in these brain regions toward the new threat. These findings provide new evidence that an aversive experience with a racial out-group member impairs the ability to flexibly and appropriately adjust fear expression towards a new threat in the environment.

Key words: Pavlovian fear conditioning; associative learning; racial attitudes and relations; stereotyping and prejudice; extinction

Introduction

Racially based threat stereotypes can have a profound impact on intergroup dynamics, leading to negative judgments and aggressive or defensive behavior towards members of a racial out-group. Behavioral research repeatedly demonstrates that Black men are stereotyped as dangerous, criminal and violent, both implicitly (Payne, 2001; Nosek *et al.*, 2002) and explicitly (Devine, 1989). These threat appraisals can predict discriminatory behaviors, such as decisions to 'shoot' unarmed Black individuals (Correll *et al.*, 2002), prime detection (and misperception) of weapons (Payne, 2001; Kubota and Ito, 2014), and bias identifying Black individuals as aggressors (Sagar and Schofield, 1980; Eberhardt *et al.*, 2004). In order to understand how to control and change negative racial biases, it is important to examine conditions that engender behavioral and cognitive flexibility. Here, we investigated whether stereotypic threat associations challenge flexibility of emotional learning when, following an aversive experience with a Black male, the Black male becomes safe and a White male becomes a new threat, i.e. reversal-learning.

Received: 12 October 2015; **Revised:** 21 March 2016; **Accepted:** 18 April 2016

One of the most straightforward techniques to examine behavioral flexibility is by testing whether an individual can update prior learning once a previously learned association changes. In reversal-learning, for instance, individuals first learn to discriminate between a reinforced stimulus and another irrelevant (unreinforced) stimulus. This association then reverses, such that the reinforced stimulus becomes irrelevant and the previously irrelevant stimulus is now reinforced. By changing stimulus-outcome associations after an initial learning experience, experimenters have been able to assess a number of factors that promote or impair behavioral and cognitive flexibility (Dias *et al.*, 1996; Bari and Robbins, 2013). Neuroimaging studies in humans implicate the striatum, midbrain, and prefrontal cortex as key neurocircuitry involved in readjusting behavior as cue-outcome associations change (Cools *et al.*, 2002; Remijnse *et al.*, 2005). Such findings accord with a body of neurophysiological research in animals showing that damage to the prefrontal cortex and striatum impairs reversal behavior (reviewed in Kehagia *et al.*, 2010; Bari and Robbins, 2013)

By combining reversal-learning with Pavlovian fear-conditioning, experimenters have been able to characterize processes related to updating threat associations in a dynamic environment, where the meaning of a danger cue can change over time (Morris and Dolan, 2004; Schiller *et al.*, 2008; Li *et al.*, 2011; Boll *et al.*, 2013). In fear-reversal tasks, a conditioned stimulus (CS1; e.g. the image of a face) initially predicts an aversive unconditioned stimulus (US; e.g. an electric shock), while a second stimulus (CS2; e.g. another face) predicts no US. This initial CS–US association establishes CS1 as a reliable indicator of impending threat, leading to a conditioned response (CR) such as increases in sympathetic arousal, including sweating, heart rate and respiration. The association between the CS and US then changes. More precisely, CS2 unexpectedly predicts the US and CS1 unexpectedly stops predicting the US. Neuroimaging research on fear-reversal shows that, upon reversal, the CR and neural activity in the amygdala and striatum flexibly shifts to the current danger cue (Schiller *et al.*, 2008).

Pavlovian fear-conditioning provides a valuable model from which to understand racial biases in general, and threat stereotypes in particular (Öhman, 2005; Amodio and Devine, 2006). In this framework, associative learning mechanisms may give rise to, or intensify preexisting implicit racial biases that are resistant to changes in behavior. Acquired race biases may then give rise to prejudicial or avoidance behavior towards racial out-group members (Lindström *et al.*, 2015). Olsson *et al.* (2005) provided evidence of intensified fear-learning to a racial out-group member using a Pavlovian fear-conditioning and extinction task. In fear-extinction, the CR gradually diminishes when the aversive US is omitted altogether. Olsson *et al.*, (2005) showed that skin conductance responses (SCR), a reflection of sympathetic arousal, persisted to a racial out-group CS as compared to a racial in-group CS throughout extinction when the US was omitted. Critically, intensified fear-learning was not expressed at the time of Pavlovian conditioning itself, when the CS–US association was initially learned; that is, individuals showed equivalent fear expression to the White CS and Black CS paired with shock. Instead, the effect of selectively intense fear-learning to a race out-group was only revealed in the time it took to extinguish SCRs, a finding in keeping with prior fear-learning studies using intrinsically threat-relevant CSs like snakes and spiders (McNally, 1987; Öhman and Mineka, 2001). In this way, deficits in behavioral flexibility are not revealed at the time of learning, but rather when a stimulus takes on ambiguous properties, which occurs when the initial stimulus-outcome association is in need of updating (Bouton, 2002).

Notably, delays in the time to extinguish fear responses to threat-relevant stimuli can be described by non-associative factors (e.g. sensitization or expectancy biases; Davey, 1992; McNally, 2015). Slower extinction rates to a racial out-group, specifically, can be described by statistical learning theory models that predict superior conditioning to novel or unfamiliar conditioned stimuli (Dayan *et al.*, 2000; Courville *et al.*, 2006; Maia, 2009). For instance, people are more likely to form an association between an unfamiliar CS and US (learning is faster and stronger), whereas prior experience/familiarity with a CS in the absence of the US interferes with learning a new CS–US association (learning is slower and weaker) (Lubow, 1973). Indeed, Olsson *et al.* (2005) found that controlling for self-reported out-group contact (analogous to familiarity) removed extinction biases to a race out-group. In sum, there are multiple mechanisms that may explain decreased flexibility of associative learning to racial out-groups.

Reversal-learning has proved a remarkably suitable technique, in a variety of forms, to characterize and understand behavioral and cognitive flexibility across a number of species and human populations (Bari and Robbins, 2013 for a review), and may be especially valuable to examine potential deficits in the ability to update racially based threat stereotypes. For instance, unlike fear-extinction, fear-reversals initiate two simultaneous learning processes: learning that CS1 is now safe whilst simultaneously learning to fear CS2. In this way, fear-reversal is behaviorally more demanding than fear-extinction, since danger is still in the environment but what signals danger has dynamically shifted (Schiller and Delgado, 2010). Reversal-learning tasks also provide an additional marker of inflexible associative learning beyond extinction; specifically, inflexible learning could be due to an inability to cease responding to the original CS (i.e., an extinction-deficit), an inability to respond to the new CS, or a combination thereof. A novel prediction from fear reversal-learning is that a threat stereotype, reinforced by a negative experience, will impair the ability to learn about a new threat in the environment. In other words, if initial learning already confirmed an implicit threat stereotype, then subjects may have trouble learning a new CS–US association in the presence of the previously reinforced CS. Such an impairment would be in line with behavioral research finding pervasive implicit negative, threat-related associations with Black individuals (Payne, 2001; Greenwald *et al.*, 2009) in comparison to objectively similarly threatening and dangerous White individuals presented around the same time (Correll *et al.*, 2002; Kubota and Ito, 2014).

