Supplementary information for

Differential roles of human striatum and amygdala in associative learning

Jian Li,[1,2] Daniela Schiller,[1,2] Geoffrey Schoenbaum,[3,4] Elizabeth A. Phelps[1,2] & Nathaniel D. Daw[1,2]

[1]Psychology Department and [2]Center for Neural Science, New York University, New York, New York 10003, and [3]Department of Anatomy and Neurobiology, and [4] Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland 20201.

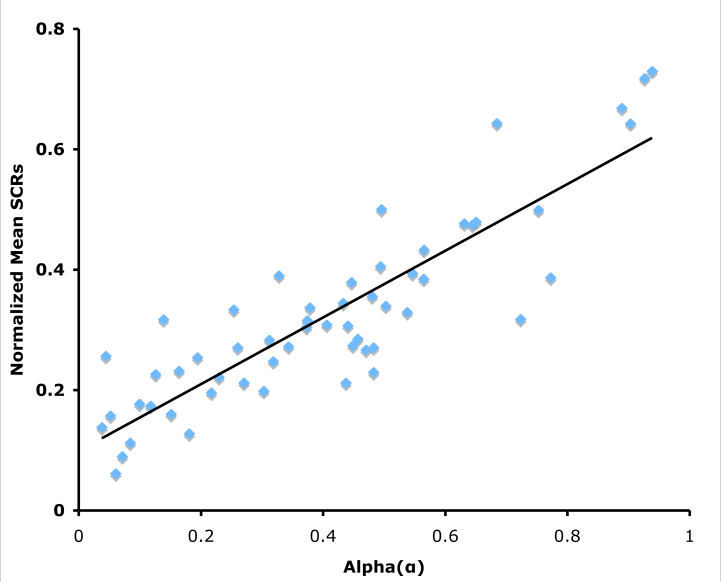
This file includes:

Supplementary data

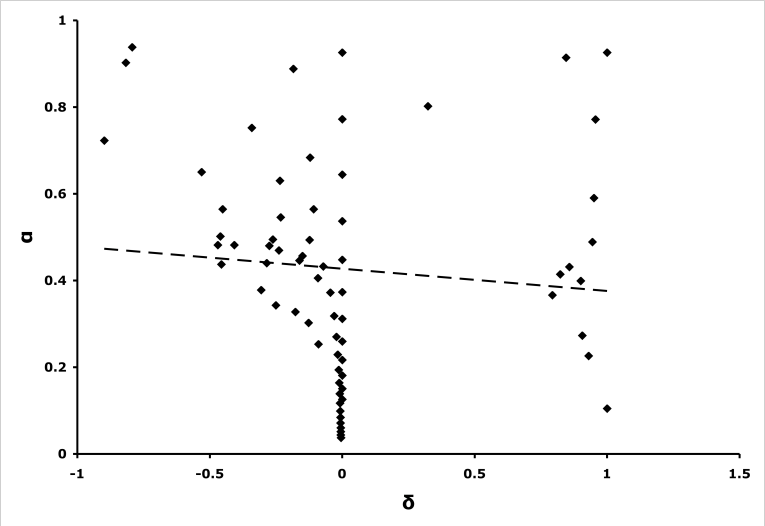
Supplementary Figures 1-3.

Supplementary Tables 1-5.

