2019

Neural computations of threat in the aftermath of combat trauma

P. Homan
Zucker School of Medicine at Hofstra/Northwell

I. Levy

E. Feltham

C. Gordon

J. Hu

See next page for additional authors

Published in final edited form as:

Nat Neurosci. 2019 March ; 22(3): 470–476. doi:10.1038/s41593-018-0315-x.

Neural computations of threat in the aftermath of combat trauma

Philipp Homan[1], **Ifat Levy**[2], **Eric Feltham**[3,4], **Charles Gordon**[3,4], **Jingchu Hu**[1], **Jian Li**[5], **Robert H. Pietrzak**[3,4], **Steven Southwick**[3,4], **John H. Krystal**[3,4], **Ilan Harpaz-Rotem**[3,4,7,*], **Daniela Schiller**[1,6,7,*]

[1]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[2]Departments of Comparative Medicine, Neuroscience and Psychology, Yale University, New Haven, CT, USA.

[3]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA.

[4]US Department of Veterans Affairs National Center for PTSD, Clinical Neurosciences Division, VA Connecticut Healthcare System, West Haven, CT, USA.

[5]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China.

[6]Department of Neuroscience and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[7]These authors contributed equally: Ilan Harpaz-Rotem, Daniela Schiller.

Abstract

By combining computational, morphological, and functional analyses, this study relates latent markers of associative threat learning to overt post-traumatic stress disorder (PTSD) symptoms in combat veterans. Using reversal learning, we found that symptomatic veterans showed greater physiological adjustment to cues that did not predict what they had expected, indicating greater sensitivity to prediction errors for negative outcomes. This exaggerated weighting of prediction errors shapes the dynamic learning rate (associability) and value of threat predictive cues. The

[*]**Correspondence and requests for materials** should be addressed to I.H. ilan.harpaz-rotem@yale.edu or D.S. daniela.schiller@mssm.edu.

Author contributions

I.L., I.H.R., and D.S. designed the study. E.F., C.G., I.L., and I.H.R. collected the data. J.H. scored the data. P.H. analyzed the data. I.L., J.L, I.H.R., and D.S. contributed to data analysis. J.H.K., R.H.P., and S.S. contributed to the interpretation of the results. P.H., I.L., I.H.R., and D.S. wrote the first draft of the manuscript. All authors contributed to the final version of the manuscript.

Code availability

The code used for the analyses is available online at: http://osf.io/rxsw2/.

Data availability

Data used to support the conclusions of this study is available online at: http://osf.io/rxsw2/.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41593-018-0315-x.

Competing interests

The authors report no competing interests.

Supplementary information is available for this paper at https://doi.org/10.1038/s41593-018-0315-x.

degree to which the striatum tracked the associability partially mediated the positive correlation between prediction-error weights and PTSD symptoms, suggesting that both increased prediction-error weights and decreased striatal tracking of associability independently contribute to PTSD symptoms. Furthermore, decreased neural tracking of value in the amygdala, in addition to smaller amygdala volume, independently corresponded to higher PTSD symptom severity. These results provide evidence for distinct neurocomputational contributions to PTSD symptoms.

Reporting Summary.

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

On returning from combat, why do some military personnel develop symptoms of PTSD and others do not? PTSD symptoms may develop after exposure to a traumatic event and are characterized by symptoms of re-experiencing, avoidance of trauma reminders, negative alterations in cognitions and mood, and alterations in arousal and reactivity[1–3]. A prominent learning theory suggests that PTSD symptoms largely reflect maladaptive associative learning during and after a traumatic event[4]. Associative learning of threat[5] is the process by which benign stimuli such as people, locations, and objects (that is, conditioned stimuli) acquire threatening properties through pairing with an aversive outcome, and have the capacity to trigger and maintain defensive responses well after the aversive event is decoupled from the conditioned stimuli. Although abnormal threat conditioning features prominently in theoretical accounts of PTSD, the manner in which learning becomes dysfunctional is less clear[4].

Accumulating evidence suggests a variety of impaired learning processes in PTSD, including overgeneralization, heightened contextual anxiety, diminished inhibition in response to safety cues, and failure to retain extinction learning[4]. These findings link PTSD to basic learning processes, but they do not disambiguate specific aspects of learning that may contribute to the disorder, such as the learning rate or the computation of aversive value. It is possible that PTSD-related abnormalities are influenced by learning parameters that we cannot directly observe, but are able to infer from observable behavior. Computational indices, which estimate such latent learning parameters, may be able to detect such differences.

Theories of associative learning, such as the Pearce-Hall learning mechanism[6], envision that learning cue-outcome associations involves tracking of several quantities: prediction errors for reinforcement, which occur when the outcome is more or less than expected (that is, surprising), and associability, reflecting the extent to which each cue has been previously accompanied by surprise. The value assigned to cues in the environment is revised in each encounter based on the prediction error. Associability dynamically guides value learning by accelerating it to cues whose predictions are poor (large prediction errors), and decelerating it when predictions become reliable. Here, we used a hybrid version of the Pearce-Hall learning model to estimate the computations performed during associative threat learning[7–9] and how the behavioral and neural tracking of these computations relate to PTSD symptom severity.

To cover the full spectrum of symptomatology, we recruited 54 combat-exposed veterans with a wide range of PTSD symptoms based on the gold-standard structured clinical interview for PTSD, the Clinician-Administered PTSD Scale (CAPS; Table 1 and Supplementary Fig. 1). Twenty-four participants had a diagnosis of PTSD and 30 participants were combat veteran controls without PTSD diagnosis. We used the threat reversal paradigm—where flexible updating of threat responses is required—together with computational modeling, to uncover latent learning parameters that are relevant for the symptomatology.

We expected that the observed threat learning behavior would be similar across different levels of symptoms (reflecting the unspecific and subtle aberrations found in threat response conditioning in PTSD in general[10,11]) but that the underlying neural computations might reveal disease-related differences.

The amygdala is a locus of associative learning in the brain[7,12], and previous work has linked PTSD symptoms with abnormal amygdala structure[13], as well as heightened amygdala reactivity to stimuli laden with emotionally negative content[14,15]. Our goal here was thus to examine whether the structural and functional implementations of specific learning computations in the amygdala relate to PTSD symptoms, and whether the threat learning-related function and volume of the amygdala contribute to PTSD symptoms in a complementary manner.

The experiment began with an acquisition phase, in which two visual stimuli (mildly angry faces) were presented consecutively in a pseudorandomized order. One of the stimuli was paired with a mild electric shock on one-third of the trials (face A), while the other was never paired with the shock (face B). The acquisition phase was immediately followed by a reversal phase, in which the contingencies were flipped such that the formerly neutral stimulus (face B) was now paired with a shock and face A became the neutral stimulus (Fig. 1a). Skin conductance response (SCR) served as the index of conditioned defensive responses.

Results

Irrespective of symptoms, veterans show successful reversal learning.

Combat-exposed veterans ($N = 54$ participants) successfully acquired and reversed threat conditioning, as assessed by the differential SCR (face A versus face B) in the two phases of the task (Fig. 1b). To test for a potential relationship between threat reversal and PTSD symptoms, we used a linear regression with threat reversal index as predictor and CAPS scores as the outcome. Reversal index was calculated by subtracting stimulus discrimination in reversal (that is, face A minus face B) from stimulus discrimination in acquisition (Fig. 1b). Controlling for irrelevant variables (age and gender), the regression revealed no significant relationship between symptoms and reversal learning ($\beta = 0.02$, $t(50) = 0.13$, two-tailed, $P = 0.894$). We also did not find evidence that PTSD symptoms were related to stimulus discrimination during threat acquisition only ($\beta = 0.03$, $t(52) = 0.22$, two-tailed, $P = 0.827$) or during the reversal phase only ($\beta = 0.02$, $t(52) = 0.12$, two-tailed, $P = 0.901$). Additional ways of categorizing veterans as highly and mildly affected did not reveal any

significant results (see Methods, ‘Sample characteristics). These results motivate the use of a computational approach that could potentially reveal latent learning differences across individuals exposed to combat trauma.

Pearce-Hall hybrid model best describes conditioned threat responses.

To estimate parameter weights for the specific computations performed during associative threat learning[7–9] and how they relate to PTSD symptom severity, we used a hybrid Rescorla-Wagner and Pearce-Hall model, which we have previously employed with the same task in healthy participants[7,8]. The computational model was informed by the Pearce-Hall learning mechanism for associability-gated learning[6]. Like the classic Rescorla-Wagner model[16], the hybrid model updates the value of each cue on each presentation of that cue, based on the discrepancy between the expected and obtained outcome, or the prediction error. The hybrid model, however, replaces the constant learning rate of the Rescorla-Wagner model by a dynamic associability parameter[6]. Associability reflects the attention that a cue receives on the basis of how accurately it has predicted outcome in the past. Unreliable cues receive more attention (higher associability) as they are likely to be unreliable in the future; and since they are unreliable, they should be updated preferentially as new information becomes available[17] (see Supplementary Material for details and for simulated parameter recovery as well as model fits and Supplementary Figs. 2 and 3).

First, to verify the suitability of the model, we conducted model comparison between several versions of reinforcement learning models. Using hierarchical Bayesian modeling we fitted three different versions of this hybrid model to the SCR data; all three outperformed the simpler Rescorla-Wagner model (deviance information criterion, DIC: 6003.91). In addition, the hybrid model with associability (α) and an additional predictor for value (V) updating (DIC: 2630.37) outperformed the models with either value alone (DIC: 2678.75) or associability alone (DIC: 2661.6) and was thus the winning model (Fig. 2a). There was no evidence that an additional scaling parameter for the reversal stage (reflecting a different prediction-error weight for the reversal stage) improved the model fit (DIC: 2772.27). Notably, similar results were obtained when using maximum likelihood estimation (MLE) as in a previous study (ref.[8]; see Supplementary Material for details and Supplementary Fig. 4). These findings indicate that the recorded SCRs during reversal learning reflect value expectations modulated by cue-specific attention. Next, we used this winning hybrid ($\alpha + V$) model to examine whether learning parameters that describe behavior and neural activity relate to PTSD symptom severity.

Symptomatic veterans assign higher weights to prediction errors.