In this study, we examined flexibility of racially-based threat associations using a combination of psychophysiology and fMRI. In a between-subjects design, White and Asian participants first learn that the image of a male (CS1) is dangerous and predicts an aversive electrical shock, and the image of another male (CS2) is safe. For one group, CS1 is Black and CS2 is White, hereafter referred to as the Black-Shocks-First (BSF) group. Initial learning in this group therefore involves a stereotypic threat association (Payne, 2001; Quillian and Pager, 2001; Correll *et al.*, 2002; Plant and Peruche, 2005). For the second group, CS1 is White and the CS2 is Black, hereafter referred to as the White-Shocks-First (WSF) group. Initial learning in the WSF group, in contrast, involves a counterstereotypic threat association (i.e., the White male is dangerous but the Black male is safe). Midway through the experiment, the contingencies unexpectedly reverse, and now CS1 is safe and the CS2 is dangerous.

In line with prior studies showing that fear-learning to racial out-groups resists extinction (Olsson *et al.*, 2005; Golkar *et al.*, 2015), we predicted persistent fear responses to the CS1 after reversal in the BSF group, but not in the WSF group. Further, the use of reversal-learning affords an additional novel prediction that the ability to shift fear responses from CS1 to CS2 will be impaired in the BSF group, but not in the WSF group. That is, once a stereotypic threat association is confirmed, we predicted that participants in the BSF group will have difficulty with the appropriate expression of fear when the US dynamically shifts between cues. Further, we predicted that activity patterns in regions commonly identified in fMRI studies of human fear-learning (Fullana *et al.*, 2015) would fail to shift towards the White male following reversal in the BSF group. This includes the amygdala, striatum, and regions of the 'central autonomic-interoceptive network' consisting of the insula, dorsal midbrain (periaqueductal gray), anterior cingulate cortex (ACC), and thalamus.

Method

Participants

Forty-one healthy right-handed adult White and Asian volunteers (mean age = 22.8; SD = 4.16) were assigned to the BSF group (N = 20, 13 females, 4 Asian) or WSF group (N = 21, 14 females, 4 Asian).[1] An additional subject was excluded due to excessive head motion. Sample size was based on previous human fear-learning research using race categories (Olsson *et al.*, 2005), and between-subjects human fear-learning research using fMRI (Dunsmoor *et al.*, 2014). All volunteers were pre-screened prior to the fMRI session to assess MRI eligibility. All subjects provided written informed consent and were paid for their participation. The experiment was approved by the University Committee on Activities Involving Human Subjects at New York University.

Implicit race bias

Implicit racial bias was assessed with the IAT (Greenwald *et al.*, 1998) during the prescreen. IAT D scores, a measure of the difference in mean response latencies between prejudice congruent (e.g., Black/Bad are paired) and prejudice incongruent trials (e.g., Black/Good are paired) divided by the pooled standard deviation of all response latencies, were calculated using the procedures recommended by Greenwald *et al.* (2003). The IAT generates a D score ranging from -2 to +2, with higher scores indicating a pro-White bias (Greenwald *et al.*, 2003). Importantly, the goal of the present study was to examine the role of implicit racial bias on fear-learning and reversal in individuals along a continuum of pro-White implicit associations, and no strong prediction was made regarding individual differences in IAT in relation to the present findings. To ensure that participants fell along a continuum of pro-White implicit associations, participants with a D score below 0.2 at prescreen were not scheduled for the fMRI experiment. Of the 55 participants pre-screened for this study, only 7 had a D score below 0.2. Mean IAT scores for BSF (Mean ± SD: 0.54 ± 0.27) and WSF group (0.55 ± 0.26) were similar.

Out-group contact

Intergroup contact was measured after the fMRI session by asking participants to estimate from 0 to 100: (i) approximately what percentage of all your close friends are Black (White)?; (ii) approximately what percentage of all your acquaintances are Black (White)?; (iii) approximately what percentage of the people you encounter on a day-to-day basis, in your neighborhood and at work or school, are Black (White)?; (iv) approximately what percentage of the people you see in the media, including television, movies, magazines, and sports, are Black (White)? We constructed intergroup contact difference scores by averaging the percentages for Black and White contact separately for each subject and then subtracting the percentage of Black contact from White contact. Out-group contact data from two subjects from the BSF group were lost due to computer error. Mean contact scores for BSF (mean ± SD: 43.72 ± 20.60) and WSF (46.04 ± 19.80) were similar.[2]

Conditioned stimuli

The CSs included photographs of two male faces from the Eberhardt Face Database series (Eberhardt *et al.*, 2006; Goff *et al.*, 2008). Eberhardt and colleagues have previously piloted these faces on prototypically (1–7 scale), attractiveness (1–7 scale), and age. Photographs were 327 × 450 pixels and displayed the actor on a gray background from the neck up in a frontal orientation. The study included three counterbalanced pairs of stimuli (a Black CS and a White CS), and each pair was equated for attractiveness, selected to be younger than 45, and selected to be prototypical of Black and White faces.

Fear-learning and reversal paradigm

The fear-learning and reversal task occurred in one continuous scanning run that included the following trial order. Habituation: four trials (2 CS1, 2 CS2) without shock to reduce initial orienting responses. Learning (acquisition): the fifth trial was a CS1 trial paired with shock signaling the start of acquisition, which included a total of 9 CS1 trials paired with shock, 12 CS1 trials unpaired (~42% reinforcement rate), and 12 CS2 trials. Reversal: a CS2 trial paired with shock signaled the start of reversal, which included a total of 9 CS2 trials paired with shock, 12 CS2 trials unpaired and 12 CS1 trials. An additional 9 CS1 and 9 CS2 trials were included after reversal in order to extinguish conditioned responses prior to a generalization test, which occurred in a separate scanning run. The extinction trials were excluded from the present analyses *a priori* and are not reported. CS+ trials paired with shock were not included in imaging or psychophysiological analysis to mitigate potential confounds in the BOLD signal and SCR introduced by the electric shock (US).

All trials were 2.5 s followed by a jittered 5.5–9.5 s (mean = 7.5 s) intertrial interval with a fixation point. Subjects rated shock expectancy on each trial on a three alternative-forced-choice scale using symbols −,?, and +, corresponding to 'no shock,' 'maybe shock,' and 'shock,' based on Boll *et al.* (2013). Subjects were not informed of the CS–US contingencies, and were told to pay attention and try to learn the association

1 Traditionally, "in-group" is defined as a group for which an individual self-identifies as being a member, and "out-group" is a group for which an individual does not identify (Tajfel, 1974). We frame our findings in terms of in-group/out-group distinctions, because excluding the small number of Asian participants does not affect the primary results.

2 Participants also completed the Modern Racism Scale (McConahay, 1986), State and Trait Anxiety Inventory (Spielberger, 1983), and Intolerance of Uncertainty (Buhr and Dugas, 2002). Groups did not differ on these factors, and we found no relationship among these measures and reported neurobehavioral results.

between the pictures and the shock. Trial order was pseudo-randomized so that no more than three CSs of the same face occurred in a row. Two stimulus presentation orders were created and counterbalanced between participants. Stimulus presentation was controlled using E-Prime 2.0.

After the reversal task, and in separate scanning run, subjects were exposed to novel images of Black and White faces as a test of fear generalization. Because the results of this generalization phase do not directly pertain to the preceding fear-learning and reversal task, we do not present the data in this report.

