Published in final edited form as:
*Nat Neurosci*. ; 14(10): 1250–1252. doi:10.1038/nn.2904.

# Differential roles of human striatum and amygdala in associative learning

**Jian Li**[1,2], **Daniela Schiller**[3], **Geoffrey Schoenbaum**[4,5], **Elizabeth A. Phelps**[1,2], and **Nathaniel D. Daw**[1,2]

[1]Psychology Department, New York University, New York, New York 10003

[2]Center for Neural Science, New York University, New York, New York 10003

[3]Departments of Psychiatry and Neuroscience, and Friedman Brain Institute, Mt. Sinai School of Medicine, New York, New York 10029

[4]Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland 20201

[5]Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland 20201

## Abstract

Although the human amygdala and striatum have both been implicated in associative learning, only the striatum's contribution has been consistently computationally characterized. Using a reversal learning task, we demonstrate that amygdala BOLD activity tracks associability as estimated by a computational model, and dissociates it from the striatal representation of reinforcement prediction error. These results extend the computational learning approach from striatum to amygdala, demonstrating their complementary roles in aversive learning.

Both the amygdala and striatum are known to be critical for associative learning. For the striatum, celebrated work in humans and other animals suggests its involvement in learning from prediction errors for reinforcement[1, 2]. Such errors occur when there is more or less reward (or punishment) than expected. Supporting this idea, the prediction error – as quantified in theories of conditioning such as the Rescorla-Wagner and temporal difference models – has helped to explain neural signaling in this system across species, including blood oxygenation level-dependent (BOLD) signals in the human striatum[2, 3].

However, BOLD activity in the amygdala has not consistently correlated with error signals, even in aversive conditioning tasks[3]. This raises the question, how might we computationally characterize learning signals in the amygdala? Such a specific characterization could shed further light on ideas about the structure's distinct contributions to associative learning. Current theories of amygdala function in humans have highlighted its role in vigilance[4] and the detection of relevant stimuli[5]. Theories of associative learning in animals – notably, the Pearce-Hall model[6] describe a more specific and potentially related function for the amygdala[7, 8]: the attentional gating of learning. These theories envision that,

Correspondence should be addressed to J.L. (lijian@nyu.edu).

AUTHOR CONTRIBUTIONS

E.A.P. and D.S. designed the study and conducted the experiment. J.L. and N.D.D. performed the data analysis. J.L., D.S., G.S., E.A.P. and N.D.D. interpreted the data and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

in order to learn cue-reinforcer associations, animals track a quantity – called associability – which reflects the extent to which each cue has previously been accompanied by surprise (positive or negative prediction errors). A cue's associability gates the amount of future learning about the cue based on whether it has been a reliable or poor predictor of reinforcement in the past. In other words, associability controls learning rates dynamically, accelerating learning to cues whose predictions are poor and decelerating it when predictions become reliable.

In non-human animals, lesion studies, and more recently, unit recordings suggest that a key neural substrate for associability is the amygdala[7–9]. To date, there is little direct evidence that the human amygdala might play an analogous role. Here, we hypothesize that the human amygdala codes for associability, which is distinct and complementary to the striatum's coding of prediction error during associative learning. Specifically, we used a computational model to examine an aversive reversal-learning task to ask whether an associability signal similar to that seen in unit recordings in non-human animals might be present in the pattern of BOLD signaling in the human amygdala during aversive learning[8].

Seventeen participants completed a Pavlovian reversal-learning task (Fig. 1a and Supplementary Methods) during which their BOLD signals and skin conductance responses (SCRs) were recorded simultaneously[10]. The experiment began with an acquisition phase, in which participants were presented with two visual stimuli (mildly angry or fearful faces; conditioned stimuli). One stimulus co-terminated with an aversive outcome (electric shock; unconditioned stimuli) on one-third of the trials (partially reinforced, CS+). The other stimulus was not paired with unconditioned stimuli (CS−). The acquisition phase was followed by an unsignaled reversal phase, in which the identities of original conditional stimuli (CS+ and CS−) switched[10]. Such a task provides a characteristic test for theories of associability, which predict that the associability of each conditioned stimulus should decline during acquisition, as the outcomes become more expected, and then increase rapidly during the reversal phase, when the outcomes are again surprising.