To understand how the model computations relate to overt PTSD symptoms, we used the best-fit model parameters. In the winning hybrid model, the prediction-error weight η, which can be seen as a learning rate for associability, is a quantity estimated for each participant from the SCR. The prediction-error weight quantifies how much weight is assigned to wrong predictions when updating trial-by-trial associability. It is possible that more symptomatic combat veterans would be more sensitive to prediction errors, and will assign higher weights to them. Indeed, we found that higher prediction-error weight was associated with higher CAPS symptoms ($\beta = 0.55$, $t(50) = 4.57$, two-tailed, $P < 0.001$; Fig.

2b; note that this association held up when using a non-parametric rank correlation test that is more robust to outliers, see Supplementary Material). This finding suggests that highly symptomatic combat veterans were more influenced by prediction errors, weighing them more strongly as they adjusted trial-by-trial attention to cues.

Symptomatic veterans show altered amygdala value computation.

During the reversal task, the value assigned to each cue is continuously updated on the basis of associability-gated prediction error. Mathematically, a value in a current trial reflects the value in the previous trial plus prediction error multiplied by associability. Associability in each trial is updated by the weighted prediction error in the previous trial (see Supplementary Material for details). As reported above, the prediction error weight was positively associated with PTSD symptoms. As the weighted prediction error shapes value, we next examined whether the neural tracking of value related to PTSD symptoms.

We focused our neural investigation on the amygdala, given its role in associative learning[7,12], value encoding[18–21,] and evidence linking PTSD symptoms with heightened amygdala reactivity to emotionally negative stimuli[14,15]. Given that amygdala morphology has also been linked with stress-related psychopathology[13,22], we examined whether amygdala neural computations and morphology are different manifestations of the same source problem (that is, redundant) or whether they incrementally explain variance in PTSD symptoms.

To address this, we calculated linear regression models including functional (value encoding based on the winning hybrid model) and structural indices for amygdala as predictors of the PTSD symptoms (for a similar analysis using the classic Rescorla-Wagner model, see Supplementary Figs. 5–7). To account for unspecific intersubject variability, these models were adjusted for age, gender, head movement and total intracranial volume (see also Supplementary Material). We found a structure-function relationship with CAPS in the right amygdala (Fig. 3a), where both volume ($\beta = -0.52$, $t(47) = -2.7$, two-tailed, $P = 0.01$); Fig. 3b) and neural activity ($\beta = -0.29$, $t(47) = -2.02$, two-tailed, $P = 0.049$; Fig. 3c) independently predicted the total CAPS score. In the left amygdala, the effect of value-dependent activity remained significant when including amygdala volume in the same model ($\beta = -0.34$, $t(47) = -2.34$, two-tailed, $P = 0.024$; Fig. 3d), but no independent effect for volume emerged ($\beta = -0.25$, $t(47) = -1.27$, two-tailed, $P = 0.211$).

We verified that the findings were comparable when restricting the study sample to the male participants and when using non-parametric rank correlations (see Supplementary Material). We also confirmed that individual differences in right amygdala volumes did not affect the effect of neural activity on CAPS symptoms (see Supplementary Material and Supplementary Fig. 8).

To further characterize the relationship between structure and function we added the interaction term to the model and found that there was no evidence for a synergistic effect between these independent variables (right amygdala: $\beta = -0.95$, $t(46) = -0.88$, two-tailed, $P = 0.385$; left amygdala: $\beta = -0.91$, $t(46) = -0.78$, two-tailed, $P = 0.437$). However, the correlation between structure and function (adjusting for total head volume) was significant

and negative (right amygdala: $\beta = -0.39$, $t(49) = -2.11$, two-tailed, $P = 0.04$; left amygdala: $\beta = -0.39$, $t(49) = -2.13$, two-tailed, $P = 0.038$). A possible explanation is a compensatory recruitment of amygdala neurons in veterans with smaller amygdala volumes, probably due to a stress-related gray matter reduction[23].

To fully characterize the brain-behavior relationship with respect to symptoms, we tested whether individual differences in prediction-error weights were associated with differences in amygdala volume. We entered the prediction-error weight (η) as an outcome measure into a linear regression and used amygdala volume as predictor, including additional regressors for age, gender and total intracranial volume. We did not find evidence that right amygdala volume ($\beta = -0.22$, $t(49) = -1.14$, two-tailed, $P = 0.26$) or left amygdala volume ($\beta = -0.19$, $t(49) = -1$, two-tailed, $P = 0.321$) were associated with prediction-error weight.

In addition to value computation, the winning hybrid model also captures prediction error and associability, both of which are associated with amygdala neural activity[8,9,17]. Since they are not strongly correlated in the hybrid model[8,24], they can be assessed separately (see also Supplementary Material and Supplementary Fig. 9). We therefore computed a second general linear model (GLM) with trial-by-trial regressors for associability, shock occurrence, and prediction error, all of which were parametric modulators of cue offset, as this is the time point when prediction error and associability are computed. We expected that tracking of associability in the amygdala[8,17], reflecting the proposed attention-gating role of this brain region, would be attenuated by PTSD symptoms. However, we did not find evidence for a relationship between amygdala neural activity and PTSD symptoms for either associability (left: $\beta = -0.14$, $t(47) = -1.01$, two-tailed, $P = 0.316$; right: $\beta = -0.06$, $t(47) = -0.43$, two-tailed, $P = 0.667$) or prediction error (left: $\beta = -0.03$, $t(47) = -0.2$, two-tailed, $P = 0.839$; right: $\beta = 0.04$, $t(47) = 0.28$, two-tailed, $P = 0.781$), suggesting that amygdala value encoding contributes to the symptoms of PTSD, whereas associability and prediction error were less influential.

All together, these findings show that lower neural tracking of value in the amygdala, in addition to smaller amygdala volumes, corresponded to higher PTSD symptom severity.

Additional brain regions tracking threat computations.

The striatum, the hippocampus and the dorsal anterior cingulate cortex (dACC) have also been implicated in the computations related to threat learning[8,9,25]. We extended our analysis to these brain regions and tested whether neural tracking of value, associability and prediction error in these regions correlated with PTSD symptoms. Using a linear mixed model with brain region and CAPS as factors and neural value computations as an outcome, we found a main effect of CAPS ($F(1,52) = 5.49$, $P = 0.023$) as well as an interaction of brain region and CAPS ($F(2, 104) = 3.14$, $P = 0.047$), driven by significant negative correlations between value tracking in the striatum (Fig. 4). These results suggest that, similar to amygdala, lower value tracking in the striatum (but not hippocampus or dACC) relates to higher symptom severity.

To test for a relationship between PTSD symptoms and neural tracking of associability and prediction error in these regions, all of which have been implicated in prediction error[26,27]

and assocability[9,25,28,29] encoding, we computed a linear mixed model with brain region, learning component and CAPS as factors, and neural activity as the dependent variable. We found an interaction of learning component and CAPS ($F(1,208) = 20.43$, $P < 0.001$), driven by negative correlations between neural tracking of associability and CAPS that were attenuated for prediction error in all three regions (Fig. 4). We confirmed that these findings for value, associability and prediction-error computation were robust to the gender imbalance, clinical heterogeneity and medication status (see Supplementary Material). We also verified that the correlations were present when testing for non-parametric rank correlations. These results indicate that the lower tracking of associability (and less so of prediction error) in the striatum, hippocampus and dACC relate to higher symptom severity.

Finally, to investigate a dissociation of associability and prediction error in amygdala and striatum as reported in a previous study[8], we tested for an interaction of region (amygdala, striatum) and learning component (associability, prediction error). To improve comparability between the current and the previous study, we ran this analysis only in veterans without a diagnosis of PTSD, and found no evidence for dissociation ($F(3, 203) = 0.58$, $P = 0.629$). We also did not find evidence that the amygdala tracked associability in the current study in veterans without PTSD.

The absence of a dissociation that was found in a previous study[8] merits an explanation. First, it is noteworthy that the current study does replicate the computational results of the previous study[8], namely the superiority of the hybrid model over the Rescorla-Wagner model. On the neural level, the previous study found an interaction of region (amygdala, striatum) and learning component (associability, prediction error) that had a medium to large effect size (Cohens $d = 0.66$, 95% confidence interval (CI): 0.12; 1.17, $t(16) = 2.71$, one-sample t-test, two-tailed, $P = 0.02$). In the current study, we found that this interaction was not significant (Cohens $d = -0.13$, 95% CI: −0.49; 0.23, $t(29) = -0.7$, one-sample t test, two-tailed, $P = 0.49$). Rather, the striatum, but not amygdala, tracked associability in addition to tracking prediction error.

Several factors could explain this result. First and foremost, the current study's population was exposed to combat trauma, therefore meeting criterion A (exposure to a traumatic event) in the clinical assessment of PTSD symptoms and, in addition, was exposed to chronic stress associated with a deployment to combat zone. One may speculate that this traumatic stress (which has been shown to affect amygdala functioning[13,22,30]) may be the root cause for a shift in tracking from the amygdala to the striatum as part of brain plasticity. Second, the current sample differed significantly from the previous sample in terms of gender ratio (M/F = 49/5 in the current study versus 9/8 in Li and colleagues; $P = 0.001$) and age range (20–52 in the current study versus 18–31 in Li and colleagues[8]). Alternatively, given that we ran this analysis only in veterans without a diagnosis of PTSD ($N = 30$), a lack of statistical power might have contributed to the non-replication.

Brain-behavior relationship.