Psychophysiology and shock

SCR collection was controlled by the BIOPAC MP-100 System using MRI-compatible electrodes (Goleta, CA). Electrodermal activity was measured throughout the experiment from the hypothenar eminence of the left palm. Data were analyzed using Matlab (MathWorks, Natick, MA), and low-pass Butterworth filtered prior to analysis. SCRs were analyzed by subtracting the mean skin conductance level 1 sec prior to CS onset from the max skin conductance level in a 0.5 to 3.5 s latency window following CS onset. Values less than .02 microsiemens were entered as 0. Raw SCRs were normalized by range-correction using each subjects maximum SCR (induced by the shock), and square-root transformed prior to analysis (Lykken and Venables, 1971). For all analyses, we excluded CS+ trials paired with shock. Subjects who did not exhibit measurable electrodermal activity were excluded from SCR analysis. Exclusionary criteria included subjects who did not show a detectable SCR on any trials including to the shock itself. Failure to evince measurable electrodermal in some individuals' responses is likely related to technical challenges of collecting SCR in the MRI environment. Seven participants were excluded based on this criterion: three from the BSF and four from the WSF group.

We derived four learning indices to examine changes in SCRs to the CS1 and CS2 in each group. As shown by Zhang *et al.* (2014), separate learning indices provide a means to directly investigate whether reversals are selectively impaired due to a failure in updating responses to CS1, CS2 or both. The first index assessed acquisition (ACQ) by calculating difference in SCRs between the CSs during the acquisition phase ($CS1_{Acquisition} - CS2_{Acquisition}$). The second index assessed reversal (REV) by calculating the difference in SCRs between the CSs during the reversal phase ($CS2_{Reversal} - CS1_{Reversal}$). Reversals (or impaired reversals) can be driven by updating responses (or persistent responses) to CS1, CS2 or both. Successful reversal to CS1 (Δ CS1) is characterized by diminished SCRs from acquisition to reversal ($CS1_{Reversal} - CS1_{Acquisition}$); successful reversals to the CS2 (Δ CS2) is characterized by *enhanced* SCRs from acquisition to reversal ($CS2_{Reversal} - CS2_{Acquisition}$). For all statistical tests, analyses were considered significant at $P < 0.05$, two-tailed.

The US was an uncomfortable 200 msec electrical shock delivered to the right wrist, connected to the Grass Medical SD9 stimulator. The intensity of shock was calibrated for each subject prior to entering the scanner to reach a level deemed highly annoying but not painful (Dunsmoor *et al.*, 2009).

Functional MRI acquisition, preprocessing and analysis

Whole-brain functional imaging was conducted on a 3T Siemens Allegra head-only scanner. Blood oxygenation level-dependent functional images were acquired parallel to the AC-PC line using a standard EPI sequence: acquisition matrix, 64×64; field of view, 192×192; flip-angle, 85°; 36 slices with interleaved acquisition; slice thickness, 3 mm; repetition time, 2 s; echo time, 30 ms.

Preprocessing and data analysis was conducted using SPM8 (Wellcome Trust Centre, www.fil.ion.ucl.ac.uk) implemented in MATLAB. Images were spatially normalized into Montreal Neurological Institute (MNI) space, voxel size resampled to $2 \times 2 \times 2$ mm, and smoothed using an isotropic 8-mm^3 Gaussian full-width half-maximum kernel. Functional images were co-registered to each participant's high-resolution T1-weighted structural scan. To account for magnetic equilibrium, the first four functional images were discarded. Images were corrected for head motion using a 3-mm movement cutoff in any dimension.

At the first-level (individual subject), separate covariates were created for the onset of $CS1_{Acquisition}$, $CS2_{Acquisition}$, $CS1_{Reversal}$, and $CS2_{Reversal}$. Stimulus duration was modeled as the reaction time on each trial (Grinband *et al.*, 2008). Additional covariates of no interest included CS+ trials paired with shock, the US, extinction trials, and 6 head motion parameters. We excluded all CS+ trials paired with shock because trials were short (2.5 s), and shock co-terminated with CS presentation. This mitigates potential confounds in the BOLD signal introduced by the shock. Events were convolved with the canonical hemodynamic response function, and a high-pass filter of 128 s was applied.

To identify regions involved in fear-learning and reversal, fMRI analysis focused on regions identified as showing a CS × Phase interaction at the group-level (listed in Table 1). Specifically, first-level contrast images were created using contrast weights 1, −1, −1, 1 for regressors $CS1_{Acquisition}$, $CS2_{Acquisition}$, $CS1_{Reversal}$, $CS2_{Reversal}$, respectively. First-level contrast images were taken to the group-level in SPM8. A one-sample *t*-test, incorporating subjects from both groups ($N = 41$), revealed regions showing the CS × Phase interaction. For whole-brain analysis, we used a voxel-wise probability of $P < 0.001$ and a minimum cluster size of 67 voxels to achieve a cluster correction of $P < 0.05$. Cluster correction was derived from 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. Parameter estimates were extracted from regions of interest to investigate patterns of learning and reversal in both groups separately. The same learning indices approach used for SCR analysis (*ACQ*, *REV*, Δ CS1, Δ CS2; see Table 2) were applied to betas from the regions identified at the group level. *A priori* regions of interest included areas traditionally implicated in prior fMRI studies of human fear-learning (see Fullana *et al.*, 2015 for meta-analysis), including the thalamus, striatum, ACC, midbrain and insula. The amygdala was incorporated as an *a priori* ROI based on its role in conditioned fear widely identified in animal neurophysiological research (Pape and Paré, 2010). Because the amygdala is frequently not identified in human fear-learning using standard univariate imaging procedures (Fullana *et al.*, 2015), but is commonly identified in fMRI studies of face processing unrelated to learning or emotional facial expression (Mende-Siedlecki *et al.*, 2013), a face localizer task was used to independently identify the amygdala.

The face localizer included four blocks each of faces, objects and scenes. Each block contained 12 pictures presented for 800 ms separated by a 200-ms blank screen, and followed by a 12-s fixation. Face-selective activity was identified at the group-level (faces > scenes + objects) at $P < 0.05$, FWE corrected for the whole-brain, which revealed bilateral activations in the left

Table 1. Regions exhibiting significant CS (CS1, CS2) × Phase (Learning, Reversal) interaction at the group-level, identified at $P < 0.001$, cluster corrected $P < 0.05$

Region	Hemisphere	MNI coordinates			Size (voxels)	Peak T	Peak Z
		x	y	z			
Insula	Left	−34	16	−2	1514	8.08	6.18
Insula	Left	−26	24	−2		7.61	5.95
Insula	Left	−38	4	14		4.10	3.72
Insula	Right	34	24	0	2197	7.79	6.04
Insula	Right	44	14	8		7.29	5.78
Precentral gyrus	Right	50	18	4		6.60	5.39
Superior frontal gyrus	Left	2	24	48	2162	6.07	5.08
Superior frontal gyrus	Right	6	14	56		5.87	4.95
Cingulate gyrus	Left	−6	14	44		5.70	4.84
Middle frontal gyrus	Right	38	2	42	384	5.59	4.77
Precentral gyrus	Right	44	8	34		4.60	4.09
Precentral gyrus	Right	40	16	32		3.61	3.33
Supramarginal gyrus	Right	64	−44	28	367	5.42	4.66
Inferior parietal lobule	Right	44	−50	48		4.53	4.04
Supramarginal gyrus	Right	50	−38	32		4.29	3.87
Medial frontal gyrus	Right	26	44	20	308	5.33	4.61
Middle frontal gyrus	Right	38	46	12		4.01	3.66
Middle frontal gyrus	Right	38	44	4		3.67	3.39
Postcentral gyrus	Left	−60	−24	22	107	4.66	4.14
Posterior insula	Right	46	−30	−6	79	4.48	4.01
Thalamus	Right	6	−12	−2	339	4.29	3.87
Midrain	Left	−4	−22	−6		4.26	3.85
Caudate head	Right	8	8	2		4.03	3.67