We first fit and validated our associability model behaviorally using a physiological measure, SCRs. Although previous work has demonstrated that SCRs correlate with cue-specific value (*V*) as predicted by a Rescorla-Wagner learning model[10], we hypothesized that these responses might reveal additional effects of associability. To test this, we compared the fit of alternative learning models to all participants' SCRs, correcting for the models' different numbers of free parameters using likelihood ratio tests (see Supplementary Methods and Supplementary Tables S1 and S2 for details). Indeed, compared to the basic Rescorla-Wagner model with a constant learning rate, value-related SCR effects were better explained by values predicted by an augmented "hybrid" Rescorla-Wagner model, which gated its learning rate dynamically according to the Pearce-Hall associability rule ($\chi^2_{34}$=104.42; $p < 0.00001$). Furthermore, since an arousal or attentional signal such as SCR might directly reflect associability (a measure of cue-specific attention) as well as value expectation, we tested whether SCR was modulated by the cue-specific associabilities learned by the model, over and above any value-related effects. This additional effect was significant ($\chi^2_{17}$=63.63; $p < 0.0001$). Both results support the hypothesis that the brain learns cue-specific associabilities and uses them to modulate predictive learning about potential aversive shocks.

In order to quantitatively identify the neural correlates of (aversive) prediction error (δ) and associability (α), we next used the fitted hybrid model to generate, for each subject, trial-by-trial time series of the estimates for δ and α. We regressed these variables on subjects' BOLD data at the time of conditioned stimuli termination (the time when, in the model, prediction error is realized and modulated by associability to gate learning; see

Supplementary Methods for details). These two time series are relatively easy to distinguish from one another, because the associability is determined not by the current prediction error, but instead by prediction errors received on previous trials with the same cue (Supplementary Figs. 1 and 2).

Based on lesion studies and electrophysiological recordings in non-human animals, we focused our search for associability-related activity on the amygdala[7–9]. We compared this activity to that of the striatum, which is associated with error-driven learning in both humans and other species[1, 2], including prediction errors for appetitive as well as aversive reinforcers[3]. As expected, BOLD activity in the bilateral ventral striatum, but not amygdala, was positively correlated with the aversive prediction error (Fig. 2a, $p < 0.05$, small volume corrected (SVC) for multiple comparisons within anatomically defined masks of the two structures). However, the opposite activation pattern emerged for associability, which was positively correlated with the bilateral amygdala, but not the ventral striatum (Fig. 2b, $p < 0.05$ SVC; see Supplementary Methods for additional discussion).

To further confirm that the striatum and amygdala were differentially engaged in representing prediction error and associability, we directly compared the mean activity within these areas (in regions defined functionally by the main effect of conditioned stimuli presentation vs. baseline during early acquisition[10], a contrast chosen so as not to bias the subsequent test for differential signaling between the regions, see Supplementary Methods). Specifically, we compared the effects of different components (α and δ) of learning signal across regions (striatum and amygdala) using a two-factor, repeated-measures ANOVA on the regression coefficients from individual subjects. We observed a significant interaction of region and model component ($F_{1, 64} = 5.75$, $p < 0.02$; Fig. 2c, note that this test does not require correction for multiple comparisons), indicating differential sensitivity to the two components (α and δ) across the two areas. In addition, a *post hoc t* test showed a larger correlation with α in the BOLD signals in bilateral amygdala than ventral striatum (paired t-test, $t_{16} = 3.03$, $p < 0.01$; Fig. 2c).

Though it has been associated with affective learning, trial-by-trial BOLD activity in the human amygdala has not consistently enjoyed a quantitative, computational interpretation comparable to prediction error in the striatum. The present results, taken together with more invasive techniques in non-human animals[7–9], are consistent with a specific functional role for the human amygdala in controlling associability during learning. This role would be complementary to prediction error signaling in mesolimbic dopamine targets, such as striatum, allowing increased processing of cues – and enhanced learning.