Prediction-error weights shape the computations of value and associability. The neural underpinnings of higher prediction-error weights, observed in the behavior of individuals with more PTSD symptoms, may therefore relate to computations of value in the amygdala

and the striatum, as well as to computations of associability in the striatum, dACC and hippocampus. While PTSD symptoms correlated positively with prediction-error weights, however, they correlated negatively with the neural tracking of value and associability. To better understand these inverse relationships we conducted a mediation analysis. This analysis revealed that the correlation between prediction-error weight and CAPS was partially mediated by the tracking of associability (but not prediction error) in the right striatum, as shown in the four steps of a mediation analysis (Fig. 5). We found that: (1) prediction-error weight positively correlated with CAPS ($\beta = 0.54$, $t(52) = 4.62$, two-tailed, $P < 0.001$); (2) prediction-error weight negatively correlated with neural activity ($\beta = -0.29$, $t(52) = -2.22$, two-tailed, $P = 0.03$); (3) neural activity negatively correlated with CAPS ($\beta = -0.43$, $t(52) = -3.43$, two-tailed, $P = 0.001$); and (4) prediction-error weight (eta) and neural activity independently predicted CAPS (eta: $\beta = 0.45$, $t(51) = 3.89$, two-tailed, $P < 0.001$; neural activity: $\beta = -0.3$, two-tailed, $t(51) = -2.55$, $P = 0.014$). Finally, we tested whether the difference between paths c and c' was significantly different from zero. To maximize statistical power, we used non-parametric bootstrapping with 5,000 draws to derive an empirical null distribution. We found that the difference between the effect with the mediator present and the effect without it (paths c and c'; Fig. 5) was indeed significant ($\beta = 0.09$, 95% CI: 0.02; 0.2; $P = 0.01$). This result indicates that our mediation model supports a significant partial mediation.

This suggests that—at least for the right striatum and associability—both higher prediction-error weights and decreased neural tracking of associability are independently related to higher CAPS symptoms. Speculatively then, the higher weight assigned to prediction errors might be a compensatory adjustment for the decreased neural tracking of associability. We did not find evidence that the neural tracking in any other region fully or partially mediated the relationship between prediction-error weights and CAPS.

Together, these findings indicate that the effect of higher prediction-error weights in individuals with higher CAPS scores was complemented by decreased striatal activity during associability computation.

Discussion

The current study found that even highly affected combat veterans were able to perform reversal learning when the SCRs were analyzed using conventional summary statistics. A more fine-grained computational analysis, however, revealed that subtle differences in latent learning components are at play: symptomatic veterans assigned more weight to prediction errors. An intuitive way of interpreting this result is in terms of attention. Highly affected individuals were more sensitive when their predictions about outcomes were wrong, and they exaggerated their adjustment to the cues that did not predict what they had expected. This behavior may be associated with the increased aversion to ambiguous losses, which was recently observed in PTSD in the context of economic decision-making. Future research will need to determine the exact relationships between decision making under uncertainty, reinforcement learning, and post-trauma symptomatology[31,32].

On the neural level, we found that the neural computations that were shaped by these altered prediction-error weights contributed to the symptoms of PTSD: aversive value encoding in the amygdala and striatum, and associability computations in the striatum, dACC, and hippocampus. Our study further indicates that the right amygdala computations contribute to the symptomatology above and beyond the effects of smaller amygdala volumes[13], suggesting additive effects of right amygdala volume and function. A model-based functional magnetic resonance imaging (fMRI) analysis such as the one used in this study can therefore not only indicate where in the brain a certain task-related activity emerges, but also which computations are probably performed.

The implication of these findings for PTSD becomes clear when the absence of behavioral differences (as indexed by reversal learning) is considered: as is well known from the behavioral (and, to a lesser extent, from the fMRI) literature, no consistent and clinically relevant differences have emerged in threat conditioning paradigms[10], which is surprising given the proposed central role of threat conditioning in the pathophysiology of PTSD[4]. A possible explanation is that behavioral measures, for example, SCR, are noisy and can indeed be interpreted as noisy realizations of deterministic learning models[33]. This suggests that the differences that are relevant for the disease may in fact be reflected by the latent parameters of the generative model rather than the noisy behavioral data.

Although all veterans were combat-exposed, only some of them developed symptoms strong enough to warrant a classical (DSM based) PTSD diagnosis. While our results do not allow us to draw causal inferences, our data do support the notion that veterans may develop more severe PTSD symptoms in response to altered neural computation of value and associability in several brain regions. Interestingly, the enhanced sensitivity to prediction errors was partially mediated by the striatal associability computations, suggesting that both increased prediction-error weight and decreased striatal tracking of associability independently contribute to PTSD symptoms. It is possible that the enhanced sensitivity to prediction errors might be the by product of the decreased neural associability tracking.

All in all, these results suggest that exploiting the combined power of computational, morphological, and functional analyses enable us to relate latent markers of learning and morphological indices to overt symptoms, as specific targets for investigating trauma-related psychopathology and its potential treatment.

Methods

Sample characteristics.

General description and excluded participants.—A total of 77 participants took part in the experiment. Due to problems with the SCR equipment or measurement problems during the functional scan, we did not obtain complete skin conductance and/or functional imaging data of 23 participants. These participants were similar compared to the included participants (see Supplementary Table 1). This was confirmed by comparing age, CAPS, BDI, STAIS, ASI, CES and education between excluded and included participants in a linear mixed model with the within subject factor metric (with the aforementioned variables as levels) and the between subject factor sample (levels: included, excluded) as well as a

random intercept. Importantly, the effects of sample ($F(1, 75.05) = 0.05$, $P = 0.832$) and the metric by sample interaction ($F(6, 434.7) = 0.13$, $P = 0.993$) were both not significant.

Thus, the full analysis was conducted on 54 combat veterans (see Table 1 for complete demographic and psychopathology details). The sample partially overlaps ($N = 30$) with the sample in a previous report[13]. The main reason to consider the partially overlapping structural data in the current study is that it increased the predictive validity of the right amygdala neural computations effect. In addition, since volume had already been shown to be predictive of CAPS symptoms[13], the current study aimed to explicitly test whether the effect of neural computation goes beyond this effect of volume.

The study was approved by the Yale University Human Investigating Committee and the Human Subjects Subcommittee of the VA Connecticut Healthcare System and compliance with all relevant ethical regulations was ensured throughout the study. All participants gave informed consent and were paid for their participation. Sample size was determined based on the assumption of a medium to large ($r = 0.4$) brain-behavior relationship between PTSD symptoms and blood-oxygenation-level-dependent (BOLD) activation. The necessary sample size was thus calculated as $N = 46$ with 80% power and $N = 61$ with 90% power.

Addressing the gender imbalance in the study sample.—Since there was a considerable gender imbalance in our study sample (49 of the 54 participants were male), we verified that all of the main results of the current study hold up when restricting the study sample to only male participants. Specifically, the effect of higher prediction-error weight predicting more CAPS symptoms remained significant ($\beta = 0.51$, $t(46) = 4.03$, two-tailed, $P < 0.001$). In addition, the effect of value computation for the right amygdala changed only minimally ($\beta = -0.27$, $t(43) = -1.78$, two-tailed, $P = 0.083$) and remained significant for the left amygdala ($\beta = -0.36$, $t(43) = -2.38$, two-tailed, $P = 0.022$); and the volume effect for the right amygdala remained significant ($\beta = -0.5$, $t(43) = -2.54$, two-tailed, $P = 0.015$).

In addition, the interaction of region and CAPS remained significant for value computation ($F(2, 94) = 3.12$, $P = 0.049$), and the interaction of learning component by CAPS remained significant for associability and prediction-error computation ($F(l, 188) = 21.94$, $P < 0.001$).

Thus, the heterogeneity introduced by gender seems to be negligible in this study, which is why we decided to keep the female participants in the sample to maximize statistical power of the otherwise relatively small study and precision of the estimated effects.

Addressing the clinical heterogeneity of the study sample.—We recruited veterans with a wide range of psychopathology, from completely healthy to pronounced PTSD, which can be seen from the distributions of PTSD (CAPS), depression (BDI), and anxiety symptoms (STAIS, ASI; Supplementary Fig. 1). Nevertheless, the reversal learning index did not differ significantly between combat veterans with and without PTSD ($t(48.46) = 0.17$, two-tailed, $P = 0.868$), between combats with high versus low PTSD (median split; $t(36.13) = -0.51$, two-tailed, $P = 0.611$), or between combat veterans with CAPS values on the extreme lower (CAPS ≤ 20; $N = 19$) or extreme higher end (CAPS ≥ 65; $N = 10$; $t(10.73) = 0.48$, two-tailed, $P = 0.638$).

To confirm that our main results were not significantly influenced by the clinical heterogeneity in our sample, we adjusted our models testing for neural computations of value, associability and prediction error for comorbidities and medication status. For value computation, the region by CAPS interaction remained significant when the model was adjusted for the number of comorbidities and medication status ($F(3,153) = 3.51$, $P = 0.017$). In addition, the effect was also robust to the adjustment for depression (BDI; $F(3, 153) = 3.73$, $P = 0.013$, state anxiety (STAIS; $F(3, 141) = 3.57$, $P = 0.016$) and anxiety sensitivity (ASI; $F(3, 150) = 3$, $P = 0.032$).

Similarly, for associability and prediction-error computation, the learning component by CAPS interaction remained significant when the model was adjusted for the number of comorbidities and medication status ($F(1, 306) = 10.41$, $P = 0.001$). In addition, the effect was also robust to the adjustment for depression (BDI; $F(1, 306) = 10.96$, $P = 0.001$), state anxiety (STAIS; $F(l, 282) = 10.9$, $P = 0.001$), and anxiety sensitivity (ASI; $F(l, 300) = 11.06$, $P < 0.001$).

Finally, we also verified that the correlations we report in the main manuscript held up when using non-parametric rank correlation tests. Specifically, we confirmed this for the correlation between prediction-error weight and CAPS ($\rho = 0.52$, two-tailed, $P < 0.001$), for the correlation between left amygdala neural value tracking and CAPS ($\rho = -0.3$, two-tailed, $P = 0.03$), the correlation between right amygdala volume and CAPS ($\rho = -0.37$, two-tailed, $P = 0.005$), the correlation between right amygdala neural value tracking and CAPS ($\rho = -0.27$, two-tailed, $P = 0.046$), the correlation between striatum neural value tracking and CAPS ($\rho = -0.35$, two-tailed, $P = 0.009$), the correlation between striatum neural associability tracking and CAPS ($\rho = -0.37$, two-tailed, $P = 0.006$), the correlation between dACC neural associability tracking and CAPS ($\rho = -0.29$, two-tailed, $P = 0.031$), and the correlation between hippocampus neural associability tracking and CAPS ($\rho = -0.29$, two-tailed, $P = 0.033$).