Table 2. Analysis of parameter estimates extracted from *a priori* regions of interest identified by the group-level CS × Phase interaction

Region	Black Shocks First group				White Shocks First group			
	CS1 = Black; CS2 = White				CS1 = White; CS2 = Black			
	*P < 0.05; **P < 0.01; ***P < 0.001; ns P > 0.05;				*P < 0.05; **P < 0.01; ***P < 0.001; ns P > 0.05;			
	ACQ	REV	ΔCS1	ΔCS2	ACQ	REV	ΔCS1	ΔCS2
A Priori regions of interest								
Insula (Left)	**	*	*	**	***	**	***	**
Insula (Right)	**	**	ns (0.064)	**	***	**	***	*
Thalamus	*	ns (0.980)	ns (0.095)	ns (0.733)	***	***	***	*
Caudate	**	ns (0.865)	ns (0.139)	ns (0.681)	**	**	**	*
Midbrain	**	ns (0.56)	ns (0.06)	ns (0.3)	***	***	***	*
Superior frontal gyrus/ACC	**	ns (0.114)	ns (0.062)	*	***	*	***	ns (0.057)
Other regions from FC/REV interaction analysis								
Inferior parietal lobule	***	**	*	*	***	ns (0.436)	**	ns (0.451)
Postcentral gyrus	*	ns (0.347)	ns (0.092)	ns (0.685)	***	ns (0.469)	***	ns (0.563)
Posterior insula	*	ns (0.276)	ns (0.160)	ns (0.104)	*	*	**	ns (0.180)
Middle frontal gyrus	***	ns (0.059)	*	*	**	*	ns (0.098)	*

($x = -20$; $y = -6$; $z = -16$; 487 voxels) and right ($x = 20$; $y = -6$; $z = -16$; 576 voxels) amygdala, and right fusiform gyrus ($x = 42$; $y = -48$; $z = -16$; 1721 voxels).

Results

Intact shock expectancy but impaired autonomic indicators of reversal following stereotypic fear-learning

As shown in Figure 1B, acquisition and reversal of Shock Expectancy was successful in both groups. Repeated-measures ANOVA of mean expectancy ratings using Phase (acquisition and reversal) and Condition (CS1, CS2) as within-subjects factors and Group (BSF,WSF) as a between-subjects factor showed a significant Phase by CS interaction ($F_{1,39} = 461.56$, $P < 0.001$ $\eta^2 = 0.922$), but no Phase by CS by Group interaction ($P > 0.6$). Thus, subjects in both groups were able to correctly identify (explicitly) the correct danger and safety stimulus during acquisition and reversal, regardless of the race identity of the CSs.

Similar to shock expectancy ratings, SCR analysis (Figure 2A; see also Supplementary Figure 1) revealed a Phase by CS interaction ($F_{1,\ 32} = 27.23$, $P < 0.001$, $\eta^2 = 0.46$). Unlike shock

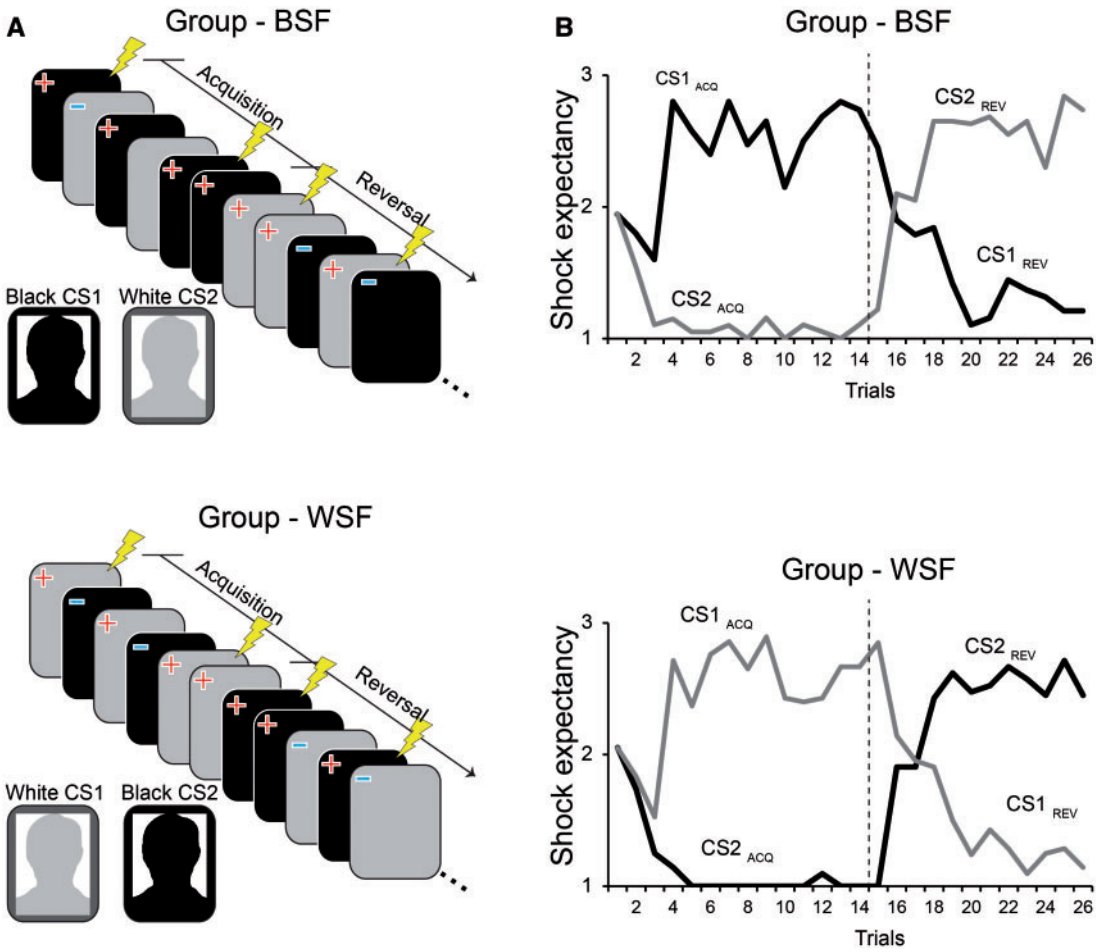

Fig. 1. Experimental design and trial-by-trial shock expectancy ratings. (A) The Black-Shocks-First (BSF) and White-Shocks-First (WSF) groups both saw the same Black and White male exemplar. For the BSF group, the Black male (CS1) was originally paired with an electric shock unconditioned stimulus (US; depicted as the lightning bolt), and the White male (CS2) was unpaired. For the WSF group, the White male (CS1) was originally paired with shock, and the Black male (CS2) was unpaired. In both groups, the CS–US association reversed midway. (B) On each trial, subjects rated expectancy using a three-point scale corresponding to 'no shock' (=1), 'maybe shock' (=2), and 'shock' (=3). These declarative expectancy ratings tracked the CS–US contingencies accurately in both groups. Dashed line indicates the start of reversal.

expectancy ratings, however, there was a significant Phase by CS by Group interaction ($F_{1,\ 32}=4.27$, $P<0.047$, $\eta^2 = 0.11$). Based on the significant interaction, we analyzed fear-learning and reversal indices in each group.

