Cognitive, Affective, & Behavioral Neuroscience (2022) 22:199–213
https://doi.org/10.3758/s13415-021-00943-4

RESEARCH ARTICLE

Value estimation and latent-state update-related neural activity during fear conditioning predict posttraumatic stress disorder symptom severity

Allison M. Letkiewicz[1] · Amy L. Cochran[2] · Anthony A. Privratsky[3] · G. Andrew James[3] · Josh M. Cisler[4]

Accepted: 9 August 2021 / Published online: 26 August 2021

Abstract

Learning theories of posttraumatic stress disorder (PTSD) purport that fear-learning processes, such as those that support fear acquisition and extinction, are impaired. Computational models designed to capture specific processes involved in fear learning have primarily assessed model-free, or trial-and-error, reinforcement learning (RL). Although previous studies indicated that aspects of model-free RL are disrupted among individuals with PTSD, research has yet to identify whether model-based RL, which is inferential and contextually driven, is impaired. Given empirical evidence of aberrant contextual modulation of fear in PTSD, the present study sought to identify whether model-based RL processes are altered during fear conditioning among women with interpersonal violence (IPV)-related PTSD (n = 85) using computational modeling. Model-free, hybrid, and model-based RL models were applied to skin conductance responses (SCR) collected during fear acquisition and extinction, and the model-based RL model was found to provide the best fit to the SCR data. Parameters from the model-based RL model were carried forward to neuroimaging analyses (voxel-wise and independent component analysis). Results revealed that reduced activity within visual processing regions during model-based updating uniquely predicted higher PTSD symptoms. Additionally, after controlling for model-based updating, greater value estimation encoding within the left frontoparietal network during fear acquisition and reduced value estimation encoding within the dorsomedial prefrontal cortex during fear extinction predicted greater PTSD symptoms. Results provide evidence of disrupted RL processes in women with assault-related PTSD, which may contribute to impaired fear and safety learning, and, furthermore, may relate to treatment response (e.g., poorer response to exposure therapy).

Keywords Computational model · Reinforcement learning · PTSD · Fear conditioning · Fear extinction · Neuroimaging

Introduction

Posttraumatic stress disorder (PTSD) is an impairing anxiety-related disorder that is marked by deficits in several aspects of learning (Jovanovic et al., 2010; Jovanovic et al., 2012; Pacella et al., 2013). For example, individuals with PTSD exhibit overgeneralization of conditioned fear from threat-related cues to approximations of threat-related cues that are safe (Kaczkurkin et al., 2017; Lopresto et al., 2016), impaired inhibition of previously learned fear associations, and impaired recall of extinction learning (for a review, see Lissek & van Meurs, 2015). Given the ubiquity of learning deficits in PTSD, identifying specific learning-related processes that are disrupted is an important research endeavor.

Computational methods have enhanced researchers' ability to circumscribe cognitive processes with greater sensitivity and precision (Price et al., 2019; Stephan & Mathys, 2014). The progression of reinforcement learning (RL) models has been particularly successful, with computational models increasingly capturing behavior and learning phenomena not well explained by prior models (Cochran & Cisler, 2019; Le

✉ Allison M. Letkiewicz
allison.letkiewicz@northwestern.edu

[1] Department of Psychiatry and Behavioral Sciences, Northwestern University, 680 N. Lakeshore Drive, Suite 1520, Chicago, IL 60611, USA

[2] Departments of Math and Population Health Sciences, University of Wisconsin, Madison, WI, USA

[3] Department of Psychiatry, Brain Imaging Research Center, University of Arkansas for Medical Sciences, Little Rock, AR, USA

[4] Department of Psychiatry, University of Wisconsin, Madison, WI, USA

Pelley, 2004; Mihatsch & Neuneier, 2002; Redish et al., 2007). For example, whereas the standard Rescorla-Wagner (RW) model, which was developed to quantitatively formalize Pavlovian RL, models simple trial-and-error ("model-free") learning, it does not capture more dynamic processes that occur during conditioning and extinction. Indeed, hybrid Pearce-Hall/RW learning models, which track trial-by-trial associability (i.e., the salience of a cue), provide a better fit for probabilistic (Brown et al., 2018) and fear-related learning (Homan et al., 2019; Li et al., 2011) than the standard RW model among healthy controls and individuals with PTSD.

In addition to data-driven model development, alignment of computational models with psychological theory is paramount (Huys et al., 2016), as this will allow for the testing and refinement of current hypotheses regarding the role of learning in disorders, such as PTSD. Whereas prior research has primarily assessed relatively simple trial-and-error learning in PTSD (Brown et al., 2018; Cisler et al., 2015; Cisler et al., 2019; Ross et al., 2018), few have included models that capture higher-level cognitive processes. Importantly, theories of anxiety and PTSD implicate abstract, higher-order cognition in fear learning and extinction (Dunsmoor & Murphy, 2015). For example, the degree to which learned fear is generalized may depend on individuals' reasoning about internal conceptual representations (Dunsmoor & Murphy, 2015). In addition to the ability to reason about, integrate, and abstract categorical representations, which is facilitated by the anterior prefrontal cortex (Davis et al., 2017), the abstraction of rules and contextual information is mediated by prefrontal regions (Cools et al., 2004; Fogelson et al., 2009). Evidence of impaired contextual modulation of fear learning among individuals with PTSD further suggests a potentially important role of complex cognitive functions in the acquisition and revision of fear (Steiger et al., 2015).

Unlike model-free (e.g., RW) and hybrid RL, which involve trial-and-error revisions of cue-outcome associations, model-based RL captures structured and dynamically shifting conditions or "rules" of learning (Daw et al., 2005; Redish et al., 2007). Model-based RL is theorized to support the development of internal models (i.e., cognitive maps)[1] that contain hypotheses about different task conditions to allow a learner to make predictions about future actions within a changing environment (Daw et al., 2005; Gläscher et al., 2010). During a situation in which a stimulus is paired with an aversive outcome (e.g., a circle is paired with a shock), an individual utilizing model-free RL would develop a single outcome expectation for that stimulus (e.g., "circles are dangerous"). By contrast, during a situation in which the pairing of an aversive outcome and a stimulus can change depending on situational factors (i.e., the abstract 'state' of the environment), an individual utilizing model-based RL would develop separate outcome expectations for the various conditions and would differentially weight their expectations depending on which situation (i.e., state) was currently relevant (e.g., "circles are dangerous in situation X, but not Y"). In PTSD, difficulty with the latter may contribute to challenges with new safety learning, which is common in PTSD (Fani et al., 2012).

Whereas model-free and hybrid RL are primarily implemented within regions of the ventral striatum, amygdala, and the salience network (Beierholm et al., 2011; Brown et al., 2018; Cisler et al., 2019; Daw et al., 2011; Gläscher et al., 2010; Ross et al., 2018), model-based RL is primarily implemented within regions of the prefrontal cortex and frontoparietal network (FPN; Beierholm et al., 2011; Gläscher et al., 2010). Supporting the possibility that individuals with PTSD may have deficits in model-based RL, prior research has documented deficits in cognitive functions that are implemented within brain regions that overlap with those that support model-based RL, such as dorsolateral prefrontal cortex, inferior frontal gyrus, and anterior prefrontal cortex (Leskin & White, 2007; Polak et al., 2012; Stein et al., 2002; Woon et al., 2017; Alvarez & Emory, 2006; Doll et al., 2016). Thus, model-based RL may be disrupted in PTSD to a greater extent than model-free processes.