Together, these results suggest that the findings of this study were robust to the clinical heterogeneity of the study sample, and that the correlations we report were robust to outliers.

Study design.

The study consisted of a threat reversal learning experiment during fMRI on a single day. Threat learning was measured with SCR; structural magnetic resonance images were acquired in the same MRI session, immediately before the task. Participants were randomly assigned to one of two trial orders (see below). Due to the study design, data collection and analysis were not performed blind to the conditions of the experiments.

Screening procedures.

Psychopathology was assessed using the Structural Clinical Interview for DSM-IV, the gold-standard CAPS for PTSD diagnosis. Exclusion criteria were mental retardation, psychosis, bipolar disorder, substance dependency (life time), drug abuse in the past year, alcohol abuse in the past 60 days, neurological disorders, learning disabilities, attention deficit hyperactivity disorder, use of antipsychotic, hypnotic or sedative medications and less than 30 days' stable dose of antidepressants. Participants currently below PTSD clinical cutoff

(that is, presence of at least one criterion B symptom, at least three criterion C symptoms, at least two criterion D symptoms, as well as criteria A, E and F met) with a history of PTSD diagnosis were also excluded (remitted PTSD). We additionally measured the combat exposure score (CES), depression with the BDI, anxiety sensitivity with the ASI and state anxiety with the STAIS. Participants underwent breathalyzer and urine tests before the experiment to further validate substance use beyond the Structural Clinical Interview for DSM-IV.

Experimental task.

We used the same task as in a previous study on threat reversal in healthy participants[7]; that is, a threat discrimination and reversal task, with delay conditioning and partial reinforcement of about 33% (Fig. la). Participants were told that they would see visual images on a screen while receiving shocks. The level of the shocks was determined by participants before the experiment. Participants inside the MRI were instructed to pay attention to the screen and try to figure out the relationship between the stimuli and the shocks. Importantly, we did not mention the two stages or the reversal of contingencies. The conditioned stimuli were two mildly angry male faces from the Ekman series.

Stimuli and apparatus.

The unconditioned stimulus was a mild electric shock to the foot (200 ms duration, 50 pulses s^{-1}). The stimuli were presented for 4 s, with a 12 s intertrial interval in which a fixation point was presented (Fig. la). During acquisition, one face (face A) was paired with the unconditioned stimulus on one-third of the trials, while the other (face B) was never paired with the unconditioned stimulus. During reversal, these contingencies switched, and face B was now paired with the unconditioned stimulus on approximately one-third of the trials and face A was not paired with the unconditioned stimulus. The order of the different trial types was pseudorandomized (no consecutive reinforced trials and no more than two consecutive trials of each kind), and the designation of faces into ‘face A’ and ‘face B’ was counterbalanced across participants. During acquisition, there were 12 presentations of each of the faces, intermixed with an additional six presentations of face A that co-terminated with the unconditioned stimulus. Reversal immediately followed acquisition, and the transition between the stages was unsignaled. This stage consisted of 16 presentations of each of the faces, intermixed with seven additional presentations of face B that co-terminated with the unconditioned stimulus. We considered the first trial in which face B co-terminated with the unconditioned stimulus as the beginning of the reversal stage (Fig. 1a).

Physiological data acquisition and analysis.

Mild shocks were delivered through a stimulating bar electrode attached to the participant’s right ankle. A BIOPAC stimulator charged by a stabilized current was used, with cable leads that were magnetically shielded and grounded through a radio frequency filter. The participants were asked to set the level of the shock themselves using a work-up procedure before scanning. In this procedure, a participant was first given a very mild shock (20 V, 200 ms, 50 pulses/s), which was gradually increased to a level the participant indicated as uncomfortable, but not painful (with a maximum level of 70 V). Skin conductance was assessed with shielded Ag-AgCl electrodes, filled with standard NaCl electrolyte gel and

attached to the middle phalanges of the second and third fingers of the left hand. The electrode cables were grounded through an radio frequency filter panel. The skin conductance signal was amplified and recorded with a BIOPAC Systems skin conductance module connected to a 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 level of SCR was assessed for each trial as the base-to-peak amplitude difference in skin conductance of the largest deflection (in microsiemens; μS) in the 0.5–4.5-s latency window after stimulus onset. The minimal response criterion was 0.02 μS. Responses below this criterion were encoded as zero. The raw skin conductance scores were square-root transformed to normalize the distributions, and scaled according to each participant's average response to the unconditioned stimulus.

Statistical analysis.

We averaged the learning effects (face A minus face B) across trials by stage (acquisition, reversal) for each participant and calculated a threat reversal index by subtracting the learning effect of reversal from the learning effect of acquisition. To assess whether participants showed successful threat reversal, we tested whether the reversal index was significantly different from zero with a one-sample *t*-test. The threshold for this analysis was set at $P < 0.05$, two-tailed. The relationship between latent learning parameters (see below) and PTSD symptomatology was estimated with a linear regression model. Data distribution was assumed to be normal but this was not formally tested. However, individual data points are shown in scatter plots throughout the manuscript.

Computational modeling.

Following classic computational learning theory[31], we assumed a deterministic learning model and a probabilistic observation model to describe the generation of our data. The deterministic learning model describes the dynamics of how internal variables gate learning, while the observation model describes how the internal variables are realized in observed data.

Pearce-Hall learning model.

Unlike the Rescorla-Wagner model (see below) that treats the learning rate as constant, the Pearce-Hall model for associability-gated learning substitutes associability for the constant learning rate. Thus, such a model incorporates prediction-error-driven value updating into an associability model, resulting in the hybrid model:

$$\begin{aligned} \delta_n &= r_n - V_n(x_n) \\ V_{n+1}(x_n) &= V_n(x_n) + \kappa\alpha_n(x_n)\delta_n \\ \alpha_{n+1}(x_n) &= \eta|\delta_n| + (1-\eta)\alpha_n(x_n). \end{aligned} \tag{1}$$

Here, x_n is the conditioned stimulus on trial n (conditioned stimulus: CS+ or CS−) and r_n as the unconditioned stimulus delivered (1 for unconditioned stimulus, 0 for no unconditioned stimulus). The punishment prediction error δ_n measures the difference between the expected

and predicted shock on trial n. The associability α for the value update is a variable. The value for the conditioned stimulus not observed on trial n remains unchanged. Since associability of trial n depends on absolute prediction errors from past but not current trials, associability $\alpha_n(x_n)$ and prediction error δ_n are relatively uncorrelated.

To derive the best fits for this model, we assumed that $V_0 = 0.5$, reflecting the assumption that getting a shock or not was equally likely for the first trial. We compared the fit of different versions of the hybrid model to the SCR data by optimizing the free parameters of each model. We assumed the likelihood of each trial's SCR S_n to be an independent and identically distributed Gaussian distribution around a mean determined by value, associability or the combination of both value and associability as predicted by the model on that trial (plus a constant term):

$$\begin{aligned} S_n &\sim \text{Normal}\left(\beta_0 + \beta_1 V_n(x_n), \sigma\right) \\ S_n &\sim \text{Normal}\left(\beta_0 + \beta_1 \alpha_n(x_n), \sigma\right) \\ S_n &\sim \text{Normal}\left(\beta_0 + \beta_1 V_n(x_n) + \beta_2 \alpha_n(x_n), \sigma\right). \end{aligned} \tag{2}$$

As can be seen, these correspond to linear regressions of value or associability, or the combination of both, to the SCR. We tested all three possible combinations (equations 1–3; Hybrid (V); Hybrid (α); Hybrid ($\alpha + V$)), all in separate fits of all free parameters.

Using Hierarchical Bayesian modeling, we first verified that we could recover simulated parameters of initial associability (α_0), κ and the associability learning rate, η (Supplementary Fig. 2). We also ruled out that an extended model with an additional scaling parameter that captured a change of the prediction-error weight for the reversal stage would fit the data better (Fig. 2a)

Rescorla-Wagner learning model.

Although we found that a hybrid model of associability and value computation outperformed a simpler Rescorla-Wagner model, we were also interested in how a basic Rescorla-Wagner model could explain value computation in the amygdala. The Rescorla-Wagner model is the standard model of error-driven predictive learning. It assumes that the expected value (V) for each trial is updated according to the learning rate and the prediction error:

$$\begin{aligned} V_{n+1}(x_n) &= V_n(x_n) + \alpha\delta_n \\ \delta_n &= r_n - V_n(x_n), \end{aligned} \tag{3}$$

Here, x_n is the conditioned stimulus on trial n (face A or face B) and r_n as the unconditioned stimulus delivered (1 for unconditioned stimulus, 0 for no unconditioned stimulus). The punishment prediction error δ_n measures the difference between the expected and predicted shock on trial n. The learning rate α for the value update is a constant free parameter. The expected value for the conditioned stimulus absent on trial n remains unchanged. To derive the best fits from the Rescorla-Wagner model, we assumed that $V_0 = 0.5$, reflecting the assumption that getting a shock or not was equally likely for the first trial. We considered that model-based value computation would be correlated with amygdala activity, and that this correlation would be more negative for individuals with higher levels of PTSD

symptoms. After verifying that simulated parameters could be recovered with the hierarchical Bayesian approach used in this study (Supplementary Fig. 5) and that the model indeed fitted the recorded SCRs (Supplementary Fig. 6), we calculated linear regression models including functional (value encoding) and structural indices for amygdala as predictors of the PTSD symptoms. To account for unspecific intersubject variability, these models were adjusted for learning rate, age, gender, head movement and total intracranial volume. For the right amygdala (Supplementary Fig. 7a), we found that volume predicted CAPS symptoms ($\beta = -0.49$, $t(46) = -2.59$, two-tailed, $P = 0.013$ and Supplementary Fig. 7b) while neural activity as a predictor did not reach statistical significance ($\beta = -0.28$, $t(46) = -2$, two-tailed, $P = 0.052$ and Supplementary Fig. 7c). In the left amygdala, the effect of value-dependent activity remained significant when including amygdala volume in the same model ($\beta = -0.35$, $t(46) = -2.42$, two-tailed, $P = 0.02$ and Supplementary Fig. 7d), but no independent effect for volume emerged ($\beta = -0.2$, $t(46) = -1.04$, two-tailed, $P = 0.303$).