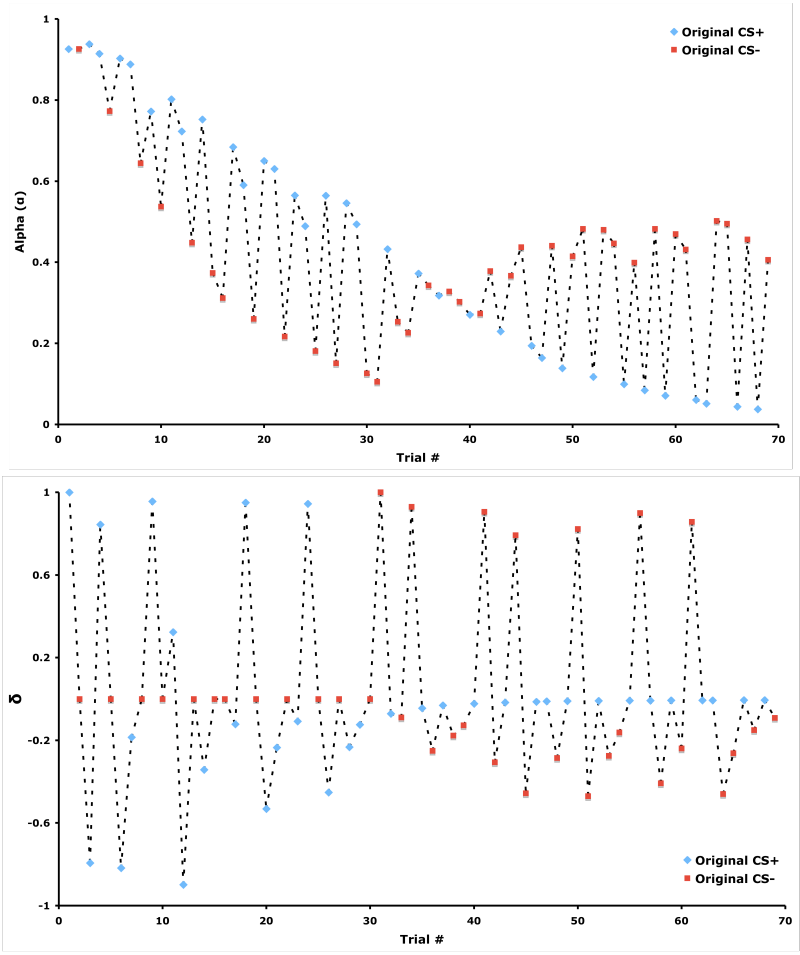
Supplementary Methods



Supplementary Figure 1. Relationship between modeled associability (α) and subjects' normalized skin conductance responses (SCRs) using the best fitting parameters for the hybrid model (see **Supplementary Methods** for details). Each data represents average SCRs for each trial averaged across all 17 subjects. The black curve is the best fitting line using least squares. This scatter-plot is provided to illustrate the fit of the optimized model; however, because the free parameters of the hybrid model were fit to optimize this correlation, it does not provide an independent measure of effect size[10].



Supplementary Figure 2. Regression analysis between associability (α) and prediction error (δ) generated using the best fitting parameters in the hybrid model ($R^2 = 0.009$, $p = 0.44$).



Supplementary Figure 3. Dynamics of associability (α) (upper panel) and prediction error (δ) (lower panel) generated from the best fitting parameters of the hybrid model. Trials labeled as original CS+ were CS+ trials in the acquisition phase but CS– (safe) trials in the reversal phase (blue); trials labeled as original CS– were CS– (safe) trials in the acquisition phase but CS+ trials in the reversal phase (red; see **Supplementary Methods** for detail).

	-LL	Vs. RW(V)	Vs. Hybrid(V)	Vs. Hybrid(α)
RW(V)	507.25	-	-	-
Hybrid(V)	455.04	$\chi_{34}^2 = 104.42$ $p < 0.00001$	-	-
Hybrid(α)	461.34	-	-	-
Hybrid(α+V)	423.23	$\chi_{51}^2 = 168.05$ $p < 0.00001$	$\chi_{17}^2 = 63.63$ $p < 0.0001$	$\chi_{17}^2 = 76.22$ $p < 0.0001$

Supplementary Table 1. Summary of the quality of individual behavioral fits from 17 subjects (56 trials per subjects), for all the 4 models we tested using likelihood ratio test (see **Supplementary Methods** for details). –LL: negative log-likelihood. For each model, -LL score is shown for fits to the SCRs of the associability time series, α, or the predicted value time series, *V* or a combination of both.