The primary goal of the present study was to build on previous work assessing model-based RL during acquisition and extinction of fear among women with assault-related PTSD. The study focused on assault-related PTSD because previous research has consistently shown that assault is a more potent risk factor for the development of PTSD than other forms of trauma (Breslau et al., 1998; Cisler et al., 2012; Frans et al., 2005; Kessler et al., 2017; Resnick et al., 1993). Additionally, because different forms of trauma predict different PTSD symptom profiles (Kelley et al., 2009), we selected participants with assault-related PTSD to increase the homogeneity of our participants, allowing us to avoid potential confounds of trauma type. Women were specifically selected for inclusion, because women are (1) twice as likely as men to develop PTSD (Kessler et al., 2005; Kilpatrick et al., 2013) and (2) at higher risk of exposure to many forms of interpersonal violence than men, including rape, sexual assault, and physical assault by an intimate partner (Iverson et al., 2013). It was hypothesized that the model-based RL model would provide a better fit for participants' behavior than the model-free and hybrid models, which do not allow a learner to develop sets of cue-outcome associations that are differentially applied and updated based on inferences about task rules (e.g., rules that differ for the acquisition and extinction context). It was further hypothesized that FPN encoding of trial-by-trial

[1] The terminology "model-based" and "model-free" refer to whether a learner develops an internal model of the environment, which occurs during model-based, but not model-free, RL. The distinction between model-free and model-based RL are explained in more detail elsewhere (e.g., Daw et al., 2005; Daw et al., 2011; Gläscher et al., 2010; Van Otterlo & Wiering, 2012; Wunderlich et al., 2012).

updates of current beliefs about task conditions, which are specific to the model-based model and contribute to differential weighting of cue value expectations based on a learner's hypotheses, would predict PTSD symptom severity. Specifically, it was anticipated that reduced FPN encoding would predict greater PTSD symptom severity, reflecting poorer contextually derived learning. Because PTSD is related to difficulties with fear extinction/safety learning (Jovanovic et al., 2009; Jovanovic et al., 2012), potential differences in the encoding of model-belief updates during acquisition versus extinction were explored. Due to high overlap between belief update and prediction error parameters during extinction, the acquisition versus extinction analyses focused on value expectations while controlling for belief updates.

Methods

Participants

A total of 103 women were enrolled as part of a larger randomized clinical trial across two study sites: (1) University of Arkansas Medical Sciences and (2) University of Wisconsin-Madison (note: n = 175 participants were assessed for eligibility, and n = 103 were enrolled; for full recruitment information, see Cisler et al., 2020). Primary inclusion criteria included female sex, aged between 21 and 50 years, and a current diagnosis of PTSD related to sexual or physical assault. Primary exclusion criteria included psychotic symptoms, pregnancy, learning disability, and medication, or magnetic resonance imaging (MRI) contraindications. Twelve women were excluded due to task visit no show, claustrophobia, or a positive drug screen, yielding 91 subjects. Of these 91 subjects, a total of 85 had either viable skin conductance responses or neuroimaging data (see Computational Modeling and Neuroimaging sections below).

Clinical Interview and Measures

The past month version of the Clinician Administered PTSD Scale for DSM-5 (CAPS-5) was used to assess for the presence of current PTSD related to interpersonal violence (i.e., assault-related PTSD; Weathers et al., 2018). The CAPS-5 is a 30-item structured clinical interview that assesses for PTSD symptoms across four clusters: reexperiencing, avoidance, negative cognitions and mood, and hyperarousal. Symptoms are rated on a scale from 0 (absent) to 4 (extreme/incapacitating). To meet criteria for current PTSD, individuals must endorse a score of 2 or above for at least one reexperiencing symptom, one avoidance symptom, two negative cognitions/mood symptoms, and two hyperarousal symptoms. In addition to a providing a categorical diagnosis, total symptom scores provided a dimensional measure of current PTSD symptoms. Although all participants had a diagnosis of PTSD, there was substantial variation in the CAPS-5 symptom severity scores (see distribution of scores in Supplemental Figure S1). The One-Word Receptive Picture Vocabulary Test, Fourth Edition (ROWPVT-4), was used as a proxy measure of intelligence quotient (IQ; Martin & Brownell, 2011). IQ was estimated to account for potential effects of individual differences in IQ on model-based RL, given that IQ deficits have previously been found to relate to poorer model-based RL (Culbreth et al., 2016). During the ROWPVT-4, participants match vocabulary words that are administered verbally to illustrations that are presented in a book (Brownell, 2000). Scores were normed according to chronological age. Additional clinical and trauma assessments were completed by participants but were not of primary interest. Follow-up tests were implemented to account for the potential the impact of these variables on results (for a description of the assessments and results of the follow-up tests, see the Supplement).

Fear Conditioning and Fear Extinction Task

Participants completed four task blocks that alternated between fear acquisition and fear extinction (Fig. 1a). The first acquisition block was preceded by a baseline (habituation) period of 12 trials (6 for presentations of each cue) without any administrations of the unconditioned stimulus (UCS).[2] The UCS was an electrotactile stimulation that was delivered to participants' lower leg. Stimulation level was set to a maximum of 50 mA, and participants' stimulation level was individually calibrated before the task at a level that was uncomfortable but not painful (approximately 7 of 10 on a Likert scale: 0 = not uncomfortable, 10 = extremely uncomfortable/painful). Triangles and circles served as the conditioned stimuli and different colored backgrounds identified the current context (i.e., the acquisition or extinction block), which were counterbalanced across participants (i.e., for half of participants, the CS+ was a triangle and for the other half the CS− was a triangle).

During each fear acquisition block, 18 conditioned safety cues (CS−) and 18 conditioned danger/threat cues (CS+) were presented for 3 seconds, with an intertrial interval of 2-6 seconds. During acquisition, the presentation of the CS+ was followed by an electrotactile stimulation (UCS) on 50% of trials, which occurred 2.5 seconds after the CS+ onset for duration of 500 msec. During each fear extinction block, there were 18 trials each of the CS− and CS+ cues, which were presented for 3 seconds, with an intertrial interval of 2-6 seconds. During extinction, no electrotactile stimulations occurred following the presentation of the CS+. The CS− and

[2] SCR responses did not differ between the CS+ and CS- cues during habituation, $t(146) = 1.28$, $p = 0.202$. Conditioned responses to the CS+ were significantly larger during acquisition than baseline, $t(146) = 7.93$, $p < 0.001$.