The ACQ index ($CS1_{\ Acquisition} - CS2_{\ Acquisition}$) was significant in the BSF group (one-sample *t*-test; $t_{16}=3.55$, $P=0.003$) and the WSF group ($t_{16}=5.16$, $P<0.001$), with no difference between groups ($P = 0.26$), indicating that fear-learning was successful in both groups regardless of the race identity of CS1 and CS2. The *REV* index ($CS2_{\ Reversal} - CS1_{\ Reversal}$) was significant in the WSF group ($t_{16}=2.96$, $P = 0.009$) but not in the BSF group ($p=.5$), and *REV* was significantly different between groups ($t_{32}=2.69$, $P = 0.01$, $d = 0.923$).

The Δ CS1 and Δ CS2 index were used to determine whether impaired fear-reversals following stereotypic fear-learning (BSF group) were a result of persistent SCRs to the Black male target, an inability to shift fear responses to the White male target, or a combination of both (Figure 2B). The Δ CS1 index ($CS1_{\ Reversal} - CS1_{\ Acquisition}$) was significant in the BSF group ($t_{16}=5.92$, $P < 0.001$) and the WSF group ($t_{16}=5.87$, $P < 0.001$), with no difference between groups ($P > 0.1$). This result demonstrates that both groups successfully managed to reduce SCRs to the original danger stimulus once it no longer signaled shock, and was contrary to the prediction that a stereotypic threat association, once confirmed, would show persistent fear responses to the Black CS1 after fear-reversal (cf., Olsson *et al.*, 2005). The Δ CS2 index ($CS2_{\ Reversal} - CS2_{\ Acquisition}$) was significant in the WSF group ($t_{16}=2.36$, $P = 0.03$), but not in the BSF group ($P = 0.44$), and Δ CS2 was significantly different between groups ($t_{32}=2.29$, $P = 0.028$, $d = 0.785$). Thus, fear-learning involving stereotypic threat associations (BSF group) resulted in a specific deficit in the ability to flexibly update autonomic fear expression to the White male CS when the stimulus went from signaling safety to signaling danger.

Impaired reversal of neural activity following stereotypic fear-learning

Group-level analysis of the CS x Phase interaction revealed activity in a number of areas identified in prior fMRI studies of fear-learning in humans, including bilateral insula, ACC extending into supplementary motor area, caudate, thalamus, and midbrain (Figure 3A, Table 1). Notably, the amygdala was not identified from the group-level contrast, even at a reduced exploratory threshold of $P < 0.005$, uncorrected. Analysis of the amygdala ROI identified from the face localizer likewise did not

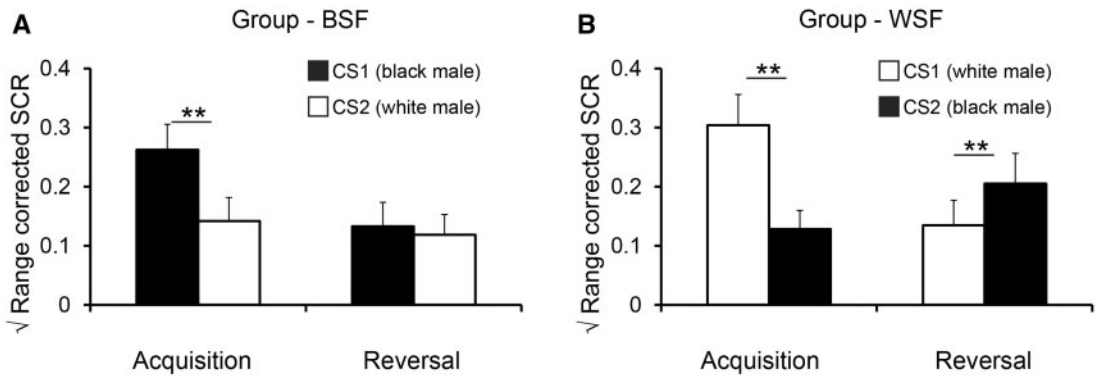

Fig. 2. Mean conditioned skin conductance responses. (A) The Black-Shocks-First (BSF) group exhibited intact learning but impaired reversal, as evidenced by enhanced SCRs to CS1 versus CS2 during Acquisition, but no difference between CS1 and CS2 during Reversal. (B) The White-Shocks-First (WSF) group showed intact learning and reversal. Error bars reflect standard errors; ** = $p < .01$.

yield greater activity for the CS1 as compared to CS2 during learning, or CS2 as compared to CS1 during reversal, in either group. This lack of amygdala engagement during fear conditioning was not entirely surprising given that the amygdala has not been identified in a number of fMRI studies of human fear conditioning that included a variety of CS types (Bach *et al.*, 2011; Fullana *et al.*, 2015).

Having identified regions of interest across all participants (2nd level in SPM), we next extracted parameter estimates from *a priori* ROIs to investigate indices of fear-reversal in each group separately. Learning-related activity was observed in each region identified at the 2nd level in the BSF and WSF groups, as indicated by greater activity to CS1 than CS2 prior to reversal (*ACQ* in Table 2). Thus, learning-related activity confirmed successful acquisition in both groups. Successful fear-reversal was observed in bilateral insula in both groups (*REV* in Table 2). However, fear-reversals were selectively impaired in the caudate, thalamus, midbrain, and ACC/SMA in the BSF group, whereas the WSF group showed significant reversals in each region. Comparing the *REV* index between groups showed significant impairments in the BSF versus WSF group in the caudate ($t_{39} = 2.37$, $P = 0.023$, $d = 0.740$), thalamus ($t_{39} = 2.85$, $P = 0.007$, $d = 0.890$), and midbrain ($t_{39} = 2.39$, $P = 0.022$, $d = 0.747$) (Figure 3B).

Based on prior findings that out-group contact moderates conditioning and extinction biases to race out-groups (Olsson *et al.*, 2005), we examined whether SCRs or neural activity associated with fear-reversal were correlated with out-group contact scores. We found no significant correlations between out-group contact scores and SCRs or neural activity, or between IAT and these behavioral and neural measures. Notably, this null finding may be a result of selecting participants with positive IAT D scores (see *Implicit Race bias* in the Method section)

Discussion

In a paradigm where stereotypic threat associations were challenged through reversal-learning, autonomic and neural responses failed to reverse if racially based threat stereotypes were initially reinforced. Neuroimaging results revealed impaired reversals in the striatum, midbrain, ACC, and thalamus; regions generally implicated in fear-learning that may be important for dynamically updating cue-outcome associations as contingencies change. In comparison, autonomic and neural responses successfully reversed if a counterstereotypic threat association (i.e. White males are dangerous but Black males are safe) changed to a stereotypic threat association, demonstrating an asymmetry in fear-reversal as a function of the initially learned threat association. We interpret these results within the framework of persistent implicit negative associations and threat stereotypes regarding race out-groups.