	-LL	Vs. RW(V)	Vs. Hybrid(V)	Vs. Hybrid(α)
RW(V)	871.52	-	-	-
Hybrid(V)	868.39	$\chi_2^2 = 6.27$ $p = 0.043$	-	-
Hybrid(α)	863.00	-	-	-
Hybrid(α+V)	861.13	$\chi_3^2 = 20.79$ $p < 0.0001$	$\chi_1^2 = 14.52$ $p < 0.0001$	$\chi_1^2 = 3.74$ $p = 0.053$

Supplementary Table 2. Quality of behavioral fits to 952 trials from 17 subjects, for all the 4 models we tested using likelihood ratio test (see **Supplementary Methods** for details). –LL: negative log-likelihood. For each model, -LL score is shown for fits to the SCRs of the associability time series, α, or the predicted value time series, *V* or a combination of both.

Hybrid(α)	
# Parameters	4
$\mathbf{V_0}$	0.000
$\boldsymbol{\alpha_0}$	0.926
η	0.166
κ	0.857

Supplementary Table 3. Best fitting parameters for the Hybrid (α) model, used to generate prediction error and associability regressors for the BOLD analysis.

	Cluster Size	T Statistics	Z Statistics	x,y,z {mm}
Inferior Frontal Cortex (L)	123	8.39	5.12	-54 18 -4
	185	5.73	4.17	58 16 -4
Ventral Thalamus	129	6.1	4.32	14 -4 16
	6	4.07	3.32	-6 -4 8
Insula (L)	12	4.7	3.67	-40 -18 -4
Insula (R)	85	6.44	4.46	40 0 -10
Postcentral Gyrus	13	5.2	3.92	68 -30 20
Superior Temporal Gyrus	16	5.14	3.89	54 -2 0
Precuneus	15	4.57	3.6	-4 -58 64
Amygdala (L)	**2**	**3.74**	**3.12**	**-18 0 -16**
Amygdala (R)	**6**	**4.68**	**3.66**	**20 4 -24**
Middle Cingulate Cortex	5	4.46	3.54	2 2 42
	3	3.9	3.22	-2 14 32
BA 7	4	3.81	3.17	6 -70 54
BA 9	3	3.97	3.26	28 54 32
BA 19	6	4.35	3.48	4 -88 28
BA 39	62	4.28	3.44	-58 -62 10
BA 42	15	4.21	3.41	-62 -24 12
BA 43	46	5.31	3.97	68 -20 20
Inferior Parietal Lobe	4	4.04	3.3	56 -50 40
	16	5.01	3.83	-56 -40 26
	5	3.92	3.23	64 -44 36
Middle Frontal Gyrus	6	4.26	3.43	44 50 -6
	4	3.81	3.17	-46 8 52
Middle Temporal Gyrus	29	5.98	4.27	64 -34 -6
Midbrain	5	4.28	3.44	18 -22 -10

Supplementary Table 4. Brain regions showing significantly positive correlations with associability (α). We report those areas surviving whole brain analysis with threshold $p < 0.001$(uncorrected).

	Cluster Size	T Statitics	Z Statistics	x,y,z {mm}
Subcallosal Gyrus	18	5.99	4.28	-20 10 -16
Medial Frontal Gyrus	32	5.96	4.26	8 42 26
	18	5.26	3.95	-14 52 16
	3	3.96	3.26	12 40 34
	3	3.96	3.26	12 40 34
Caudate Head	**39**	**5.18**	**3.91**	**-12 8 2**
Superior Frontal Gyrus	5	4.62	3.63	14 50 -18
	7	3.92	3.23	-18 56 -8
Inferior Frontal Gyrus	2	4.04	3.3	-48 14 -2
	3	3.85	3.19	-28 22 -14
BA 6	6	4.42	3.52	-10 28 38
BA 9	8	4.92	3.78	-36 26 40
BA 10	19	4.81	3.73	12 52 -4
BA 19	7	5.76	4.18	-36 -82 30
BA 40	5	4.29	3.45	-40 -42 50
Precentral Gyrus	6	4.18	3.38	-24 -20 68
Anterior Cingulate	3	3.97	3.27	0 36 2
Posterior Cingulate Gyrus	4	4.2	3.4	-8 -44 34

Supplementary Table 5. Brain regions showing significantly positive correlations with prediction error (δ). We report those areas surviving whole brain analysis with threshold $p < 0.001$(uncorrected).

