nature neuroscience

Differential roles of human striatum and amygdala in associative learning

Jian Li[1,2], Daniela Schiller[3], Geoffrey Schoenbaum[4–6], Elizabeth A Phelps[1,2] & Nathaniel D Daw[1,2]

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 found that amygdala blood oxygen level–dependent activity tracked associability as estimated by a computational model, and dissociated 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 is not consistently correlated with error signals, even in aversive conditioning tasks[3]. This raises the question of how we might 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, such as 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, to learn cue-reinforcer associations, animals track a quantity, known as associability, that 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 on the basis of 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 nonhuman animals, lesion studies and, more recently, unit recordings have indicated that an important neural substrate for associability is the amygdala[7–9]. To date, there is little direct evidence that the human amygdala might have an analogous role. We hypothesized 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 and asked whether an associability signal similar to that seen in unit recordings in nonhuman animals might be present in the pattern of BOLD signaling in the human amygdala during aversive learning[8].

We asked 17 participants to complete a Pavlovian reversal-learning task (**Fig. 1a** and **Supplementary Methods**) and simultaneously recorded their BOLD signals and skin conductance responses (SCRs)[10]. The experiment began with an acquisition phase, in which participants were presented with two visual stimuli (mildly angry faces, conditioned stimulus). One stimulus co-terminated with an aversive outcome (electric shock, unconditioned stimulus) on one-third of the trials (partially reinforced). The other stimulus was not paired with an unconditioned stimulus. The acquisition phase was followed by an unsignaled reversal phase, in which the identities of the original conditioned and unconditioned stimuli were switched[10]. This 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 SCRs (**Fig. 1b**). Although previous work has found 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 1** and **2**). Indeed, compared with the basic Rescorla-Wagner model using 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, given that an arousal or attentional signal such as a SCR might directly reflect associability (a measure of cue-specific attention) as well as value expectation, we tested whether SCRs were modulated by the cue-specific associabilities learned by the model, over and above any value-related effects. This additional effect was

[1]Department of Psychology, New York University, New York, New York, USA. [2]Center for Neural Science, New York University, New York, New York, USA. [3]Departments of Psychiatry and Neuroscience, and Friedman Brain Institute, Mt. Sinai School of Medicine, New York, New York, USA. [4]Department of Anatomy and Neurobiology, University of Maryland School of Medicine, Baltimore, Maryland, USA. [5]Department of Psychiatry, University of Maryland School of Medicine, Baltimore, Maryland, USA. [6]National Institute on Drug Abuse Intramural Research Program, Baltimore, Maryland, USA. Correspondence should be addressed to J.L. (lijian@nyu.edu).

Received 12 April; accepted 7 July; published online 11 September 2011; doi:10.1038/nn.2904

significant ($\chi^2_{17} = 63.63$, $P < 0.0001$). Both of these results support the hypothesis that the brain learns cue-specific associabilities and uses them to modulate predictive learning about potential aversive shocks.

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; **Supplementary Methods**). These two time series were relatively easy to distinguish from one another, as the associability was determined not by the current prediction error, but instead by prediction errors received on previous trials with the same cue (**Supplementary Figs. 1–3**).

On the basis of lesion studies and electrophysiological recordings in nonhuman animals, we focused our search for associability-related activity on the amygdala[7–9]. We compared amygdala 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 both appetitive and aversive reinforcers[3]. As expected, BOLD activity in the bilateral ventral striatum, but not in the amygdala, was positively correlated with the aversive prediction error ($P < 0.05$, small volume corrected (SVC) for multiple comparisons within anatomically defined masks of the two structures; **Fig. 2a**). However, the opposite activation pattern emerged for associability, which was positively correlated with the bilateral amygdala, but not the ventral striatum ($P < 0.05$, SVC; **Fig. 2b** and **Supplementary Methods**).

To further confirm that the striatum and amygdala were differentially engaged in representing prediction error and associability, we directly compared the mean activity in these areas (in regions defined functionally by the main effect of conditioned stimuli presentation versus 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 ($F_{1,64} = 5.75$, $P < 0.02$, note that this test does not require correction for multiple comparisons; **Fig. 2c**), 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 in ventral striatum (paired *t*-test, $t_{16} = 3.03$, $P < 0.01$; **Fig. 2c**).

Although 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 that of prediction error in the striatum. Our results, taken together with more invasive techniques in nonhuman 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 nonhuman 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 signals 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, as the results of lesion studies

-

(**Supplementary Tables 4** and **5**). These results leave open the question of 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 that used reward learning tasks in nonhuman animals and found similar roles for the amygdala and the striatum in the computation of associability and prediction error, respectively[8]. In the context of these findings, our results suggest that what distinguishes these two value-learning regions may not be the nature of the reinforcer, but rather the computational contribution to the learning signal[3,14,15].

Note: Supplementary information is available on the Nature Neuroscience website.

ACKNOWLEDGMENTS

We thank P. Glimcher, R. Rutledge and E. DeWitt for discussions and comments. This research was supported by a McKnight Foundation Scholar Award, Human Frontiers Science Program grant RGP0036/2009-C, US National Institutes of Health (NIH) grant MH087882 (part of the CRCNS program, to N.D.D.), a James S. McDonnell Foundation grant and NIH grant MH080756 to E.A.P., and NIH grants DA015718 and AG027097 to G.S. This work was also supported by a Seaver Foundation grant to the Center for Brain Imaging.

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.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Schultz, W., Dayan, P. & Montague, P.R. *Science* **275**, 1593–1599 (1997).
2. O'Doherty, J.P., Dayan, P., Friston, K., Critchley, H. & Dolan, R.J. *Neuron* **38**, 329–337 (2003).
3. Delgado, M.R., Li, J., Schiller, D. & Phelps, E.A. *Phil. Trans. R. Soc. Lond. B* **363**, 3787–3800 (2008).
4. Davis, M. & Whalen, P.J. *Mol. Psychiatry* **6**, 13–34 (2001).
5. Phelps, E.A. in *The Human Amygdala* (eds. Whalen, P. & Phelps, E.) 204–219 (Guilford Press, New York, 2009).
6. Pearce, J.M. & Hall, G. *Psychol. Rev.* **87**, 532–552 (1980).
7. Holland, P.C. & Gallagher, M. *Trends Cogn. Sci.* **3**, 65–73 (1999).
8. Roesch, M.R., Calu, D.J., Esber, G.R. & Schoenbaum, G. *J. Neurosci.* **30**, 2464–2471 (2010).
9. Belova, M.A., Paton, J.J., Morrison, S.E. & Salzman, C.D. *Neuron* **55**, 970–984 (2007).
10. Schiller, D., Levy, I., Niv, Y., LeDoux, J.E. & Phelps, E.A. *J. Neurosci.* **28**, 11517–11525 (2008).
11. Courville, A.C., Daw, N.D. & Touretzky, D.S. *Trends Cogn. Sci.* **10**, 294–300 (2006).
12. Preuschoff, K. & Bossaerts, P. *Ann. NY Acad. Sci.* **1104**, 135–146 (2007).
13. Behrens, T.E.J., Woolrich, M.W., Walton, M.E. & Rushworth, M.F.S. *Nat. Neurosci.* **10**, 1214–1221 (2007).
14. Robbins, T.W., Cador, M., Taylor, J.R. & Everitt, B.J. *Neurosci. Biobehav. Rev.* **13**, 155–162 (1989).
15. Baxter, M.G. & Murray, E.A. *Nat. Rev. Neurosci.* **3**, 563–573 (2002).