Learning-related activity, prior to fear-reversal, included a number of regions regarded as part of a 'central autonomic-interoceptive network' (Fullana *et al.*, 2015). It is notable that SCRs and learning-related BOLD activity was similar prior to reversal, regardless of whether the initial danger stimulus was a Black male or a White male. This finding is also noteworthy given the literature on race-related neural responses independent of affective learning manipulations. That is, many studies report enhanced BOLD activity to race out-groups in the ACC, fusiform gyrus, and amygdala (e.g., Lieberman *et al.*, 2005) that is linked predominately to the strength of implicit race biases (see Kubota *et al.*, 2012). Importantly, prior studies show that the context and task goals under which research subjects view other people can affect the strength of neural activity evoked by an out-group versus in-group member; for example, amygdala responses to racial out-group members are diminished when subjects focus attention on social individuating (i.e. vegetable preference) versus social categorizing (i.e. age) features (Wheeler and Fiske, 2005). In the present study, the task goal involved learning what individual face predicted shock. After subjects learned which individual face predicted shock, the BSF group showed a selective inability to then update this learning to express fear to the appropriate new danger stimulus.

The results from fear-learning (prior to reversal) are also in keeping with previous fear-learning literature showing equivalent SCRs to a Black and White male CS paired with shock (Olsson *et al.*, 2005; Molapour *et al.*, 2015) and equivalent neural activity to a Black and White male CS in fear-learning neurocircuity, including the insula, ACC and amygdala (Molapour *et al.*, 2015). However, unlike prior studies (Olsson *et al.*, 2005; Navarrete *et al.*, 2009), we did not observe persistent fear responses to the Black CS once it stopped predicting the US. It is at this point that the distinction between extinction and reversal tasks becomes apparent, and why we contend that reversal-learning offers a unique benefit to characterize associative learning deficits.

To restate issues introduced earlier, extinction deficits to threat-relevant stimuli (including, perhaps, racial out-group members) could reflect factors or processes unrelated to associative learning, per se. This includes different rates of habituation to threat-relevant CSs (Öhman *et al.*, 1974) or a

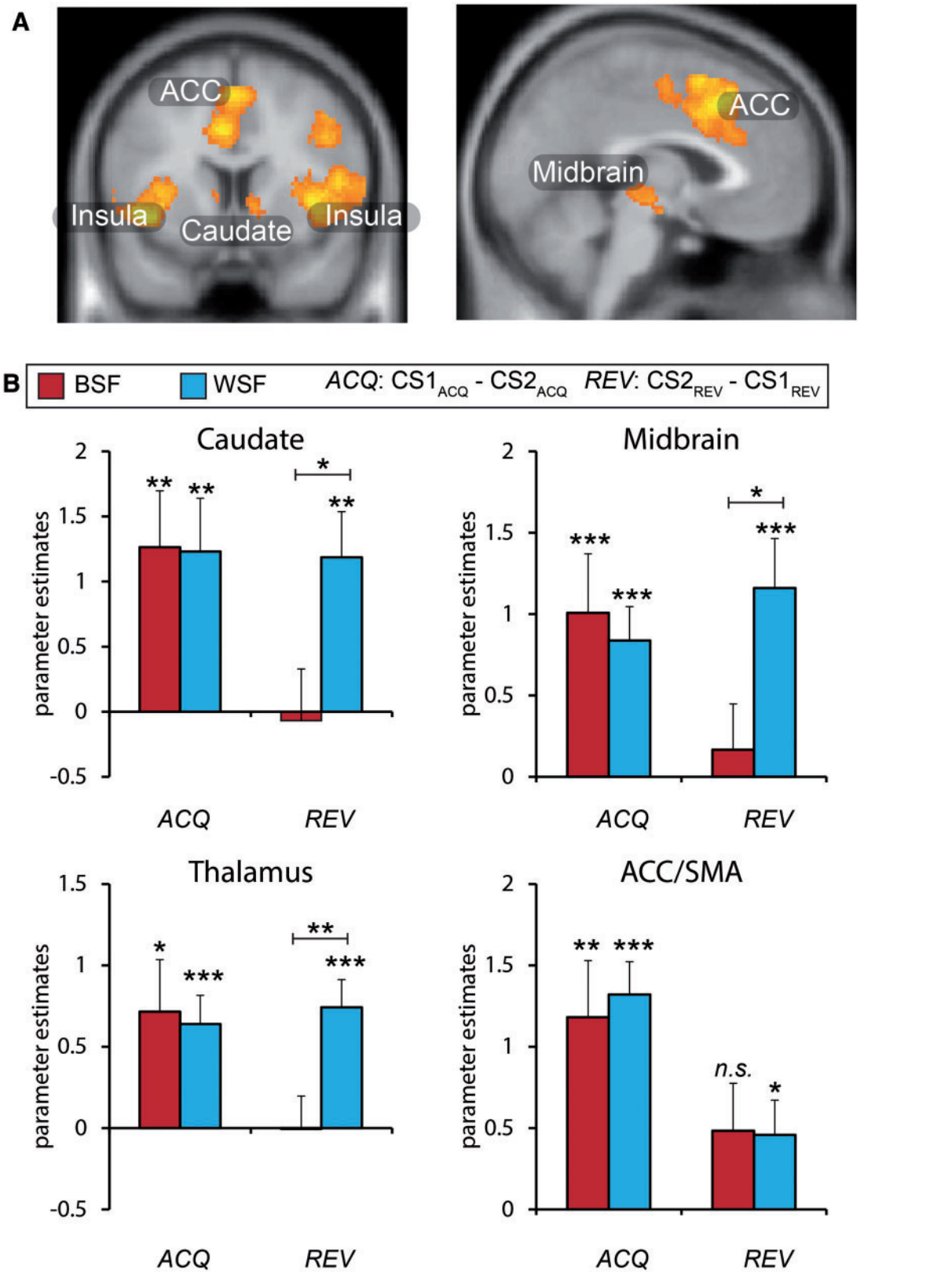

Fig. 3. Neuroimaging results. (A) Regions exhibiting increased activity to the current threat cue versus the current danger cue. This analysis revealed clusters in the insula, midbrain, anterior cingulate (ACC) extending into the supplementary motor area (SMA), and caudate across participants (see Table 1 for full list of regions). (B) Mean parameter estimates from regions of interest show intact acquisition (ACQ) in both groups, but impaired reversals (REV) selectively in the BSF group. Error bars reflect standard errors; $^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$. Whole brain activity displayed at $P < 0.001$, cluster corrected $P < 0.05$. BSF = Black-Shocks-First group; WSF = White-Shocks-First group.

pre-existing expectancy bias that threat-relevant CSs are more likely to be associated with an unpleasant outcome (Tomarken *et al.*, 1989; Davey, 1992), among other non-associative factors (McNally, 1987; Lovibond *et al.*, 1993; McNally, 2015). Persistent fear to an out-group member during extinction may also reflect phenomena well-described by statistical learning theories that learning is stronger to novel or unfamiliar stimuli (Dayan *et al.*, 2000; Courville *et al.*, 2006; Dunsmoor *et al.*, 2015). In this way, conditioning that involves a picture of a less familiar looking out-group member is stronger and more persistent than conditioning that involves a picture of a more familiar looking in-group member. Superior conditioning—as demonstrated through delayed extinction—might therefore reflect intact rather than impaired associative learning (Maia, 2009).