Supplementary Methods

Subjects. Seventeen healthy right-handed subjects (9 males) whose ages were between 18 and 31 years old were recruited for the reversal-learning task. All participants provided informed consent and were paid a flat fee ($40) for their participation. The experiment was approved by the NYU University Committee on Activities Involving Human Subjects (UCAIS).

Experiment Design. We re-analyzed the data from a previously published[1] fear reversal-learning task with partial reinforcement (**Fig. 1a**). Subjects were instructed that they would view visual images (faces) on a computer screen while possibly receiving electric shocks. During the task, participants were instructed to focus on the computer screen and try to figure out the relationship between the visual stimuli and the shocks. Participants were not briefed about the fact that there were two stages (acquisition and reversal) of the experiment or when the reversal would occur[1].

Conditioned stimuli (CSs) were two mildly angry male faces from the Ekman series. Previous studies have shown that these stimuli effectively elicit fear conditioning[2]. The unconditioned stimulus (US) was a mild electric shock to the wrist (200ms duration, 50 pulses/s). The CSs were presented for 4s, and they co-terminated with the shocks. A 12s inter-trial interval (ITI) was inserted between trials with a fixation point at the center of the screen (**Fig. 1a**). In the acquisition phase, one face (face A) was paired with the US on one-third of the trials (CS+), and the other (face B) was never paired with the US (CS–). In the reversal phase, these reinforcement contingencies were reversed such that face B now was paired with the US on approximately one-third of the trials (new CS+) and face A became the safe stimulus (new CS–). The order of the trial types was fixed across subjects (this experiment design allowed us to average per-trial SCRs across subjects in **Fig. 1b** and supplemental Figure 1), and was generated pseudorandomly with two constraints: no consecutive reinforced trials and no more than two consecutive trials of the same kind. The designation of different faces (faces A & B) as CS+ and CS– was counterbalanced across subjects. The acquisition phase contained 18 CS+ trials, 6 of which ended with the delivery of US, and 12 CS– trials. The unsignaled reversal phase immediately followed the acquisition phase with 16 CS– trials and 23 CS+ trials (7 of which ended with the US

delivery). We designate the first trial in which the original CS– co-terminated with the electric shock as the beginning of the reversal phase (**Fig. 1a**).

Physiological stimuli and assessment. Mild shocks were delivered through a stimulating bar electrode attached with a Velcro® strap to the participants' right wrists. A Grass Medical Instruments stimulator charged by a stabilized current was used to deliver shocks, with cable leads that were magnetically shielded and grounded through a filter. Prior to scanning, participants were instructed to set the magnitude of the shock themselves using a work-up procedure in which subjects were first given a very mild shock (10V, 200ms, 50 pulses/s) that was gradually increased until the subject indicated the level of shock was "uncomfortable, but not painful." Skin conductance responses (SCR) were assessed with shielded Ag-AgCl electrodes, filled with standard NaCl electrolyte gel, that were attached to the middle phalanges of the second and third fingers of the left hand. Signals were amplified and recorded with a BIOPAC Systems skin conductance module connected to an Apple Macintosh computer. Data were continuously recorded at a rate of 200 samples per second. An off-line analysis of the analog skin conductance waveforms was conducted with AcqKnowledge software (BIOPAC Systems). The magnitude of SCRs was assessed for each trial as the peak-to-peak amplitude difference in skin conductance of the largest deflection (in micro-Siemens) in the 0.5-4.5s latency window after stimulus onset. The minimal response criterion was 0.02 μS. Responses below this criterion were interpreted as zero. The raw SCR scores were square root transformed to normalize the distributions, and scaled according to each subject's mean square-root-transformed US response[1].