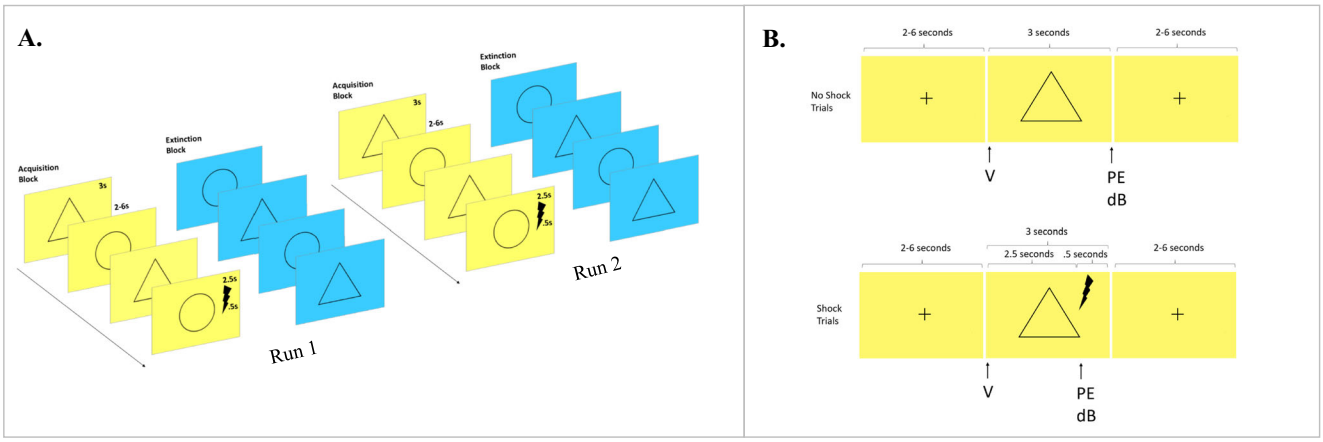


Fig. 1 **a** Schematic representation of the acquisition and extinction blocks of the Fear Conditioning and Fear Extinction Task. **b** Representation showing the temporal mapping of value estimations (V), prediction errors (PE), and latent state belief updates (dB) during the no shock and shock trials. For the neuroimaging analyses, the onset phase was parametrically modulated by trial-by-trial value expectations (*V*t,c) and the outcome phase was parametrically modulated by trial-by-trial PEs (positive and negative; PEt,c) and latent-state belief updates (dBt,c)

CS+ stimuli were pseudorandomly presented during each block, and participants completed a total of 156 task trials.

Skin Conductance Response Acquisition and Preprocessing

Following an approach used in previous studies (Homan et al., 2019; Li et al., 2011), participants' SCR data was used to test the fit of several RL models. SCR has previously been shown to map onto value expectations and associability during RL (Li et al., 2011). More specifically, anticipatory SCR scales with the degree to which individuals expect that an outcome will occur for a given cue (e.g., delivery of an electrotactile stimulation), with larger SCR responses reflecting greater expectation of an outcome. Model fit was tested by minimizing the error between model estimated trial-wise value expectations and participants' trial-wise SCR.

SCR data were acquired from participants' left hand with the BIOPAC MP150 Data Acquisition System using the EDA100C module with the MECMRI-TRANS (MRI compatible) cable system. BIOPAC AcqKnowledge 4.3 software recorded SCR data at a rate of 2,000 Hz at the Arkansas site and 1,000 Hz at the Wisconsin site. Data were preprocessed using an approach that is consistent with our prior studies (Cisler et al., 2020; Privratsky et al., 2020) and contemporary recommendations on modeling skin conductance data (Bach, 2014; Bach et al., 2010; Bach et al., 2013; Bach & Friston, 2013). This pipeline used a 10-ms median filter, unidirectional butterworth filter with 0.0159 hz and 5-hz low- and high-pass frequencies, and by downsampling to 10 hz. Next, trial-by-trial SCR responses were estimated using a forward convolution model of SCR and were normalized to individuals' maximum SCR response. Seventeen participants were excluded from computational modeling analyses due to flat responding, an excessive number of artifacts, or missing SCR data, yielding 74 participants whose data were included in the computational modeling. The amount of SCR data loss (19%) is comparable to prior fear extinction studies using SCR (Garfinkel et al., 2014; Haaker et al., 2013; Raij et al., 2018).

Computational Models

To identify whether a model-free, hybrid, or model-based RL model (from here on referred to simply as model-free, hybrid, and model-based models) provided a better fit to participants' SCR data, several models were tested. The primary set of models that were tested against the model-based model included a standard RW model and a hybrid model, both of which have previously been used to estimate learning parameters from SCR data during fear conditioning tasks and the latter of which has been found to provide a better fit than the standard RW model (Li et al., 2011). Additional versions of the standard RW model were tested for completeness (see the Supplement).

Model-Free Model

Model-free RL was assessed with several versions of the Rescorla-Wagner (RW) model. The standard RW assumes that a learning agent keeps track of associative strengths representing the learner's expectations for an outcome following the presentation of a cue. For a learning agent, we let a continuous variable $V_{t,c}$ denote the associative strength (i.e., value expectation) on trial t for observed cue c. We also let a binary variable $outcome_t$ denote the outcome on trial t, with $outcome_t$ equal to 1 if the participant received a shock and 0 otherwise. Associative strengths are updated via prediction errors (PE), given by the difference between what happened

(outcome) and what was expected, scaled by the learning rate. If cue c is presented on trial t, then the PE is denoted by $\delta_t = outcome_t - V_{t,c}$, and the associative strength of cue c is updated as follows: $V_{t+1,c} = V_{t,c} + \alpha_t \delta_t$, where α_t is a constant on trial t known as a learning rate. For the standard RW model, the same learning rate is used for all trials and associative strengths are only updated for presented cues, i.e., $\alpha_t = \alpha$ for all t and $V_{t+1,c} = V_{t,c}$ for cue c not presented on trial t.

Hybrid Model

The hybrid model builds on the RW model: $V_{t+1,c} = V_{t,c} + \kappa\alpha_t \times \delta_t$, where $V_{t,c}$ is the value expectation of cue (c) on the current trial (t), κ is a learning rate, α_t is the associability for the current trial, and δ_t is the PE for the current trial (Le Pelley, 2004; Li et al., 2011). As with the RW model, $\delta_t = outcome_t - V_{t,c}$. Unlike the RW model, a constant learning rate (κ) scales an associability parameter that changes from trial to trial and is defined as $\alpha_{t+1} = \eta|\delta_t| + (1\text{-}\eta)\alpha_t$. The free parameter, η, scales the prior magnitude of PE and its additive inverse scales the prior associability (i.e., weighted PE). A higher η reflects greater weighting of the prior PE relative to the prior associability.

Model-Based Model

The latent state (LS) model was used to capture model-based RL (Cochran & Cisler, 2019; Letkiewicz et al., 2020). This model was previously found to explain learning phenomena characterized by contextually based learning better than other widely used models (e.g., renewal, spontaneous recovery). Latent states are unobserved task rules/conditions that, in aggregate, define a learning environment. Each latent state contains sets of associations between cues and outcomes and a learner must infer which associations are currently most applicable. The LS model builds on the RW model by using integer l (i.e., the latent state) to index these sets of associations, where $V_{t,c,l}$ is the current value strength of option c for latent state l: $V_{t+1,c,l} = V_{t,c,l} + \alpha_{t,c,l} \times \delta_{t,l}$. For the LS model, the learning rate ($\alpha_{t,c,l}$) is specific to the cue (c) on the current trial (t) for latent state l. The learning rate is proportional to a quantity that captures the degree to which a learner believes that the current task conditions are captured by a given latent state, referred to as latent-state beliefs, $p_{l,t}$. The PE is specific to the latent state, whereby $\delta_{t,l} = outcome_t - V_{t,c,l}$ for cue c on trial t for latent state l. Trial-by-trial expectations $V_{t,c}$ of cue c are computed by taking a weighted average of $V_{t,c,l}$ with weights $p_{l,t}$. Following an outcome on a given trial, beliefs about current task conditions are updated (*delta* beliefs, dB). Larger updates in latent-state beliefs reflect larger changes in a learner's internal model of the current task rules (see Supplement for additional details).