A failure to *reverse* fear expression and neural activity in the BSF group, in contrast, suggests that the initial stereotypic threat-learning experience interfered in some way with the ability to learn the new CS–US association. Thus, it is a selective failure in updating learning to be in line with the new situation, a finding not attributable solely to persistent fear to the Black CS, slower habituation to the Black CS, or experimental demand characteristics. Nor can these results be attributed to low-level perceptual differences between Black and White CSs, as both groups showed similar acquisition regardless of which CS acted as the threat or safety signal. One possible explanation of this selective impairment to shift responses towards the new CS in the BSF group is that the originally unreinforced stimulus (White male) acted as an especially potent safety signal by virtue of being compared to the stereotypic threat stimulus (Black male). Once the shock shifted to the White male, safety value that had accrued to the White male limited the ability for this stimulus to evoke a conditioned fear response, in a manner similar to a conditioned inhibitor (Christianson *et al.*, 2012).

In prior fMRI studies of reversal-learning, regions involved in acquisition of new learning reverse as contingencies shift (Cools *et al.*, 2002; Morris and Dolan, 2004; Schiller *et al.*, 2008; Boll *et al.*, 2013)—a finding replicated here if learning reversed from an initial counterstereotypic to a stereotypic threat association (WSF group). Elements of this circuitry, including the dorsal midbrain/periaqueductal gray and the striatum, are particularly important for updating learning based on aversive prediction errors (Delgado *et al.*, 2008; Schiller *et al.*, 2008; Li *et al.*, 2011; McNally *et al.*, 2011; Roesch *et al.*, 2012; Roy *et al.*, 2014). Research relevant for the striatum's role in prejudice is scant, but its role in intergroup biases has been shown during assessment of intergroup trust (Stanley *et al.*, 2012), and when perceivers lack prior experience with the out-group (Van Bavel *et al.*, 2011).

In conclusion, the present results offer insight into social factors influencing the implicit expression of intensified fear-learning, and provide new evidence that aversive experiences with a member of a racial out-group may impair future learning about the aversive nature of an in-group member. This finding adds to our understanding about the basic learning mechanisms that may contribute to the persistence of stereotypes and in-group favoritism. Specifically, these results go beyond detailing the effects of race-based threat stereotypes on perceptions of out-group members, and open the way for an area of research that investigates how out-group stereotypes may interfere with the ability to learn about members of one's own social group.

Funding

This study was supported by NIH RO1 MH097085 to E.A. Phelps. J.E. Dunsmoor is supported by NIMH K99MH106719 and J. Li is supported by the National Natural Science Foundation of China 31322022. C.A.O. Coelho is supported by BEPE FAPESP #2013/10907-3.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

Acknowledgments

The authors would like to thank Oriel FeldmanHall for helpful comments on the manuscript, and Leeann Ozer for assistance on behavioral piloting.

References

Amodio, D.M., Devine, P.G. (2006). Stereotyping and evaluation in implicit race bias: evidence for independent constructs and unique effects on behavior'. *Journal of Personality and Social Psychology*, **91**, 652.

Bach, D.R., Weiskopf, N., Dolan, R.J. (2011). A stable sparse fear memory trace in human amygdala. *Journal of Neuroscience* **31**, 9383–9.

Bari, A., Robbins, T.W. (2013). Inhibition and impulsivity: behavioral and neural basis of response control. *Progress in Neurobiology* **108**, 44–79.

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans'. *European Journal of Neuroscience*, **37**, 758–67.

Bouton, M.E. (2002). Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biol Psychiatry*, **52**, 976–86.

Buhr, K., Dugas, M.J. (2002). The intolerance of uncertainty scale: psychometric properties of the English version'. *Behav Res Ther* **40**, 931–45.

Christianson, J.P., Fernando, A.B., Kazama, A.M., Jovanovic, T., Ostroff, L.E., Sangha, S. (2012). Inhibition of fear by learned safety signals: a mini-symposium review'. *The Journal of Neuroscience*, **32**, 14118–24.

Cools, R., Clark, L., Owen, A.M., Robbins, T.W. (2002). Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging'. *The Journal of Neuroscience*, **22**, 4563–7.

Correll, J., Park, B., Judd, C.M., Wittenbrink, B. (2002). The police officer's dilemma: using ethnicity to disambiguate potentially threatening individuals. *Journal of Personality and Social Psychology*, **83**, 1314.

Courville, A.C., Daw, N.D., Touretzky, D.S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, **10**, 294–300.

Davey, G.C.L. (1992). An expectancy model of laboratory preparedness effects. *Journal of Experimental Psychology-General*, **121**, 24–40.

Dayan, P., Kakade, S., Montague, P.R. (2000). Learning and selective attention. *Nature Neuroscience*, **3**, 1218–23.

Delgado, M.R., Li, J., Schiller, D., Phelps, E.A. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **363**, 3787–800.

Devine, P.G. (1989). Stereotypes and prejudice: their automatic and controlled components. *J Pers Soc Psychol*, **56**, 5.

Dias, R., Robbins, T., Roberts, A. (1996). Dissociation in prefrontal cortex of affective and attentional shifts'. *Nature*, **380**, 69–72.

Dunsmoor, J.E., Kragel, P.A., Martin, A., LaBar, K.S. (2014). Aversive Learning Modulates Cortical Representations of Object Categories'. *Cerebral Cortex*, **24**, 2859–72.

Dunsmoor, J.E., Mitroff, S.R., LaBar, K.S. (2009). Generalization of conditioned fear along a dimension of increasing fear intensity. *Learning & Memory*, **16**, 460–9.

Dunsmoor, J.E., Niv, Y., Daw, N.D., Phelps, E.A. (2015). Rethinking extinction. *Neuron*, **88**, 47–63.

Eberhardt, J.L., Davies, P.G., Purdie-Vaughns, V.J., Johnson, S.L. (2006). Looking deathworthy perceived stereotypicality of black defendants predicts capital-sentencing outcomes. *Psychological Science*, **17**, 383–6.

Eberhardt, J.L., Goff, P.A., Purdie, V.J., Davies, P.G. (2004). Seeing black: race, crime, and visual processing. *Journal of Personality and Social Psychology*, **87**, 876.

Fullana, M., Harrison, B., Soriano-Mas, C., *et al.* (2015). Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Mol Psychiatry*, **21**, 500–8.

Goff, P.A., Eberhardt, J.L., Williams, M.J., Jackson, M.C. (2008). Not yet human: implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology*, **94**, 292.

Golkar, A., Björnstierna, M., Olsson, A. (2015). Learned fear to social out-group members are determined by ethnicity and prior exposure. *Name*. Frontiers in Psychology, **6**, 123.

Greenwald, A.G., McGhee, D.E., Schwartz, J.L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, **74**, 1464.

Greenwald, A.G., Nosek, B.A., Banaji, M.R. (2003). Understanding and using the implicit association test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, **85**, 197.