Model fitting and selection. To test whether the behavior was consistent with the hypothesized learning mechanisms, and to validate and fit a model for subsequent fMRI analysis, we compared the fit of several reinforcement learning models to the trial-by-trial skin conductance responses. These included the Rescorla-Wagner model of learning by prediction errors and augmented hybrid models that gate prediction error driven learning by associability (as suggested by the Pearce-Hall model).

Models

We define x_n as the conditioned stimulus on trial n (CS+ or CS–) and r_n as the US delivered (1 for US, 0 for no US). All the models define value (i.e., shock) predictions $V_n(x)$ for each stimulus and trial. The punishment prediction error $\delta_n = r_n - V_n(x_n)$ measures the difference between the expected and predicted shock on trial n.

Rescorla-Wagner model

The Rescorla-Wagner model is a standard account of error-driven predictive learning. In our implementation, the values are initialized to V_0, a free parameter, then on each trial n the value of the observed CS x_n is updated according to the prediction error:

$$V_{n+1}(x_n) = V_n(x_n) + \alpha\delta_n$$

Here, the learning rate α for the value update is a constant free parameter. The value for the CS not observed on trial n remains unchanged.

Hybrid model

Note that our implementation of the Rescorla-Wagner model treats the learning rate for the value update as constant. It is natural to replace this assumption with the empirically well supported Pearce-Hall mechanism for associability-gated learning[3, 4], by substituting the Pearce-Hall associability for the constant learning rate. Equivalently, such a model incorporates Rescorla-Wagner's empirically well supported notion of error-driven value update into the Pearce-Hall associability model. The resulting hybrid model is:

$$V_{n+1}(x_n) = V_n(x_n) + \kappa\alpha_n(x_n)\delta_n$$
$$\alpha_{n+1}(x_n) = \eta|\delta_n| + (1-\eta)\alpha_n(x_n)$$

Note that trial n's associability α_n depends on (absolute) prediction errors from past trials, but not the current one. This makes δ_n and $\alpha_n(x_n)$ relatively uncorrelated to one another (important for seeking separate neural correlates), and also means that δ_n is not "double counted" in the value update.

Model fitting and comparison

Given any particular setting of the free parameters, when applied to the sequence of stimuli and outcomes actually experienced by the subjects, each model defines a trial-by-trial time series of CS values V_n, and, in the case of the hybrid model, also a second time series of associabilities or dynamic learning rates, $\alpha_n(x_n)$. In principal, either or both of these quantities might be reflected in the trial-by-trial skin conductance responses. Previous work, including on this dataset, suggested that SCRs might be encoding CS values V_n[1]. We thus focus our model fitting and comparison on the hypothesis that the inclusion of associabilities $\alpha_n(x_n)$ explains additional variance in SCRs, either indirectly through their effects on the values, V_n, learned by the hybrid model, or directly, with associability as an additional correlate explaining variance in SCRs, beyond that explained by value alone.

We examined both of these questions by comparing the fit of different models to the SCR data. We optimized the free parameters of each model to maximize the likelihood of the sequence of SCRs measured following the CS. We modeled the likelihood of each trial's SCR S_n as independent and identically distributed (i.i.d.) Gaussian distribution around a mean determined by the scaled value (or associability, or the combination of both value and associability) predicted by the model on that trial plus a constant term:

[1] $S_n \sim \mathrm{N}(\beta_0 + \beta_1 V_n(x_n), \sigma)$ or
[2] $S_n \sim \mathrm{N}(\beta_0 + \beta_1 \alpha_n(x_n), \sigma)$ or
[3] $S_n \sim \mathrm{N}(\beta_0 + \beta_1 V_n(x_n) + \beta_2 \alpha_n(x_n), \sigma)$

These are equivalent to linear regressions from the values or associabilities, or the combination of both, to the SCRs. For the Rescorla-Wagner model (RW(V)),we used V_n as the independent variable (since α is constant); for the Hybrid model, we tested all three possible combinations (Equations 1-3; Hybrid(V); Hybrid(α); Hybrid(α+V)), all in separate fits of all free parameters.