Analyses

Computational Modeling

Skin conductance responses that were acquired during the Fear Conditioning and Extinction Task were used to identify optimal participant model parameter estimates. For each participant, model parameters were estimated by fitting computational models to SCR data from participants without any missing data (n = 74) via maximum likelihood estimation. Following convention, a square root transformation was applied the SCR (prior to the transformation, SCR values were rescaled between 0 and 1). Normalized, square root transformed SCR values were regressed onto trial-by-trial linear value estimation terms ($V_{t,c}$). Skin conductance responses were also regressed onto associability (α_t) for the hybrid model and onto updates in latent state beliefs (dB_t) for the LS model. Regression error was assumed to follow a normal distribution with mean zero and unknown variance. Maximum likelihood estimation was performed by minimizing squared regression error summed only over trials in which a shock was omitted (Li et al., 2011) using fmincon in Matlab (The MathWorks, Inc., Natick, MA). For each participant and RL model, estimation yielded regression coefficients and RL model parameters. Resulting log-likelihood values were compared to identify the best-fitting model. Additional regression parameters were included in exploratory analyses to identify whether the inclusion of non-linear terms would capture large trial-by-trial changes in SCR not readily captured by linear terms, thereby yielding better model fit (results provided in the Supplement).

Neuroimaging

All neuroimaging analyses focused on the LS model parameters (see Supplement for neuroimaging acquisition and preprocessing details). Trial-by-trial dBs, which characterize model-based RL updates, were carried forward to voxelwise and independent component analysis (ICA) to identify brain regions and networks that support implementation/encoding of latent-state updates among women with PTSD. Additionally, analyses focused on identifying whether PTSD symptom severity modulated dB-related encoding. Because individual-level parameters have previously been shown to be too noisy to yield reliable neuroimaging results (i.e., they exhibit high levels of error variance), learning parameters (e.g., V, PE, dB) were averaged across participants and the resulting trial-by-trial mean parameters were used in the neuroimaging analyses in accordance with previous research (Daw et al., 2005; Daw et al., 2011; Li et al., 2011; Schönberg et al., 2007; Schönberg et al., 2010). Of the 91participants who were eligible for the present study, 77 participants had viable neuroimaging data. However, one participant

was excluded due to missing clinical variables (final sample: n = 76; see the Supplement for information regarding the overlap between the computational modeling and neuroimaging samples). See Table 1 for demographic and clinical characteristics of participants included in the computational modeling and/or neuroimaging analyses (n = 85).

Voxelwise Analyses Participants' voxelwise time courses were regressed onto the design matrix using AFNI (3dREML; Cox, 1996). The design matrix included the stimulus onset and outcome phases of the task. The onset phase was parametrically modulated by trial-by-trial value expectation ($V_{t,c}$) from the LS model, and the outcome phase was parametrically modulated by trial-by-trial PE (positive and negative; $PE_{t,c}$) and latent-state belief updates ($dB_{t,c}$) from the LS model (Fig. 1b). To account for the potential impact of the electrotactile stimulation on neural activity, the design matrix also included a "shock" regressor. Because PE and the occurrence of the shock are highly correlated (Erdeniz et al., 2013), PE-related neural activity was not interpreted. Beta coefficients for the onset phase modulated by V and the outcome phase modulated by dB were included in the second-level analyses. Linear mixed effects models (LMEs) were implemented using MATLAB (fitlme; The MathWorks, Inc., Natick, MA) to test for main effects of neural activity during dB and V, controlling for age, IQ, and study site:

$$\mathrm{dB} \sim \mathrm{age} + \mathrm{IQ} + \textit{study site} + (1|\textit{subject}) \quad (1)$$

$$\mathrm{V} \sim \mathrm{age} + \mathrm{IQ} + \textit{study site} + (1|\textit{subject}) \quad (2)$$

Additionally, a model tested for potential unique effects of dB and V-related neural activity on CAPS-5 symptom severity. PE was included as a predictor, but PE results were not interpreted (for the reasons stated above):

$$\textit{CAPS} \sim \mathrm{V} + \mathrm{dB} + \mathrm{PE} + \mathrm{age} + \mathrm{IQ} + \textit{study site} + (1|\textit{subject}) \quad (3)$$

A set of follow-up analyses assessed for differential effects of acquisition versus extinction on the results of models 2 and 3. The design matrix included separate regressors for trial onset modulated by value expectation during acquisition ($V_{ta,c}$) and extinction ($V_{te,c}$), and separate regressors for outcome modulated by dB during acquisition ($dB_{ta,c}$) and extinction ($dB_{te,c}$). Because correlations between PE during extinction ($PE_{te,c}$) and $V_{te,c}$ and between $PE_{te,c}$ and $dB_{te,c}$ were highly correlated ($|r| > 0.70$), PE was not included in these analyses. Additionally, given that outcome-related neural activity associated with PE could not be separated from that of dB, LME tests focused on whether neural activity during V significantly differed during acquisition versus extinction (contrast coded: acquisition = 1, extinction = −1), controlling for age, IQ, and study site:

$$\mathrm{V} \sim \textit{contrast} + \mathrm{age} + \mathrm{IQ} + \textit{study site} + (1|\textit{subject}) \quad (4)$$

Separate models tested for potential unique effects of V and dB-related neural activity on CAPS symptom severity during acquisition and extinction.

$$\textit{CAPS} \sim \mathrm{V}_{\textit{acquisition}} + \mathrm{dB}_{\textit{acquisition}} + \mathrm{age} + \mathrm{IQ} + \textit{study site} + (1|\textit{subject}) \quad (5)$$

$$\textit{CAPS} \sim \mathrm{V}_{\textit{extinction}} + \mathrm{dB}_{\textit{extinction}} + \mathrm{age} + \mathrm{IQ} + \textit{study site} + (1|\textit{subject}) \quad (6)$$

Table 1 Participant demographic and clinical characteristics

	Participants	Arkansas site	Wisconsin site
n	85	39	46
Age, mean (SD)	33.7 (8.7)	36.1 (8.4)*	31.6 (8.5)
Race/ethnicity (%)			
White	74	69	78
Black/African American	17	23	11
Asian	0	0	0
Hispanic, Latina	4	0	7
Pacific Islander	0	0	0
Native American	0	0	0
Other	5	8	4
IQ, Mean (SD)	98.8 (20.6)	88.1 (21.6)*	107.5 (15.0)
CAPS symptom severity, mean (SD)	42.4 (11.2)	42.2 (11.5)	40.9 (10.9)
CAPS current # symptoms	14.3 (2.8)	13.8 (2.7)	14 (2.9)
PCL-C	43.6 (13.8)	45.6 (13.7)	41.9 (14.7)

IQ Intelligence Quotient, One-Word Receptive Vocabulary Test, *CAPS* Clinician- Administered PTSD Scale, *PCL-C* PTSD Checklist - Civilian Version