Our results also link work on the amygdala's role in associative learning in non-human animals with research in humans on cortical representations of uncertainty and their control over learning rates. Bayesian theories predict that several sorts of uncertainty should, jointly, determine learning rates, according to computations only approximated by the Pearce-Hall rule[11]. Correlates of such quantities have been reported in cingulate and insular cortices[12, 13], near areas where BOLD also correlated with associability in our analysis (Supplementary Table 4). We hypothesize that cortical uncertainty signals may reflect predecessor variables contributing to the computation of net associability in amygdala, since lesion studies support the amygdala's causal role as a key hub in learning rate control [7]. However, no single study has yet manipulated all of the different factors necessary to distinguish the many types of uncertainty that might contribute to associative learning.

In the current study, we extend the computational characterization of learning signals in the human brain from the striatum (prediction error) to the amygdala, which we found correlates with associability. Our results leave open the question whether associability coding in

human amygdala is specific to aversive tasks, or to other features of our experiment such as the use of mildly aversive (angry) faces as conditioned stimuli. However, our findings complement previous research using reward learning tasks in non-human animals showing similar roles for the amygdala and the striatum in the computation of associability and prediction error, respectively[8]. In the context of this animal literature, the present results suggest that what distinguishes these two value-learning regions may not be nature of the reinforcer, but rather the computational contribution to the learning signal[3, 14, 15].

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We thank P. Glimcher, R. Rutledge and E. DeWitt for discussions and comments. This research was supported by a McKnight Foundation Scholar Award and NIH MH087882, part of the CRCNS program to ND; a James S. McDonnell Foundation grant and NIH MH080756 to EAP; a NIH DA015718 and AG027097 to GS; and a Blavatnik award to DS. This work was also supported by a Seaver Foundation grant to the Center for Brain Imaging (CBI).

## References

1. Schultz W, Dayan P, Montague PR. Science. 1997; 275:1593–1599. [PubMed: 9054347]
2. O'Doherty JP, Dayan P, Friston K, Critchley H, Dolan RJ. Neuron. 2003; 38:329–337. [PubMed: 12718865]
3. Delgado MR, Li J, Schiller D, Phelps EA. Philos Trans R Soc Lond B Biol Sci. 2008; 363:3787–3800. [PubMed: 18829426]
4. Davis M, Whalen PJ. Mol Psychiatry. 2001; 6:13–34. [PubMed: 11244481]
5. Phelps, EA. The Human Amygdala. Whalen, P.; Phelps, E., editors. Guilford Press; New York: 2009. p. 204-219.
6. Pearce J, Hall G. Psychol Rev. 1980:532–552. [PubMed: 7443916]
7. Holland P, Gallagher M. Trends Cogn Sci. 1999; 3:65–74. [PubMed: 10234229]
8. Roesch MR, Calu DJ, Esber GR, Schoenbaum G. J Neurosci. 2010; 30:2464–2471. [PubMed: 20164330]
9. Belova MA, Paton JJ, Morrison SE, Salzman CD. Neuron. 2007; 55:970–984. [PubMed: 17880899]
10. Schiller D, Levy I, Niv Y, LeDoux JE, Phelps EA. J Neurosci. 2008; 28:11517–11525. [PubMed: 18987188]
11. Courville AC, Daw ND, Touretzky DS. Trends Cogn Sci. 2006; 10:294–300. [PubMed: 16793323]
12. Preuschoff K, Bossaerts P. Ann N Y Acad Sci. 2007; 1104:135–146. [PubMed: 17344526]
13. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS. Nat Neurosci. 2007; 10:1214–1221. [PubMed: 17676057]
14. Robbins TW, Cador M, Taylor JR, Everitt BJ. Neurosci Biobehav Rev. 1989; 13:155–162. [PubMed: 2682402]
15. Baxter MG, Murray EA. Nat Rev Neurosci. 2002; 3:563–573. [PubMed: 12094212]