Likelihoods were pooled over all trials, but omitting trials in which a shock was delivered in order to avoid possible contamination of the predictive response by shock-related responses. (Although shock trials were omitted from the regression onto the SCRs, they were included in the computation of V_n and α_n.)

The key questions concerning whether associability impacts SCR can each be posed statistically in terms of the comparison of fit between a more complicated model, and a more restricted one that is nested within it via a parametric restriction. Thus, we used classical likelihood ratio tests of the null hypothesis that the improvement in fit of the more complicated model relative to the simpler one was better than that expected by chance (i.e., overfitting, given the additional parameters included in the more complicated model). For example, RW(V) is nested in the Hybrid(V) model by setting $\eta = 0$ and $\kappa = 1$ (the comparison tests the hypothesis that values learned with associability explain the SCRs better than values learned with a constant learning rate); additionally, Hybrid(V) and Hybrid(α) are nested in Hybrid(α+V) with β_1 or β_2 = 0 (these comparisons test whether the addition of either variable as an additional covariate improves fit relative to a model explaining SCRs as well as possible on the basis of the other variable only) respectively.

We fit these models separately to each individual subject's SCRs (i.e., using a separate set of free parameters for each subject), and performed likelihood ratio tests on the data likelihoods aggregated across subjects (**Supplementary Table 1**). Although the small number of trials limits power on the individual level, we verified that the aggregate results were supported in a reasonable proportion of individuals considered individually (In particular Hybrid(α+V) fit better than Hybrid(V) for 6/17 subjects at $p<.05$ and trended so at $p<.1$ for an additional 1/17; compared to RW(V), the full Hybrid(α+V) showed a significant improvement for 6/17 subjects and a trend for 3 more).

It's been shown previously that when free parameters are fit separately to each individual subject in this way, the resulting parametric regressors (prediction errors, etc.) are too noisy to achieve reliable fMRI effects[5-8]. Accordingly, we repeated all behavioral model fits taking the models' free parameters as fixed across subjects, which is a simple and effective way of regularizing the free parameters. The major

results (**Supplementary Table 2 & Fig. 1**) held up in this case; in particular, the comparisons between Hybrid(α+V) and Hybrid(V), and between Hybrid(V) and RW(V), confirmed the importance of including associability to explain SCRs. Indeed, in this case, the Hybrid(α+V) model was not a significantly better fit than than the simpler Hybrid(α) model. We thus adopted the parameters fit with the Hybrid(α) model for subsequent imaging analyses (**Supplementary Table 3)**.

Imaging acquisition and analysis. A 3T Siemens Allegra head-only scanner and Siemens standard coil (Siemens) were used for MRI data acquisition. Functional images were collected using a single-shot gradient echo EPI sequence (TR = 2000ms, TE = 25ms, FOV = 192cm, flip angle = 75^0, bandwidth = 4340 Hz/px, echo spacing = 0.29 ms) after T1-weighted (256 × 256 matrix, 176 1-mm sagittal slices) anatomical images were acquired. Thirty-nine contiguous oblique-axial slices (3 × 3 × 3 mm voxels) parallel to the AC-PC line were obtained.

Analysis of the imaging data was conducted using SPM8 (Wellcome Trust Center for Neuroimaging; http://www.fil.ion.ucl.ac.uk/spm/) and xjView (http://www.alivelearn.net/xjview8/). Motion effects were corrected by aligning images in each run to the first volume using a 6-parameter rigid body transformation. Mean functional images were then coregistered to the structural image, and normalized into MNI template space using a 12-parameter affine transformation (SPM8 "segment and normalize" estimated from the structural). Normalized functional images were re-sampled into 2 × 2 × 2 voxel resolution. A Gaussian kernel with a full width at half maximum (FWHM) of 6mm was applied for spatial smoothing.