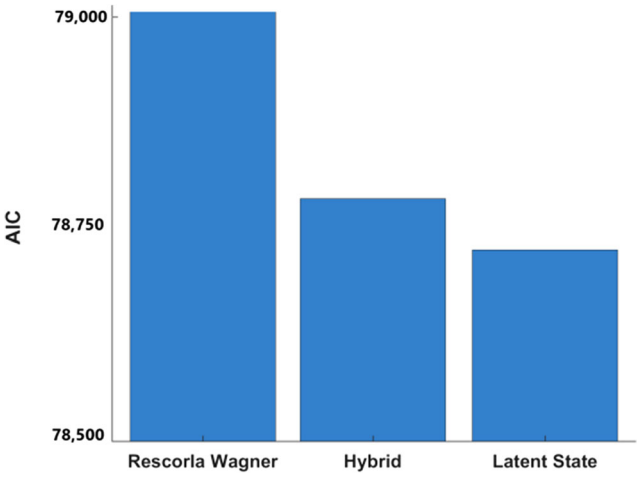


Fig. 2 Summed Akaike Information Criterion (AIC) values across participants showing that the Latent State model outperformed the Rescorla Wagner and Hybrid models (note: lower AIC values reflect better model fit)

Voxelwise comparisons were implemented within a sample specific grey matter mask and cluster-level thresholding controlled for voxelwise comparisons using an uncorrected $p < 0.001$ and cluster size $k \geq 18$, which was identified using AFNIs 3dClustSim.

Independent Component Analysis ICA was used to identify temporally coactivated spatially distributed neural large-scale networks and was implemented using the Group ICA of fMRI Toolbox (GIFT; Calhoun et al., 2001) in Matlab R2016a. A model order of 35 was selected to balance the tradeoff between component estimation reliability and interpretability. Thirteen of the 35 components were identified as functional networks theoretically related to learning or PTSD, including a left and right FPN that were of primary interest (22 networks that represented either motion artifact, CSF, or networks of non-interest such as motor cortex were excluded). Additionally, follow-up analyses were implemented with the remaining 11 networks (see Supplemental Figure S2). ICA timecourses were regressed onto the same design matrices described above using AFNI (3dREML; Cox, 1996) and resulting beta coefficients were included in the second-level analyses. The same series of LMEs described above were implemented for 1) each FPN and 2) follow-up networks using Matlab (fitlme; The MathWorks, Inc.). Bonferroni correction was applied for the two FPN networks ($p < 0.025$) and for the post-hoc analyses across the eleven additional networks ($p < 0.005$). Additionally, following an approach used by Erdeniz et al. (2013), several GLMs were fitted to participants' ICA time courses to identify whether the inclusion or removal of the PE and/or shock regressors altered the main effects of dB and/or V (see Supplement).

Results

Model Fit

The standard RW and hybrid models, which are nested models, were formally tested using a log-likelihood ratio test. Similar to previous studies (Boll et al., 2013; Homan et al., 2019; Li et al., 2011), the hybrid model outperformed the standard RW model, $\chi^2 = 515.43$, df = 148, $p < 0.001$. It also outperformed the additional RW models that were tested (see Supplement). Because the LS model does not contain the terms included in the hybrid model, a log-likelihood ratio test

Table 2 Regions associated with trial-by-trial changes in latent-state beliefs (dB) ouring outcome

Region	Cluster size (mm^3)	Peak t-value	Center of mass coordinates		
			X	Y	Z
R. IFG/Insula	1664	11.60	42	-8	16
R. Paracentral Lobule	1600	9.53	1	-23	59
L. Insula Lobe	1132	12.81	-34	-6	10
R. Calcarine Gyrus	153	6.81	18	-60	7
L. Postcentral/Precentral Gyrus	149	7.13	-43	-7	48
R. Inferior Occipital Gyrus	135	-5.51	41	-75	-3
L. Cerebellum	125	6.74	-5	-38	-20
Superior Orbital Gyrus	103	-6.89	-4	59	-21
L. Calcarine Gyrus	70	6.09	-15	-68	7
L. Superior Frontal Gyrus	63	-5.43	-14	39	47
Precuneus	57	-6.15	-2	-56	27
Cerebellum	34	5.47	1	-58	-36
L. Cerebellum	31	5.18	-31	-53	-28
R. Cuneus	29	4.61	13	-77	33
L. Inferior Occipital Gyrus	24	-6.48	-46	-70	-9
L. Middle Frontal Gyrus	23	-4.58	-33	15	47
L. Angular Gyrus	23	-4.72	-46	-66	29
L. Postcentral Gyrus	18	-4.60	-38	-28	53

Table 3 Regions associated with trial-by-trial associative strength (V) during stimulus onset

Region	Cluster size (mm^3)	Peak *t*-value	Center of mass coordinates		
			X	Y	Z
Paracentral Lobule	252	6.64	4	-23	66
R. Middle Frontal Gyrus	29	4.99	43	-3	50
L. Thalamus	26	5.42	-6	-19	7
L. Precuneus	21	-5.16	-7	-65	44

was not performed to compare these models. A comparison between Akaike Information Criterion values, which were summed across participants, revealed that the LS model provided a better fit than the standard RW and hybrid models (Fig. 2), as well as the additional RW models that were tested (Supplemental Figure S2).

Main Effects

Voxelwise Analyses

Table 2 lists the brain regions in which neural activation predicted latent-state belief updates. Activity in the left inferior frontal gyrus, left and right insula, right paracentral lobule, left and right calcarine gyrus, left postcentral/precentral gyrus, cerebellum, and right cuneus were positively related to latent-state belief updates (Fig. 3a). Activity in left and right inferior occipital gyrus, superior orbital gyrus, left superior frontal gyrus, precuneus, left middle frontal gyrus, and left angular gyrus were negatively related to latent-state belief updates. Table 3 lists the brain regions in which neural activation predicted value expectation. Greater value expectation-related activity emerged in the paracentral lobule, right middle frontal gyrus, and left thalamus (Fig. 3b). Lower value expectation-related activity emerged in the left precuneus.

Table 4 lists the brain regions in which the parameter-related neural activity uniquely predicted PTSD symptom severity. Figure 3c shows that lower latent state update-related activation within the right calcarine gyrus/right posterior cingulate cortex predicted higher CAPS scores. As shown in Fig. 3d, greater value expectation-related activation within the left angular gyrus/inferior parietal lobule predicted higher CAPS scores.

ICA

Table 5 lists the ICA networks that predicted trial-by-trial latent-state belief update encoding. Reduced encoding was evident in the left frontoparietal network (FPN), as well as the limbic, dorsomedial prefrontal cortex/posterior cingulate cortex, and hippocampal networks (Fig. 4a). Increased encoding was evident in the pre-supplementary motor area and striatal networks. Table 5 lists the ICA networks that predicted trial-by-trial value estimation encoding. Increased encoding emerged in the pre-supplementary motor area and