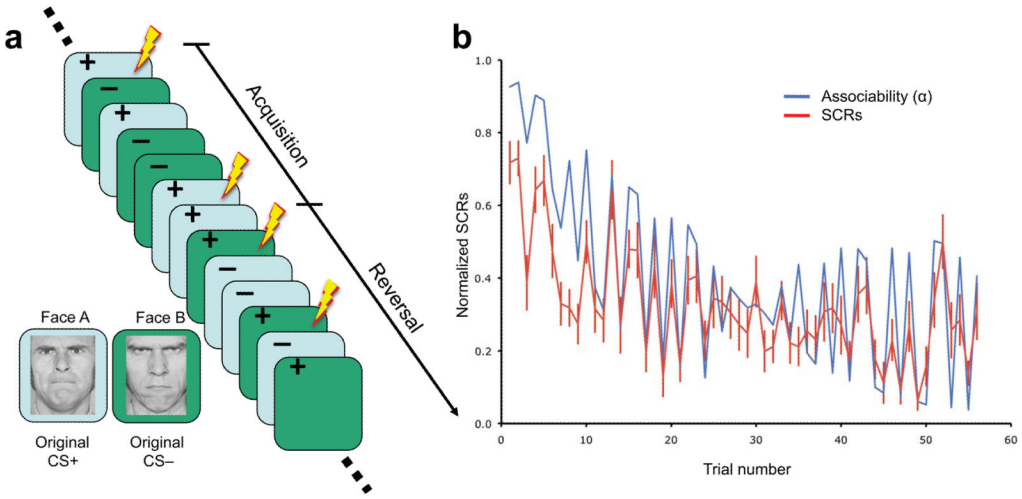


**Figure 1.**

Experimental design and behavioral model fit. (**a**) Experiment timeline illustration. Acquisition phase consisted of presentations of the CS+ which was partially associated with electric shock and CS− not associated with shock. In the reversal phase, the reinforcement contingencies for the original CS+ and CS− were switched. (**b**) Average SCRs across subjects (red) and the best-fit associability trace (blue).

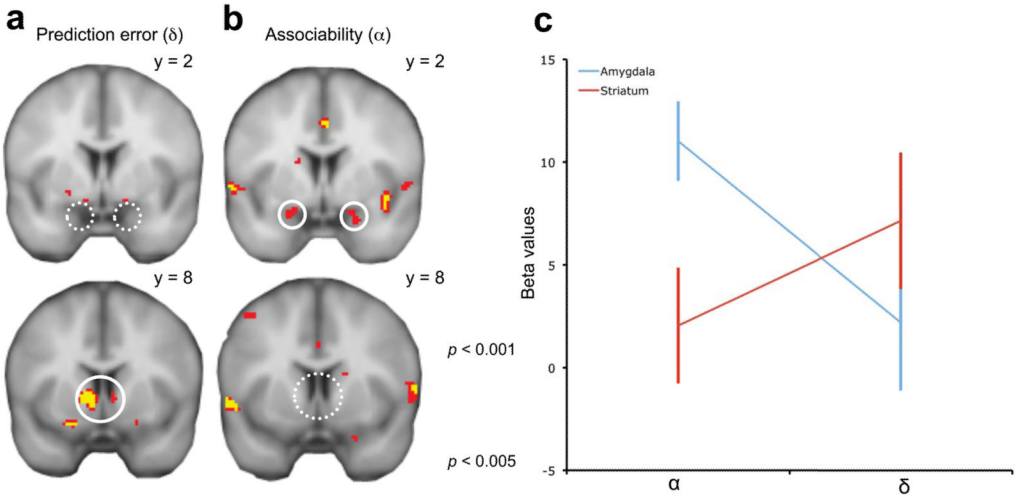


**Figure 2.**

Neural correlates of associability and prediction error term. (**a**) BOLD activities in the ventral striatum but not amygdala correlate with prediction error. (**b**) BOLD activities in the bilateral amygdala but not ventral striatum correlate with associability regressor ($p < 0.05$, SVC; results are shown at uncorrected thresholds to display the full extent of the activation. (**c**) Differential representations of associability (**α**) and prediction error (**δ**) in striatum and amygdala BOLD activities (± s.e.m.) identified by an two-way ANOVA.