We also confirmed that the results were robust to the specific ROI definition of the amygdala and that individual differences in right amygdala volumes did not affect the effect of neural activity on CAPS symptoms: We repeated our analysis of right amygdala value computation using the winning hybrid ($\alpha + V$) model. We used the individual amygdala segmentations as computed by Freesurfer as masks for the ROI analysis of the right amygdala in SPM. After running the recon-all pipeline in Freesurfer, we converted and binarized the subcortical segmentation of each individual to NIfTI format. We then applied the individual normalization parameters calculated by SPM during the SPM preprocessing pipeline to warp the Freesurfer segmentation to the Montreal Neurological Institute space. A figure (Supplementary Fig. 8) shows two illustrative participants with individual amygdala masks (estimated in Freesurfer and indicated in red) projected on to their T1-weighted brain anatomy in SPM.

Given that we found an effect of volume for the right amygdala, we thus extracted the mean beta estimates of these individual right amygdala masks and entered the estimates in a multivariable regression model, using CAPS as the dependent measure and the beta estimates together with amygdala volume as predictors, adjusting for age, gender and intracranial volume. We found a similar effect for the neural activity compared to the original findings in the right amygdala ($\beta = -0.3$, $t(48) = -2.07$, two-tailed, $P = 0.044$), suggesting that the BOLD effects were correctly estimated.

To further characterize the relationship between structure and function we added the interaction term to the model and found that there was no evidence for a synergistic effect between these independent variables (right amygdala: $\beta = -0.7$, $t(45) = -0.71$, two-tailed, $P = 0.482$; left amygdala: $\beta = -0.02$, $t(45) = -0.02$, two-tailed, $P = 0.988$). Moreover, the correlation between structure and function was not significant (right amygdala: $r(52) = 0.05$, two-tailed, $P = 0.742$; left amygdala: $r(52) = 0.04$, two-tailed, $P = 0.797$). Further, a mediation analysis with amygdala volume as a mediator of the association between value activity and CAPS symptoms did not show evidence for full or partial mediation; instead, inclusion of amygdala volume did in fact improve the predictive validity of neural activity. A possible explanation is a compensatory recruitment of amygdala neurons in veterans with smaller amygdala volumes, probably due to a stress-related gray matter reduction[23].

We also tested a potential difference in learning rates between acquisition and reversal and additionally tested an extended version of the Rescorla-Wagner model. We added an additional scaling parameter ρ, which captures the change in the learning rate during the reversal stage. For acquisition, we thus used the classical Rescorla-Wagner model:

$$V_{n+1}(x_n) = V_n(x_n) + \alpha\delta_n \\ \delta_n = r_n - V_n(x_n), \tag{4}$$

and the extended model for reversal:

$$V_{n+1}(x_n) = V_n(x_n) + \rho\alpha\delta_n \\ \delta_n = r_n - V_n(x_n), \tag{5}$$

where ρ is the scaling parameter. We performed a model comparison between the two models, computing the DIC that captures the goodness of fit of a Bayesian hierarchical model with lower values meaning better fits[34]. Notably, we found that the simpler model provides a better fit to the data (extended model: DIC = 6223.11; simpler model: DIC = 6003.91).

Finally, we investigated whether fitting V_0, the initial value, as an additional free parameter would improve the model fit and found that the resulting DIC was higher than the one from the original simpler model; we thus kept the simpler model with V_0 fixed at 0.5.

Hierarchical Bayesian model fitting.

Bayesian modeling versus MLE.—We used Hierarchical Bayesian analysis (HBA) to obtain estimates of the free parameters in our computational models. The advantage of HBA compared to MLE is that individual differences are accounted for but information across individuals is pooled so that individual estimates are pulled toward the population mean (an effect sometimes referred to as shrinkage[35]).

In MLE, on the other hand, point estimates are obtained that maximize the likelihood of the data for each individual separately[36]. Individual ML estimates, however, may suffer from noise and may be unreliable when faced with an insufficient amount of data[37]. In addition, there is no guarantee that non-linear optimization algorithms commonly used (such as fmincon in MATLAB) will provide a set of parameter values that uniquely maximize the log-likelihood; indeed, premature stops of the algorithm are common, which provide a local (and thus sub-optimal) instead of a global maximum of the likelihood function[36]. Group-level analysis of MLE, which treats a group as a single subject and estimates a single set of parameters for a whole group of individuals, is commonly used to generate fMRI regressors for model-based analyses as it is supposed to generate more reliable estimates needed for fMRI[33], but it inevitably ignores individual differences and does not allow for group comparisons.

Hybrid model comparison using MLE.—Nevertheless, to replicate the model comparison reported in Li and colleagues more directly[8], we also performed a model comparison of the hybrid models using MLE as in the aforementioned study, and ranked all

three hybrid models as well as the Rescorla-Wagner model according to their Bayesian Information Criterion (BIC). Consistent with the results in Li and colleagues[8], the model with the lowest BIC was the hybrid ($\alpha + V$) model, outperforming the other hybrid models as well as the Rescorla-Wagner model (Supplementary Fig. 4a). More specifically, direct comparisons using likelihood ratio tests revealed that the Hybrid (V) model outperformed the Rescorla-Wagner analysis model ($\chi^2 = 408.11$, d.f. = 108, $P < 0.001$), and the Hybrid (α + V) outperformed the Rescorla-Wagner model ($\chi^2 = 877.61$, d.f. = 162, $P < 0.001$), the Hybrid (V) model ($\chi^2 = 469.5$, d.f. = 54, $P < 0.001$) and the Hybrid (α) model ($\chi^2 = 348.54$, d.f. = 54, $P < 0.001$).

We did not find evidence that the individual MLE model fits interacted with the PTSD symptomatology; the correlation between model parameters and symptoms was essentially flat for each of the four models (Supplementary Fig. 4b).

Details on the Bayesian modeling procedure.—To perform HBA, we used the probabilistic programming language Stan v.2.15.1 (Stan Development Team, 2014), which makes use of Markov chain Monte Carlo (MCMC) sampling algorithms termed Hamiltonian Monte Carlo. Hamiltonian Monte Carlo provides an efficient sampling algorithm even for multilevel models and highly correlated parameters[38].

For the Rescorla-Wagner model, individual parameters were assumed to be drawn from group-level normal distributions. Normal and half-Cauchy distributions were used for the priors of the group-level means and standard deviations, respectively[39,40]. We used weakly informative priors[35] to minimize the influence of those priors on the posterior distributions with our relatively small sample size. As the learning rate α is bounded between 0 and 1, we used the inverse probit transformation (the cumulative distribution function of a unit normal distribution) to convert unconstrained values into this range. The mathematical relationship between the probability density function and the cumulative density function of the unit normal distribution guarantees for this transformation that the converted prior will be uniformly distributed between 0 and 1. Stan provides a fast approximation of the inverse probit transformation (the Phi_approx function) to achieve this. The learning rate was thus declared as follows:

$$\mu_{\alpha'} \sim \text{Normal}\ (0, 1)$$

$$\sigma_{\alpha'} \sim \text{half-Cauchy}\ (0, 5)$$

$$\alpha' \sim \text{Normal}\left(\mu_{\alpha'}, \sigma_{\alpha'}\right)$$

$$\alpha = \text{Probit}^{-1}(\alpha')$$

where μ'_α and σ'_α are hyper-parameters that dictate the distribution of α' and sequentially α.

A total of 2,000 samples were drawn after 1,000 burn-in samples for each of four chains (resulting in a total of 8,000 samples). To assess the convergence of the chains for each parameter, we used the Gelman-Rubin test[41] that calculates an $\hat{R}$ statistic, with $\hat{R}$ values close to 1.00 indicating that the MCMC chains have converged to the target distributions.

Notably, the $\hat{R}$ values obtained for all model parameters were 1.00, and visual inspection of MCMC chains confirmed the mixing of MCMC samples. In addition, effective sample sizes (ESS) of model parameters, which are associated with autocorrelation and mixing of MCMC chains (with a smaller ESS indicating higher autocorrelation), were typically greater than 1,000 (out of 8,000 total samples). The minimum ESS of hyper-parameters was 592. Visual inspection of the parameters with smaller ESS confirmed their convergence to the target distributions.

For the hybrid models, priors of individual parameters were again assumed to be drawn from group-level normal distributions, but normal and half-normal distributions were used for the priors of the group-level means and standard deviations, respectively.

Parameter recovery tests.

To further verify the plausibility of our model, we used simulated data to test whether simulated parameters could be recovered (recovery tests). We generated true parameter values, simulated synthetic behavioral data based on the parameters and recovered their parameter values using the HBA described in the previous section. Results confirmed that the model was successful at recovering the simulated parameters (Supplementary Figs. 2 and 5).

Statistical analysis.

For the Rescorla-Wagner model, we used the individual posterior means of the learning rate to calculate the trial-wise expected value for each participant. These values were used as parametric modulators in the model-based fMRI analysis. For the hybrid model, trial-wise expected value, associability and prediction error were used as parametric modulators in the model-based fMRI analysis (see below).

Structural magnetic resonance imaging and analysis.

A Siemens Trio TIM 3T and 12-channel receiver array head coil were used for data acquisition. High-resolution T1-weighted anatomical images ($1 \times 1 \times 1$ mm^3) were acquired with an MPRAGE pulse sequence (voxel size $1 \times 1 \times 1$ mm^3; repetition time = 2.5 s; echo time = 2.77ms; flip angle = 7°; 256 × 256 matrix, 176 sagittal slices of 1 mm).

Blinded to the clinical status, image processing and segmentation were conducted using the automated Freesurfer recon-all pipeline (http://surfer.nmr.mgh.harvard.edu). Freesurfer transforms brains from native space to standard space to perform subcortical segmentation, and then transforms them back to native space to extract individual amygdala volumes. We thus used those extracted measures of amygdala volume for each participant and restricted

the analysis to this a priori defined ROI. Amygdala volume measures were then used as predictors in multivariable linear regressions (see below).