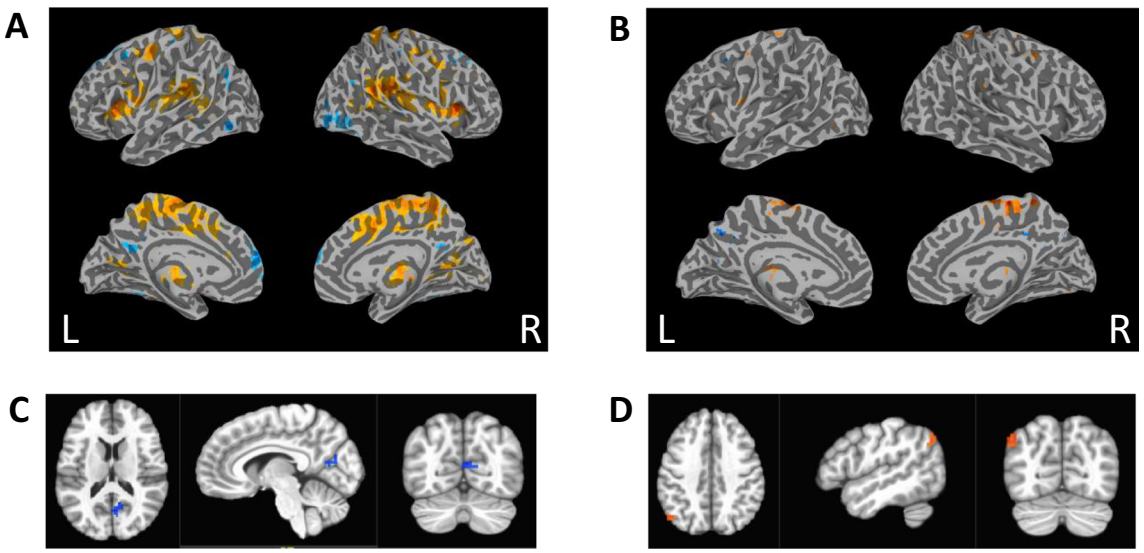


Fig. 3 Brain regions associated with (**a**) latent-state belief updates (dB) and (**b**) value estimations (V). Parameter-related neural activity that uniquely predicted PTSD symptom severity for (**c**) latent-state updates and (**d**) value estimations. L = left. Warm colors = positive z-values

Table 4 Parameter-related activity uniquely predictive of clinician-administered PTSD scale severity

Region	Parameter	Cluster size (mm^3)	Peak *t*-value	Center of mass coordinates		
				X	Y	Z
Left Angular Gyrus/IPL	V	27	4.68	-46	-66	36
Right Calcarine Gyrus/Right PCC	dB	22	-3.75	8	-62	15

striatal networks (Fig. 4b). Results for all networks are provided in the Supplement. A similar pattern of results emerged across study sites (Figure S5).

Table 6 lists the networks in which parameter-related encoding uniquely predicted PTSD symptom severity on the CAPS. Increased encoding of value expectation within the left FPN predicted greater PTSD symptom severity, $t(68) = 3.58$, $p < 0.001$. No other results held above correction.

Task Phase Effects (Acquisition vs. Extinction)

Voxelwise Analyses

Table 7 lists the brain regions in which neural activation differentially predicted value expectation during the acquisition versus extinction task phases. Greater activity during extinction than acquisition emerged in the middle cingulate cortex, thalamus, left fusiform gyrus, right lingual gyrus, left calcarine gyrus, left middle occipital gyrus, cuneus, and right lingual gyrus.

ICA

Table 8 lists the ICA networks that exhibited differential value expectation encoding during the acquisition versus extinction task phases. A significant effect of task phase emerged for value expectation encoding within both FPNs. Specifically, greater value expectation encoding was evident in the left and right FPN during acquisition relative to extinction, $t(146) = 2.53$, $p = 0.013$, and $t(146) = 2.62$, $p < 0.010$, respectively (Fig. 5). Significantly greater value expectation encoding was also evident in medial/lateral prefrontal cortex network during acquisition versus extinction, $t(146) = 3.38$, $p < 0.001$, whereas lower value estimation encoding was evident in hippocampal network during acquisition versus extinction, $t(146) = -2.92$, $p < 0.004$. Results followed a similar pattern across study site (Figure S6).

Table 9 lists the networks in which value expectation-related encoding uniquely predicted PTSD symptom severity on the CAPS during acquisition and extinction. As shown in Fig. 6a, during acquisition, but not extinction, greater value expectation encoding in the left FPN uniquely predicted higher PTSD symptom severity, $t(69) = 2.99$, $p = 0.004$. As shown in Fig. 6b, during extinction, but not acquisition, lower value expectation encoding in the dorsomedial prefrontal cortex/posterior cingulate cortex network uniquely predicted higher PTSD symptom severity, $t(69) = -4.05$, $p < 0.001$. A similar pattern of results was evident across study site (Figure S7).

Discussion

Results in this sample of adult women with IPV-related PTSD revealed that the latent-state model provided a better fit for

Table 5 Networks associated with trial-by-trial encoding of latent-state updates (dB) and value estimation (V)

Parameter	ICA Network	t-value	*p*-value
dB	L. Frontoparietal	-3.64	<0.001
	Pre-Supplementary Motor Area	6.70	<0.001
	Limbic	-7.44	<0.001
	Hippocampus	-3.14	0.003
	Dorsomedial PFC/Posterior Cingulate Cortex	-6.26	0.001
	Insula/Middle Frontal Gyrus	2.45	0.017
	Striatum	10.66	<0.001
V	Pre-Supplementary Motor Area	4.28	<0.001
	Limbic	-2.24	0.029
	Striatum	3.86	<0.001

Note: Networks in gray font did not survive Bonferroni correction

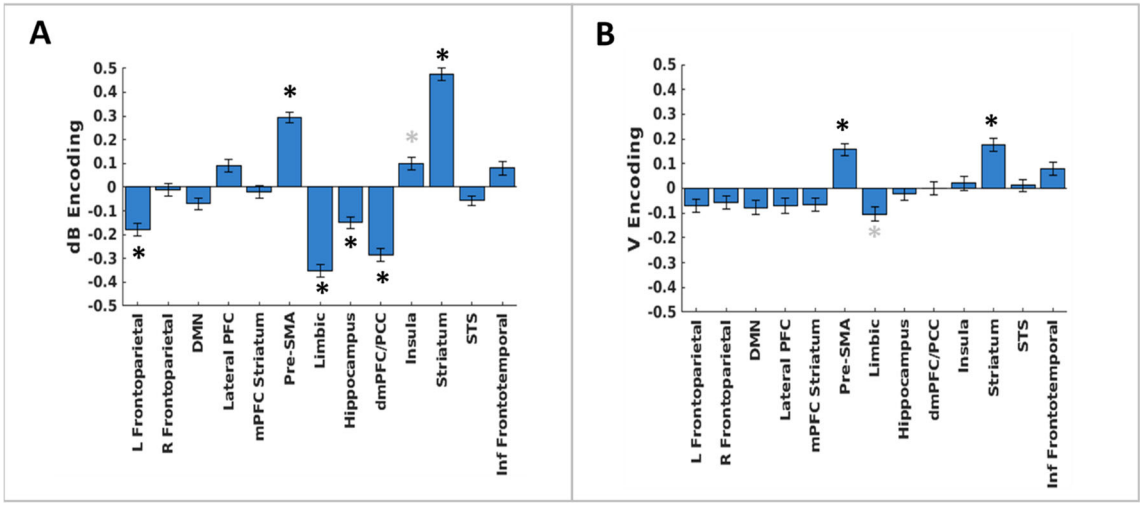


Fig. 4 Depiction of the main effects of encoding during trial-by-trial changes in (**a**) latent-state beliefs (dB) and (**b**) value estimation (V) for GLM model 1 (full model: dB, PE, V, and shock regressor included)

Table 6 Parameter-related activity uniquely predictive of clinician-administered PTSD scale severity

ICA Network	Parameter	*t*-value	*p* value
L. Frontoparietal	V	3.58	0.0007
Default mode	V	2.01	0.049
Dorsomedial PFC/posterior cingulate cortex	V	2.02	0.048
Inferior frontotemporal	V	-2.37	0.021

Note: Networks in gray font did not survive Bonferroni correction

participants' physiological responses during the fear conditioning task than the standard RW and hybrid models. Notably, the latter model is considered a "gold standard" for modeling fear conditioning responses, because it has been found to fit fear conditioning responsivity better than standard RW models in several previous studies (Homan et al., 2019; Li et al., 2011), which was replicated in the current study. However, it was outperformed by the latent-state model in the present study. This suggests that the latent-state model captures learning dynamics during fear conditioning and extinction that are not readily measured by the hybrid model, such as learning that varies as a function of task conditions.