Greenwald, A.G., Poehlman, T.A., Uhlmann, E.L., Banaji, M.R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology*, **97**, 17.

Grinband, J., Wager, T.D., Lindquist, M., Ferrera, V.P., Hirsch, J. (2008). Detection of time-varying signals in event-related fMRI designs. *Neuroimage* **43**, 509–20.

Kehagia, A.A., Murray, G.K., Robbins, T.W. (2010). Learning and cognitive flexibility: frontostriatal function and monoaminergic modulation. *Current Opinion in Neurobiology*, **20**, 199–204.

Kubota, J.T., Banaji, M.R., Phelps, E.A. (2012). The neuroscience of race. *Nature Neuroscience*, **15**, 940–8.

Kubota, J.T., Ito, T.A. (2014). The role of expression and race in weapons identification. *Emotion*, **14**, 1115.

Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., Daw, N.D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, **14**, 1250–2.

Lieberman, M.D., Hariri, A., Jarcho, J.M., Eisenberger, N.I., Bookheimer, S.Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, **8**, 720–2.

Lindström, B., Golkar, A., Olsson, A. (2015). A clash of values: Fear-relevant stimuli can enhance or corrupt adaptive behavior through competition between Pavlovian and instrumental valuation systems'. *Emotion*, **15**, 668.

Lovibond, P.F., Siddle, D.A.T., Bond, N.W. (1993). Resistance to extinction of fear-relevant stimuli - preparedness or selective sensitization. *Journal of Experimental Psychology-General*, **122**, 449–61.

Lubow, R.E. (1973). Latent inhibition. *Psychological Bulletin*, **79**, 398–407.

Lykken, D.T., Venables, P.H. (1971). Direct measurement of skin conductance: a proposal for standardization. *Psychophysiology*, **8**, 656–72.

Maia, T.V. (2009). Fear conditioning and social groups: statistics, not genetics. *Cognitive Science*, **33**, 1232–51.

McConahay, J.B. (1986) *Modern racism, ambivalence, and the modern racism scale*. In J. Dovidio and S. Gaertner (Eds.), Prejudice, discrimination, and racism (pp. 91–126). Orlando, FL: Academic Press.

McNally, G.P., Johansen, J.P., Blair, H.T. (2011). Placing prediction into the fear circuit. *Trends in Neurosciences*, **34**(6), 283–92.

McNally, R.J. (1987). Preparedness and phobias: a review. *Psychological Bulletin*, **101**, 283.

McNally, R.J. (2015). The Legacy of Seligman's Phobias and Preparedness (1971). *Behavior Therapy*. Available: http://dx.doi.org/10.1016/j.beth.2015.08.005.

Mende-Siedlecki, P., Verosky, S.C., Turk-Browne, N.B., Todorov, A. (2013). Robust selectivity for faces in the human amygdala in the absence of expressions. *Journal of Cognitive Neuroscience*, **25**, 2086–106.

Molapour, T., Golkar, A., Navarrete, C.D., Haaker, J., Olsson, A. (2015). Neural correlates of biased social fear learning and interaction in an intergroup context. *Neuroimage*, **121**, 171–83.

Morris, J., Dolan, R. (2004). Dissociable amygdala and orbitofrontal responses during reversal fear conditioning. *Neuroimage* **22**, 372–80.

Navarrete, C.D., Olsson, A., Ho, A.K., Mendes, W.B., Thomsen, L., Sidanius, J. (2009). Fear extinction to an out-group face the role of target gender. *Psychological Science*, **20**, 155–8.

Nosek, B.A., Banaji, M., Greenwald, A.G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, **6**, 101.

Öhman, A. (2005). Conditioned fear of a face: A prelude to ethnic enmity?. *Science*, **309**, 711–3

Öhman, A., Eriksson, A., Fredriksson, M., Hugdahl, K., Olofsson, C. (1974). Habituation of the electrodermal orienting reaction to potentially phobic and supposedly neutral stimuli in normal human subjects. *Biological Psychology*, **2**, 85–93.

Öhman, A., Mineka, S. (2001). Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological Review*, **108**, 483–522.

Olsson, A., Ebert, J.P., Banaji, M.R., Phelps, E.A. (2005). The role of social groups in the persistence of learned fear'. *Science*, **309**, 785–7.

Pape, H.C., Paré, D. (2010). Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. *Physiological Reviews*, **90**, 419–63.

Payne, B.K. (2001). Prejudice and perception: the role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology*, **81**, 181.

Plant, E.A., Peruche, B.M. (2005). The consequences of race for police officers' responses to criminal suspects'. *Psychological Science*, **16**, 180–3.

Quillian, L., Pager, D. (2001). Black neighbors, higher crime? The role of racial stereotypes in evaluations of neighborhood crime1'. *American Journal of Sociology*, **107**, 717–67.

Remijnse, P.L., Nielen, M.M., Uylings, H.B., Veltman, D.J. (2005). Neural correlates of a reversal learning task with an affectively neutral baseline: an event-related fMRI study'. *Neuroimage*, **26**, 609–18.

Roesch, M.R., Esber, G.R., Li, J., Daw, N.D., Schoenbaum, G. (2012). Surprise! Neural correlates of Pearce–Hall and Rescorla–Wagner coexist within the brain'. *European Journal of Neuroscience*, **35**, 1190–200.

Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G.E., Wager, T.D. (2014). Representation of aversive prediction errors in the human periaqueductal gray'. *Nature Neuroscience*, **17**, 1607–12.

Sagar, H.A., Schofield, J.W. (1980). Racial and behavioral cues in black and white children's perceptions of ambiguously aggressive acts. *Journal of Personality and Social Psychology*, **39**, 590.

Schiller, D., Delgado, M.R. (2010). Overlapping neural systems mediating extinction, reversal and regulation of fear. *Trends in Cognitive Sciences*, **14**, 268–76.

Schiller, D., Levy, I., Niv, Y., LeDoux, J.E., Phelps, E.A. (2008). From fear to safety and back: reversal of fear in the human brain. *Journal of Neuroscience*, **28**, 11517–25.

Spielberger, C.D. (1983) *Manual for the State-Trait Anxiety Inventory*. Palo Alto, CA: Consulting Psychologists Press.

Stanley, D.A., Sokol-Hessner, P., Fareri, D. S., *et al.* (2012). Race and reputation: perceived racial group trustworthiness influences the neural correlates of trust decisions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 744–53.

Tajfel, H. (1974). Social identity and intergroup behaviour. *Social Science Information/Sur Les Sciences Sociales*, **13**, 65–93.

Tomarken, A.J., Cook, M., Mineka, S. (1989). Fear-relevant selective associations and covariation bias. *Journal of Abnormal Psychology*, **98**, 381–94.

Van Bavel, J.J., Packer, D.J., Cunningham, W.A. (2011). Modulation of the fusiform face area following minimal exposure to motivationally relevant faces: evidence of in-group enhancement (not out-group disregard). *Journal of Cognitive Neuroscience*, **23**, 3343–54.

Wheeler, M.E., Fiske, S.T. (2005). Controlling racial prejudice social-cognitive goals affect amygdala and stereotype activation. *Psychological Science*, **16**, 56–63.

Zhang, Z., Manson, K.F., Schiller, D., Levy, I. (2014). Impaired associative learning with food rewards in obese women. *Current Biology*, **24**, 1731–6.