Notably, excessive head motion was found to be associated with reduced estimates of gray matter thickness and volume compared to age- and gender-matched samples and consequently with inflated effect sizes[42]. Following a recent suggestion that participants' head movement during functional imaging sequence may provide a proxy for head movement during the structural sequence (where no head movement is recorded), we calculated the total head movement in mm during fMRI for each participant and included this scalar as a covariate in the statistical analysis[42]. Importantly, this covariate was used as a proxy measure for head movement during the anatomical session. The movement during the fMRI session was regressed out in the fMRI design matrix. Note that the exclusion of this additional regressor in the multivariable regression did not alter the main results of our analysis.

fMRI and analysis.

Functional images were acquired using a single-shot gradient echo planar imaging sequence (repetition time = 2,000 ms; echo time = 25 ms; field of view =192 cm; flip angle = 75°; bandwidth = 4,340 Hz per pixel; echo spacing = 0.29 ms). Forty contiguous oblique-axial slices ($3 \times 3 \times 3$ mm^3 voxels) parallel to the anterior-commissure-posterior-commisure line were obtained.

Analysis of the imaging data were conducted using SPM 12 (http://fil.ion.ac.ak/spm12). After discarding the first eight volumes, native-space images were realigned, slice-time corrected and co-registered to each subject's structural scan. Structural image preprocessing included segmentation, bias correction and spatial normalization; these normalization parameters were also used to normalize the functional images. Finally, functional images were smoothed with a Gaussian kernel (4mm full-width at half-maximum).

Model-based fMRI analysis.

We conducted a computational analysis using the hybrid Pearce-Hall learning model, with the fMRI regressors derived from the fits to the SCR data. Cue onset and offset were modeled as two discrete events, and each expected value (V) regressor was included as a parametric modulator of the stimulus onset event. In addition, the occurrence of a shock (0 for trials with no shock, 1 for trials with a shock) and prediction error were modeled as parametric modulators of cue offset.

Six regressors modeling affine head-motion parameters were also included in the GLM. All events were convolved with a canonical gamma-variate hemodynamic response function. The contrast of interest was the correlation of expected value V, corresponding to the expectation of a shock on each trial, with the BOLD response in the brain. We thus computed images of this contrast for each participant and used the contrast images as input for the ROI-based analyses. Our ROI-based analysis focused on the amygdala. We defined the amygdala ROIs functionally, using an independent contrast of conditioned stimuli (both face A and face B) versus baseline and a relatively loose contrast of $P < 0.001$. For each ROI, we extracted the mean beta estimates obtained from the GLM for the correlation of

expected value y with the BOLD response. The beta estimates were then entered as predictors in multivariable linear regressions.

To assess the independent contributions of structural and functional indices on PTSD symptoms, we then calculated multivariable linear regressions including both the structural and functional indices as predictors, and the symptoms as measured with the CAPS as outcome measures. For each ROI, the structural index was the volume, the functional indices were the extracted mean beta estimates obtained from the computational GLM for the correlation of expected value y with the BOLD response. These models were adjusted for learning rate, age, gender, head movement and total intracranial volume to adjust for unspecific inter subject variability.

In addition to the amygdala, we extended our analysis to the dACC, the striatum and hippocampus. The dACC was defined with an independent functional contrast (face A versus face B in acquisition, corresponding to conditioned stimulus: CS+ versus CS−); since Li and colleagues[8] found the strongest activation in the caudate nucleus, we used a contrast of shock occurrence versus baseline in a sample of healthy control participants who underwent the same task to ensure reliable activation[43,44]. Like the functional contrast for the dACC, this contrast for the striatum was also independent of the computational contrasts. The hippocampal ROI was defined anatomically using the WFU Pickatlas[45]. We then used two separate GLMs to examine neural activity related to value encoding (GLM 1) as well as associability and prediction error (GLM 2). Replicating the design of a previous study in reversal learning and value encoding for GLM l[9], we included value encoding as parametric modulator of stimulus onset and included shock occurrence and prediction error as additional regressors of stimulus offset. For GLM 2, we followed the design of the previous study by Li and colleagues[8] and included associability, shock occurrence and prediction error as parametric modulators of stimulus offset.

We then extracted the beta estimates of the ROIs and computed separate linear mixed models for value computation as well as associability and prediction error, respectively Neural activity was used as the dependent variable in these models, and predictors for region and CAPS (model 1, predicting value computation) as well as learning component (model 2; predicting associability and prediction error) were entered as predictors. In addition, we included a random intercept and a random slope for region to account for the within-subject correlations.

We also performed a computational analysis using the simpler Rescorla-Wagner learning model, with the fMRI regressors derived from the fits to the SCR data. Cue onset and offset were modeled as two discrete events, and each expected value (V) regressor was included as a parametric modulator of the stimulus onset event. In addition, the occurrence of a shock (0 for trials with no shock, 1 for trials with a shock) and prediction error were modeled as parametric modulators of cue offset, but are not considered in this study due to algebraic collinearity with the V regressor. This means that the parametric regressor of interest was expected value, which modulated cue onset, while the regressors for shock outcome and prediction error were included in the design matrix (modulating cue offset) but are not considered in this study.

While this setup is in line with previous studies using the same reversal paradigm together with computational modeling[8,9], we manually confirmed that that the amount of collinearity in the fMRI design matrix was acceptable. We calculated the variance inflation factor (VIF) for the value regressor. The VIF reflects how much the variance of the estimated regression coefficient is increased by the correlation among the model regressors. Its square root quantifies how larger the standard error is compared with what it would be if the regressor were uncorrelated with the model regressors. While it is common practice to consider a VIF > 10 as problematic, it is important to note that even in the presence of collinearity the regression coefficients will be unbiased. As the term VIF suggests, what is affected is the variance of the estimates, resulting in increased noise and reduced statistical power. However, we verified that the VIF of the value regressor was below five for each participant, with a mean VIF across participants of 1.5 (s.d. 0.13) suggesting that collinearity was not an issue for this design. Supporting this conclusion, the effects for amygdala value computation on CAPS symptoms remained significant when including the amygdala activation during shocks in the same model (left: $\beta = -0.34$, $t(46) = -2.32$, two-tailed, $P = 0.025$; right: $\beta = -0.31$, $t(46) = -2.14$, two-tailed, $P = 0.038$).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The main source of funding for this work was provided by NIMH 105535 R01 grant awarded to I. Harpaz-Rotem and D. Schiller (MPI) and funding provided by the Clinical Neurosciences Division of the National Center for PTSD. Additional support was provided by Klingenstein-Simons Fellowship Award in the Neurosciences to D. Schiller, The Brain and Behavior Research Foundation 23260 to I. Harpaz-Rotem, Chinese NSF grant 31421003 to J. Li, and the Swiss National Science Foundation grant SNF 161077 to P. Homan. The analytic work was supported in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

References

1. American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, 4th edn (American Psychiatric Publishing, Arlington, 2013).
2. Pietrzak RH, Tsai J, Harpaz-Rotem I, Whealin JM & Southwick SM Support for a novel five-factor model of posttraumatic stress symptoms in three independent samples of Iraq/Afghanistan veterans: a confirmatory factor analytic study. J. Psychiatr. Res 46, 317–322 (2012). [PubMed: 22154134]
3. Harpaz-Rotem I, Tsai J, Pietrzak RH & Hoff R The dimensional structure of posttraumatic stress symptomatology in 323,903 U.S. veterans. J. Psychiatr. Res 49, 31–36 (2014). [PubMed: 24275548]
4. Lissek S & van Meurs B Learning models of PTSD: theoretical accounts and psychobiological evidence. Int. J. Psychophysiol 98, 594–605 (2015). [PubMed: 25462219]
5. Pavlov I Conditioned Reflexes (Oxford Univ. Press, Oxford, 1927).
6. Pearce JM & Hall G A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol. Rev 87, 532–552 (1980). [PubMed: 7443916]
7. Schiller D, Levy I, Niv Y, LeDoux JE & Phelps EA From fear to safety and back: reversal of fear in the human brain. J. Neurosci 28, 11517–11525 (2008). [PubMed: 18987188]
8. Li J, Schiller D, Schoenbaum G, Phelps EA & Daw ND Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci 14,1250–1252 (2011). [PubMed: 21909088]