Contrary to predictions, latent-state belief updates were associated with reduced (as opposed to increased) activity in regions previously associated with model-based RL and cognitive control, including the left FPN and left superior frontal

Table 7 Regions associated with trial-by-trial value expectation (V) during stimulus onset, acquisition versus extinction

Region	Cluster size (mm^3)	Peak t-value	Center of mass coordinates		
			X	Y	Z
Middle Cingulate Cortex	107	-5.68	-1	-40	35
Thalamus	62	-5.05	-1	-21	9
L. Fusiform Gyrus	60	-4.56	-40	-51	-14
R. Lingual Gyrus	49	-5.13	9	-44	3
L. Calcarine Gyrus	47	-4.45	-18	-64	9
L. Middle Occipital Gyrus	36	-5.27	-31	-84	5
Cuneus	28	-4.53	6	-67	24
R. Lingual Gyrus	27	-3.89	17	-58	3

Table 8 Networks associated with trial-by-trial encoding of value estimation (V) during acquisition versus extinction

Parameter	ICA Network	*t*-value	*p* value
V	L. Frontoparietal	2.53	0.013
	R. Frontoparietal	2.62	0.010
	Medial/Lateral PFC	3.38	<0.001
	Medial PFC/Striatum	-2.44	0.016
	Limbic	2.78	0.006
	Hippocampus	-2.92	0.004
	Dorsomedial PFC/Posterior Cingulate Cortex	2.51	0.013

Note: Regions in gray font did not hold above Bonferroni correction

gyrus. Similar to previous research on model-based RL, latent-state belief updates were related to increased striatal network activity (Daw et al., 2011; McDannald et al., 2011; but also see Gläscher et al., 2010). Latent-state belief updates were also related to increased activity of the left and right insula, which are implicated in punishment and loss-related learning (Palminteri et al., 2012). Although the bilateral anterior insula are components of the salience network, which is often associated with model-free PE encoding (Cisler et al., 2019; Preuschoff et al., 2008), this is not the first study to identify insula activation during model-based RL (Lee et al., 2014). Heightened insula activity during updates to latent-state beliefs may reflect sensitivity to unexpected and changing environmental demands that signal increased need for cognitive control processes during learning (Jiang et al., 2015).

Although the latent-state model provided a better overall fit to the SCR data than the other computational models that do not capture model-based RL processes, PTSD symptom severity was not predicted by FPN-related encoding of latent-state belief updates. Instead, higher PTSD symptom severity was uniquely related to reduced activity during latent-state updates in the right calcarine gyrus/posterior cingulate cortex, which are implicated in visual processing, visual imagery, and the focus of attention (Klein et al., 2000; Leech & Sharp, 2014). By contrast, increased value estimation-related encoding in the left FPN and activity of the angular gyrus/intraparietal lobule uniquely predicted greater PTSD symptom severity. It was further revealed that increased encoding within the left FPN during acquisition predicted greater PTSD symptom severity during acquisition, whereas reduced encoding within the dorsomedial prefrontal cortex (dmPFC/PCC) predicted greater PTSD symptom severity during extinction. Atypical FPN and dmPFC/PCC activity have been identified in previous studies of PTSD, including fear conditioning and extinction studies (for a review, see Suarez-Jimenez et al., 2020), although the direction of these effects is somewhat mixed in PTSD. It is important to note that, in contrast with prior studies that did not separate RL and non-RL sources of variance, we were able to identify effects for distinct RL processes using our model-based model. Heightened encoding of value estimation during the acquisition blocks (i.e., when threat expectancies are high) among cognitive control regions may contribute to enhanced representation of fear, while reduced encoding of value estimation during the extinction blocks (i.e., when threat expectancies are low) of the dmPFC/PCC may contribute to difficulties revising stimulus-outcome associations within a safe environment (Wang et al., 2014).

A notable limitation of the present study is the lack of a control comparison group. Although altered neural activity during latent-state belief updates did not emerge within cognitive control regions or networks in relation to PTSD