For statistical analysis, we constructed two impulse events for each trial, at the times of CS onsets and CS offsets (potential US onsets). The first event was included to control the overall BOLD variance. To study parametric effects related to learning, we included three parametric regressors modulating the CS offset impulse event: (1) Associability $\alpha_n(x_n)$, (2) outcome identity r_n (1 = shock; 0 = no shock) and (3) prediction error $\delta_n(x_n)$ as parametric regressors modulating the outcome impulse event. The associability and prediction error time series were generated using the

hybrid model (Hybrid(α)) at the best fitting setting of the free parameters, and were the focus of our analysis (**Supplementary Fig. 3**). The associability and prediction error time series were not mutually correlated (R^2 = .009; **Supplementary Fig. 2**), allowing their effects to be examined separately. The outcome identity was included as a dummy variable. Note that we modeled both of these effects at the onset of outcome revelation rather than at the onset of CS since the prediction error is computed when the outcome information is available, it's also at the outcome period when prediction error is combined with the associability to influence the value update.

We then convolved all these regressors with SPM8's canonical hemodynamic response function, computed parameter estimates for each subject, and took these estimates to the group random effects level for statistical testing[9] (**Supplementary Tables 4 & 5**).

To examine activity patterns in the amygdala and striatum more closely, we defined ROIs using an independent contrast. Since previous work showed that both striatum and amygdala are highly active during the early acquisition trials (10 CS+ & 7 CS– trials)[1], we created a separate general linear model (GLM) which included box-car predictors that started with the onsets of CSs and lasted the length of each trial (4s) and additional predictors for trials terminating with a shock. Both regressors were convolved with a standard hemodynamic response function. We used a relatively loose threshold ($p < 0.005$, unc) to generate striatum and bilateral amygdala masks (CS vs. baseline). We extracted the average beta values for associability (α) and PE (δ) within these masks and conducted a repeated-measure ANOVA with factors of model component and ROI. Note that the test of an interaction by region cannot be biased by the selecting contrast, since the same contrast was used to define the masks in both regions.

References

1. Schiller, D., Levy, I., Niv, Y., LeDoux, J. E. & Phelps, E. A. From fear to safety and back: reversal of fear in the human brain. *J. Neurosci.* **28**, 11517-11525 (2008).
2. Morris, J. S., Ohman, A. & Dolan, R. J. Conscious and unconscious emotional learning in the human amygdala. *Nature* **393**, 467-470 (1998).
3. Sutton, R. *Adapting Bias by Gradient Descent: An Incremental Version of Delta-Bar-Delta* (In *Proceeding of Tenth National Conference on Artificial Intelligence*, MIT Press, 1992).
4. Le Pelley, M. E. The role of associative history in models of associative learning: a selective review and a hybrid model. *Q. J. Exp. Psychol. B* **57**, 193-243 (2004).
5. Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B. & Dolan, R. J. Cortical substrates for exploratory decisions in humans. *Nature* **441**, 876-879 (2006).
6. Schonberg, T., Daw, N. D., Joel, D. & O'Doherty, J. P. Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *J. Neurosci.* **27**, 12860-12867 (2007).
7. Schonberg, T. *et al*. Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. *Neuroimage* **49**, 772-781 (2010).
8. Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P. & Dolan, R. J. Model-based influences on humans' choices and striatal prediction errors. *Neuron* **69**, 1204-1215 (2011).
9. Friston, K. J., Frith, C. D., Frackowiak, R. S. & Turner, R. Characterizing dynamic brain responses with fMRI: a multivariate approach. *Neuroimage* **2**, 166-172 (1995).
10. Vul, E., Harris, C., Winkielman, P. & Pashler, H. Puzzlingly High Correlations in fMRI Studies of Emotion, Personality, and Social Cognition. *Perspectives on Psychological Science* **4**, 274-290 (2009).