9. Atlas LY, Doll BB, Li J, Daw ND & Phelps EA Instructed knowledge shapes feedback-driven aversive learning in striatum and orbitofrontal cortex, but not the amygdala. eLife 5, e15192 (2016). [PubMed: 27171199]
10. Duits P et al. Updated meta-analysis of classical fear conditioning in the anxiety disorders. Depress. Anxiety 32, 239–253 (2015). [PubMed: 25703487]
11. Browning M, Behrens TE, Jocham G, O'Reilly JX & Bishop SJ Anxious individuals have difficulty learning the causal statistics of aversive environments. Nat. Neurosci 18, 590–596 (2015). [PubMed: 25730669]
12. LeDoux JE Emotion circuits in the brain. Annu. Rev. Neurosci 23, 155–184 (2000). [PubMed: 10845062]
13. Pietrzak RH et al. Amygdala-hippocampal volume and the phenotypic heterogeneity of posttraumatic stress disorder: a cross-sectional study. JAMA Psychiatry 72, 396–398 (2015). [PubMed: 25692480]
14. Admon R et al. Human vulnerability to stress depends on amygdala's predisposition and hippocampal plasticity. Proc. Natl Acad. Sci. USA 106, 14120–14125 (2009). [PubMed: 19666562]
15. Neumeister P et al. Specific amygdala response to masked fearful faces in post-traumatic stress relative to other anxiety disorders. Psychol. Med 48, 1209–1217 (2018). [PubMed: 28950918]
16. Rescorla R & Wagner A in Classical Conditioning II: Current Research and Theory (eds Black AH & Prokasy WF) (Appleton-Century-Crofts, New York, 1972).
17. Roesch MR, Esber GR, Li J, Daw ND & Schoenbaum G Surprise! Neural correlates of Pearce-Hall and Rescorla-Wagner coexist within the brain. Eur. J. Neurosci 35, 1190–1200 (2012). [PubMed: 22487047]
18. Jin J, Zelano C, Gottfried JA & Mohanty A Human amygdala represents the complete spectrum of subjective valence. j. Neurosci 35, 15145–15156 (2015). [PubMed: 26558785]
19. Belova MA, Paton JJ, Morrison SE & Salzman CD Expectation modulates neural responses to pleasant and aversive stimuli in primate amygdala. Neuron 55, 970–984 (2007). [PubMed: 17880899]
20. Klavir O, Genud-Gabai R & Paz R Functional connectivity between amygdala and cingulate cortex for adaptive aversive learning. Neuron 80, 1290–1300 (2013). [PubMed: 24314732]
21. Genud-Gabai R, Klavir O & Paz R Safety signals in the primate amygdala. J. Neurosci 33, 17986–17994 (2013). [PubMed: 24227710]
22. Morey RA et al. Mid-Atlantic MIRECC Workgroup. Amygdala volume changes in posttraumatic stress disorder in a large case-controlled veterans group. Arch. Gen. Psychiatry 69, 1169–1178 (2012). [PubMed: 23117638]
23. Wrocklage KM et al. Cortical thickness reduction in combat exposed U.S. veterans with and without PTSD. Eur. Neuropsychopharmacol 27, 515–525 (2017). [PubMed: 28279623]
24. Raio CM, Hartley CA, Orederu TA, Li J & Phelps EA Stress attenuates the flexible updating of aversive value. Proc. Natl Acad. Sci. USA 114,11241–11246 (2017). [PubMed: 28973957]
25. Roesch MR, Calu DJ, Esber GR & Schoenbaum G Neural correlates of variations in event processing during learning in basolateral amygdala. J. Neurosci 30, 2464–2471 (2010). [PubMed: 20164330]
26. Schultz W Dopamine neurons and their role in reward mechanisms. Curr. Opin. Neurobiol 7, 191–197 (1997). [PubMed: 9142754]
27. O'Doherty JP, Dayan P, Friston K, Critchley H & Dolan RJ Temporal difference models and reward-related learning in the human brain. Neuron 38, 329–337 (2003). [PubMed: 12718865]
28. Preuschoff K & Bossaerts P Adding prediction risk to the theory of reward learning. Ann. N.Y. Acad. Sci 1104, 135–146 (2007). [PubMed: 17344526]
29. Behrens TE, Woolrich MW, Walton ME & Rushworth MF Learning the value of information in an uncertain world. Nat. Neurosci 10, 1214–1221 (2007). [PubMed: 17676057]
30. Etkin A & Wager TD Functional neuroimaging of anxiety: a meta-analysis of emotional processing in PTSD, social anxiety disorder, and specific phobia. Am. J. Psychiatry 164, 1476–1488 (2007). [PubMed: 17898336]

31. Ruderman L et al. Posttraumatic symptoms and aversion to ambiguous losses in combat veterans. Depress. Anxiety 33, 606–613 (2016). [PubMed: 27000639]
32. Brown VM et al. Associability-modulated loss learning is increased in post-traumatic stress disorder. eLife 7, e30150 (2018). [PubMed: 29313489]
33. Daw ND in Decision Making, Affect, and Learning: Attention and Performance XXIII (eds Delgado MR, Phelps EA & Robbins TW) (Oxford Univ. Press, New York, 2011).
34. Spiegelhalter D, Best N, Carlin B & Van der Linde A Bayesian measures of model complexity and fit (with discussion). J. R. Stat. Soc. B 64, 583–639 (2002).
35. Gelman A et al. Bayesian Data Analysis 3rd edn (CRC Press, Boca Raton, 2013).
36. Myung IJ Tutorial on maximum likelihood estimation. J. Math. Psychol 47, 90–100 (2003).
37. Ahn W-Y, Haines N & Zhang L Revealing neurocomputational mechanisms of reinforcement learning and decision-making with the hbayesdm package. Comput. Psychiatr 1, 24–57 (2017). [PubMed: 29601060]
38. Gelman A, Lee D & Guo J Stan: A probabilistic programming language for Bayesian inference and optimization. J. Educ. Behav. Stat 40, 530–543 (2015).
39. Ahn W-Y et al. Decision-making in stimulant and opiate addicts in protracted abstinence: evidence from computational modeling with pure users. Front. Psychol 5, 849 (2014). [PubMed: 25161631]
40. Gelman A Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian. Anal 1, 515–534 (2006).
41. Gelman A & Rubin DB Inference from iterative simulation using multiple sequences. Stat. Sci 7, 457–472 (1992).
42. Savalia NK et al. Motion-related artifacts in structural brain images revealed with independent estimates of in-scanner head motion. Hum. Brain. Mapp 38, 472–492 (2017). [PubMed: 27634551]
43. Cools R, Clark L, Owen AM & Robbins TW Defining the neural mechanisms of probabilistic reversal learning using event-related functional magnetic resonance imaging. J. Neurosci 22, 4563–4567 (2002). [PubMed: 12040063]
44. Dodds CM et al. Methylphenidate has differential effects on blood oxygenation level-dependent signal related to cognitive subprocesses of reversal learning. J. Neurosci 28, 5976–5982 (2008). [PubMed: 18524902]
45. Maldjian JA, Laurienti PJ, Kraft RA & Burdette JH An automated method for neuroanatomic and cytoarchitectonie atlas-based interrogation of fMRI data sets. Neuroimage 19, 1233–1239 (2003). [PubMed: 12880848]

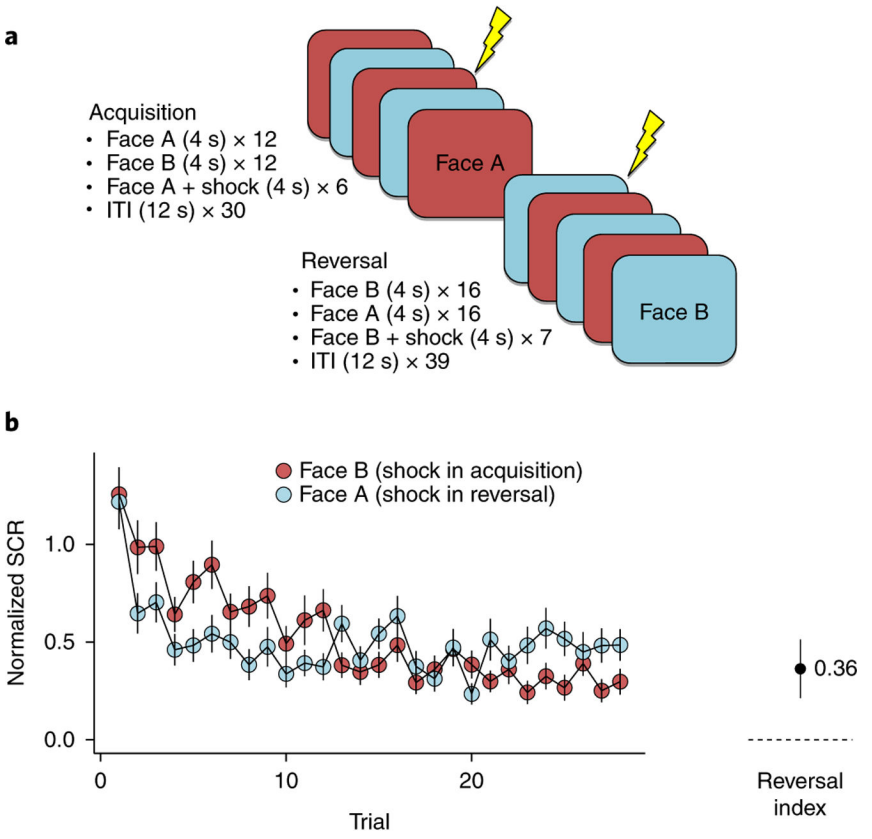

Fig. 1|. Experimental overview

a, Experimental design. The experiment consisted of 69 trials and a reinforcement rate of 33%. Stimuli were presented for 4 s in two pseudorandomized orders, followed by an intertrial interval of 12 s. During acquisition, face A was paired with a shock in about one-third of the trials and face B was not paired. Reversal started after 30 trials without previous instructions or warnings. During reversal, face B was now paired with a shock in about one-third of the trials, whereas face A was not paired anymore (ITI, intertrial interval), **b,** Time course of threat reversal learning. Mean normalized SCRs with standard errors ($N = 54$ participants). Participants showed successful threat reversal, indicated by a significant interaction of stage by stimulus; that is, a reversal index (subtracting stimulus discrimination (face A—face B) in reversal from stimulus discrimination in acquisition) with 95% confidence intervals that is significantly different from zero in a one-sample t test, two-sided ($t(53) = 4.75$, $P < 0.001$).

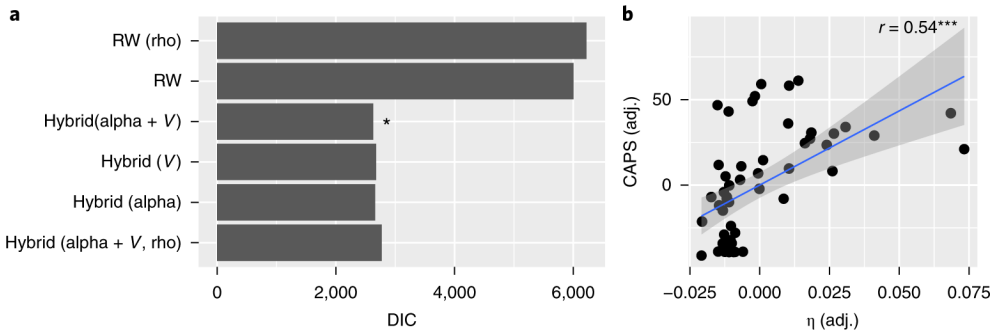

Fig. 2 |. Computational model comparison and relationship to PTSD symptoms.

a, All three versions of the hybrid model informed by the Pearce-Hall learning mechanism outperformed the simpler Rescorla-Wagner (RW) model. In addition, the hybrid model with associability and value outperformed the models with either value only or associability only and was thus the winning model (indicated with an asterisk). An extension of the Rescorla-Wagner or hybrid model with a scaling parameter ρ (rho) for the reversal stage, reflecting the potential change of learning during the reversal stage, did not perform better than the hybrid model with alpha, (associbaility), and V (value), which we thus kept as the winning model, **b**, Prediction-error weight η predicts symptoms as assessed with the CAPS. Using the best-fit model parameters, we found that a higher prediction-error weight η (which captures the learning rate for associability) predicted more CAPS symptoms. A partial correlation is shown after adjustments for age and gender and a Pearson correlation coefficient with a two-tailed significance test. Error shadings correspond to standard errors, adj., adjusted for all other parameters in the model; ***$P < 0.001$.

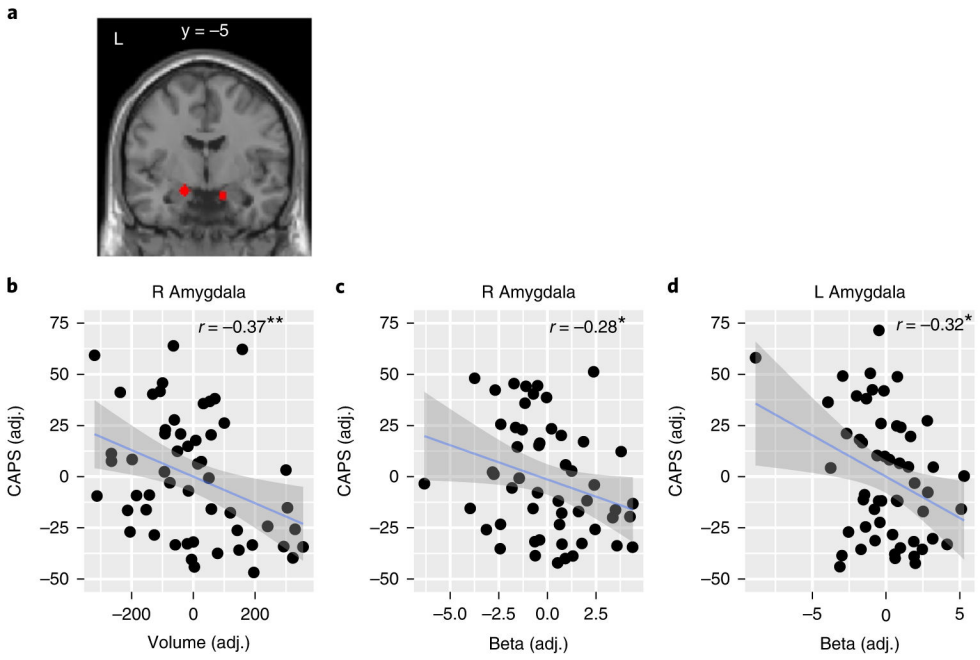

Fig. 3 |. Amygdala structure and value computation contribute to PTSD symptoms using a hybrid computational model of associability and value encoding.

a, Region of interest (ROI) used in the computational imaging analysis. The amygdala (red) was defined functionally, using the contrast of conditioned stimuli (both CS+ and CS−) versus baseline, **b**–**d**, Amygdala volume and value-dependent neural activity independently contribute to PTSD symptoms. Partial correlations are shown ($N = 54$ participants) with Pearson correlation coefficients and two-tailed significance tests. Right amygdala volume and activity as well as left amygdala activity correlated negatively with PTSD symptoms as measured with CAPS. Thus, lower value tracking in the amygdala and smaller amygdala volume correspond to higher symptom severity. Regressions were adjusted for age, gender, head movement, and total intracranial volume. Error shadings correspond to standard errors, adj., adjusted; $^{**}P < 0.01$; $^{*}P < 0.05$.

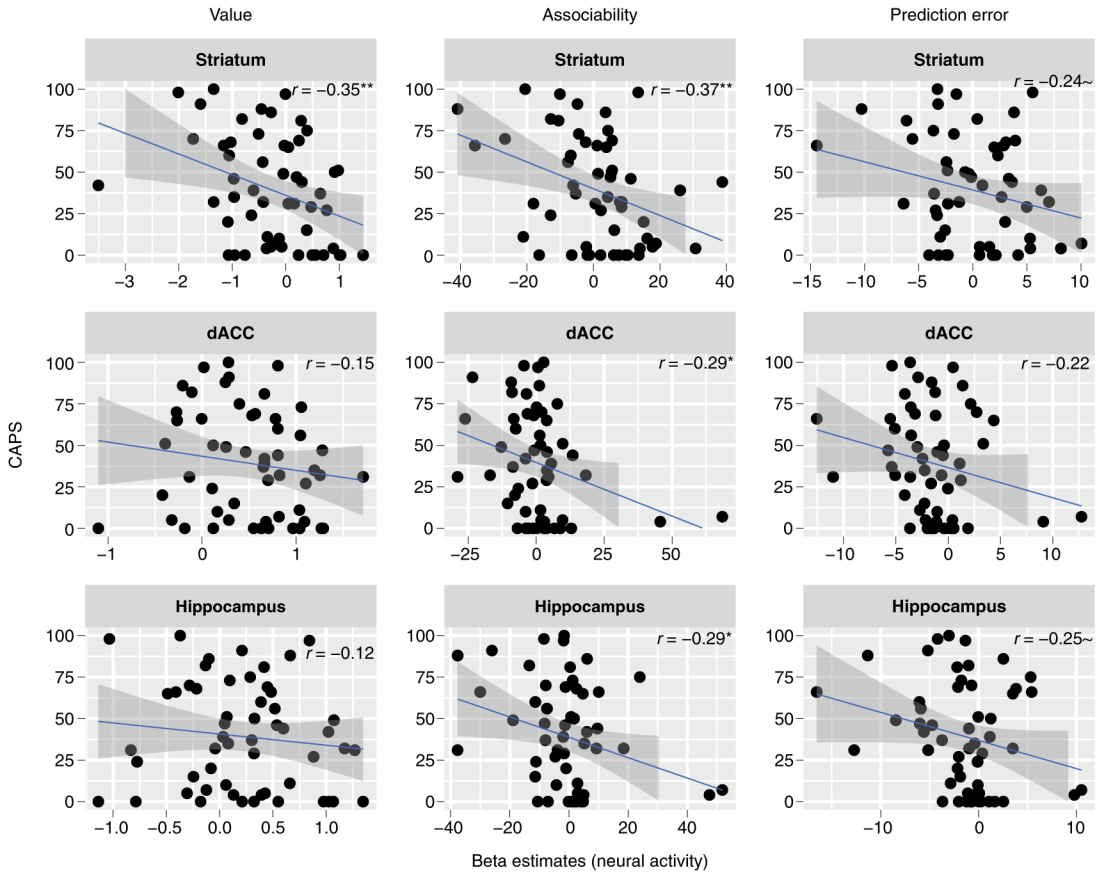

Fig. 4 |. Neural computations of value, associability and prediction error and their relationship to CAPS symptoms for different ROIs.

We found negative correlations for value encoding as well as associability that were attenuated for prediction error ($N = 54$ participants). Pearson correlation coefficients are shown with two-tailed significance tests. Error shadings correspond to standard errors. $^{**}P < 0.01$; $^{*}P < 0.05$; $\sim P < 0.1$.

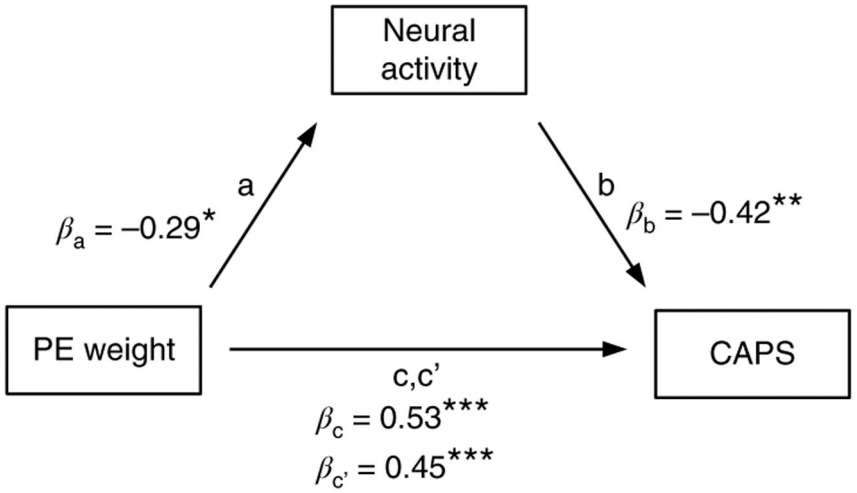

Fig. 5 |. Associability-related neural activity in the right striatum partially mediates the relationship between prediction-error weights and CAPS.

Standardized regression coefficients are shown (N = 654 participants) and their statistical significance tested with one-sample t tests, two-tailed. Both prediction-error weights and striatal neural tracking of associability independently predicted PTSD symptoms as measured with CAPS when included as predictors in the same model. PE, prediction error; *** $P < 0.001$; ** $P < 0.01$; * $P < 0.05$.

Table 1 |

Sample characteristics

Characteristic	*N*	Mean	s.d.
Males	49		
Females	5		
Age	54	32.8	8
Education	51	3.7	1.2
ASI	52	20.7	13.3
BDI	53	16	13.3
CAPS	54	39.2	32
CES	51	16.8	6.5
STAIS	49	40.8	13.7
Medicated	18		
Comorbidities:			
MDD	17		
Past alcohol abuse	7		
Panic disorder	5		
Past cannabis abuse	4		
Generalized anxiety	3		
Social phobia	3		
Anxiety disorder NOS	2		
Adjustment disorder	1		
Anxiety disorder GMC	1		
Dysthymic disorder	1		
Other DSM-IV Axis I disorder	1		
Past cocaine abuse	1		
Past opiates abuse	1		
Specific phobia	1		

Education was a categorical variable, defined as: 1, 8th grade or less; 2, some high school; 3, high school graduate or General Education Diploma; 4, some college; 5, college graduate; 6, advanced graduate degree. *Abbreviations:* ASI, Anxiety Sensitivity Index; BDI, Beck Depression Inventory; CES, combat exposure score; STAIS, State Anxiety subscale of the Spielberger State-Trait Anxiety Inventory; MDD, major depressive disorder; NOS, not otherwise specified; GMC, due to general medical condition; s.d., standard deviation; DSM, Diagnostic and Statistical Manual of Mental Disorders.