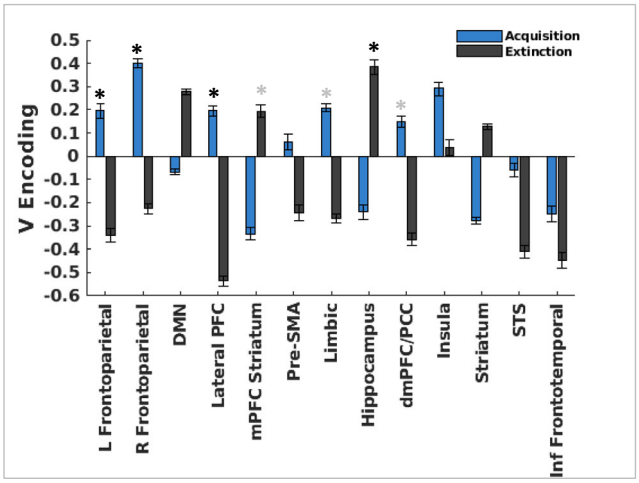
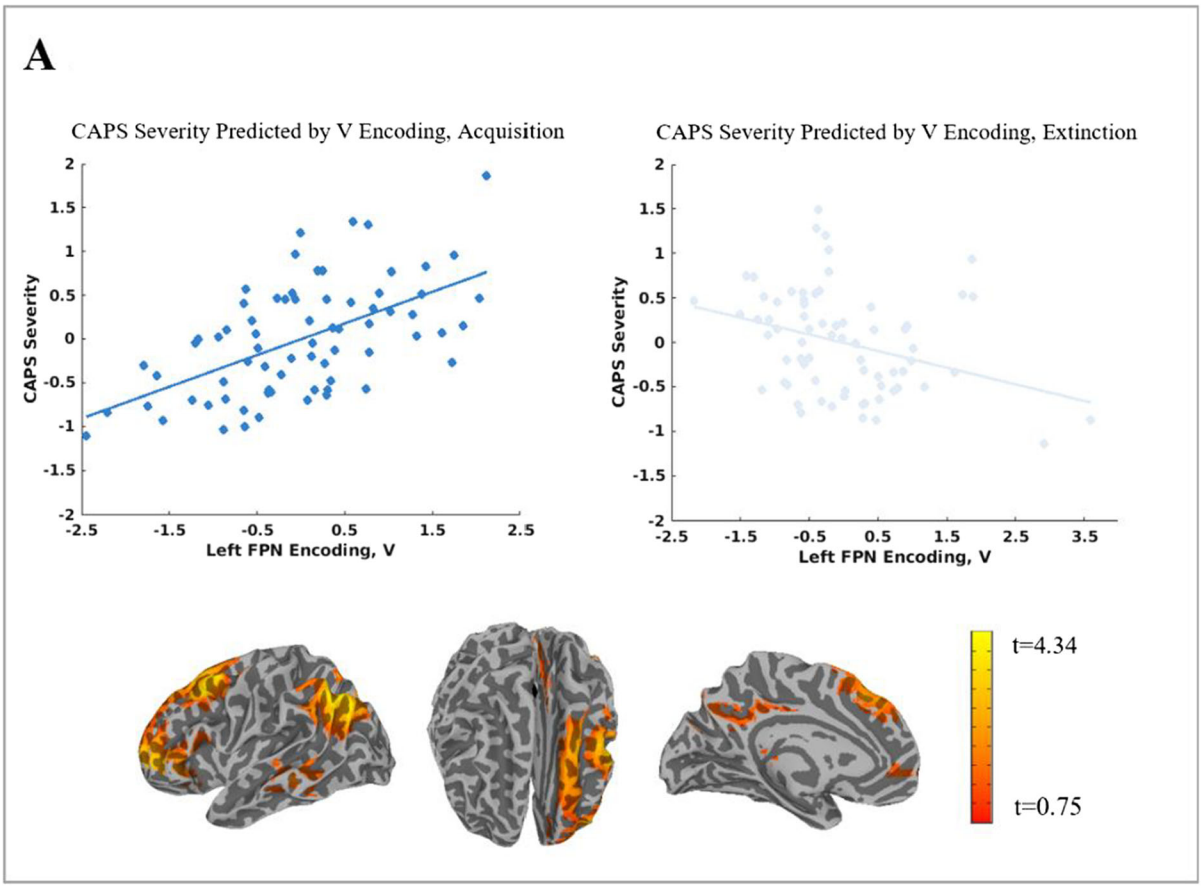
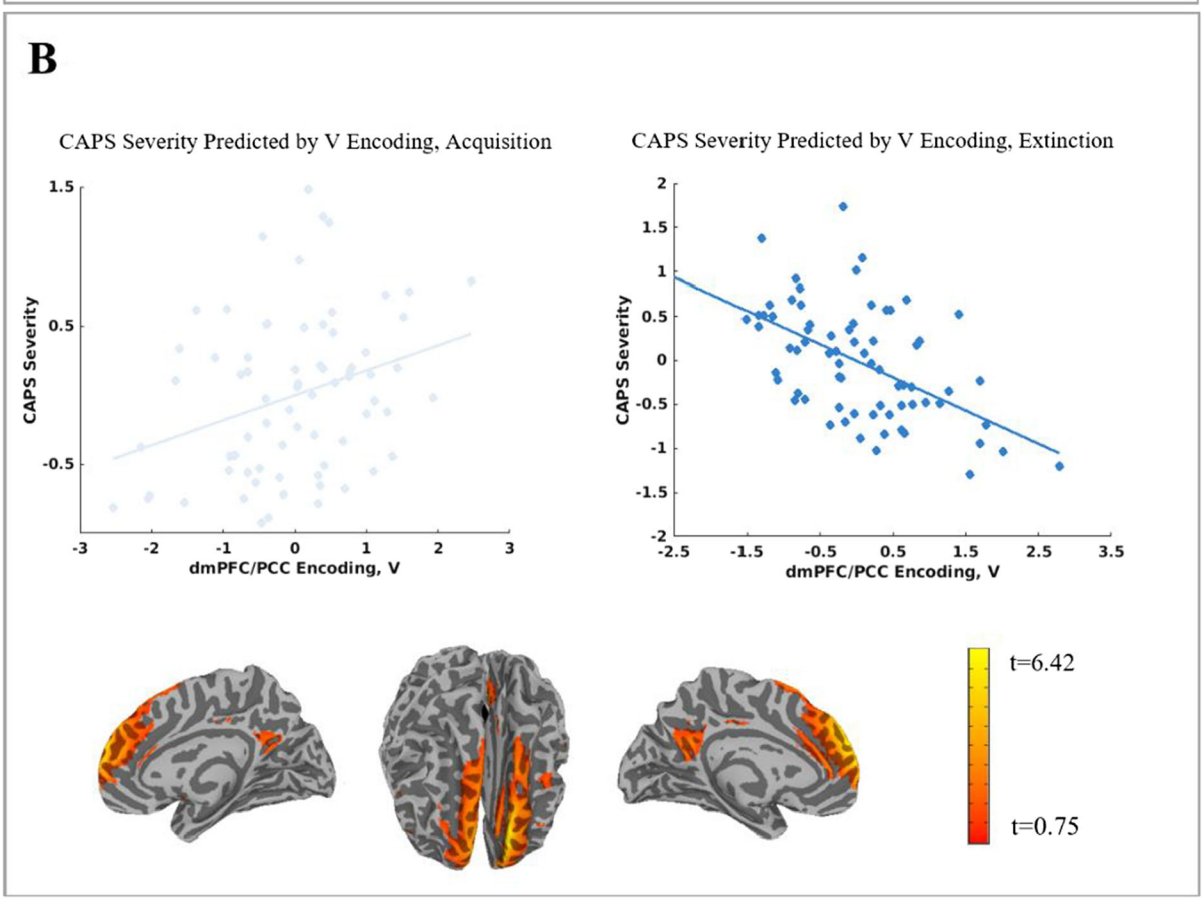


Fig. 5 Depiction of the effects of acquisition versus extinction on encoding during trial-by-trial changes in value estimation (V)

Table 9 V-related activity uniquely predictive of clinician-administered PTSD scale severity

ICA Network	*t*-value	*p* value
Acquisition		
L. Frontoparietal	3.45	0.001
Extinction		
Dorsomedial PFC/PCC	-4.05	<0.001


A
CAPS Severity Predicted by V Encoding, Acquisition
CAPS Severity
Left FPN Encoding, V
CAPS Severity Predicted by V Encoding, Extinction
CAPS Severity
Left FPN Encoding, V
t=4.34
t=0.75


B
CAPS Severity Predicted by V Encoding, Acquisition
CAPS Severity
dmPFC/PCC Encoding, V
CAPS Severity Predicted by V Encoding, Extinction
CAPS Severity
dmPFC/PCC Encoding, V
t=6.42
t=0.75

◀ **Fig. 6** **a** Scatterplots depicting relationships between left frontoparietal network encoding and PTSD symptom severity during (1) acquisition and (2) extinction, and a graphical depiction of the left frontoparietal ICA network. **b** Scatterplots depicting relationships between dorsomedial prefrontal cortex network encoding and PTSD symptom severity during (1) acquisition and (2) extinction, and a graphical depiction of the dorsomedial prefrontal cortex network ICA network. The dark blue scatterplots represent the relationships that reached the level of significance. The light blue (non-significant) scatterplots are provided for reference

symptoms, we did not test whether individuals with PTSD exhibit poorer encoding of latent-state belief updates within FPNs relative to individuals without PTSD. It is possible that PTSD symptoms scaled with value expectations, rather than latent-state belief updates, within the left FPN because overall individuals with PTSD generally had difficulty engaging contextual learning processes during fear conditioning. It will be important to establish the typical pattern of neural activity and encoding of value expectations and latent-state belief updates to further contextualize the meaning of present results. It also will be important to establish whether results extend to non-IPV related PTSD.

Overall, results provide some evidence that model-based RL processes that are altered during fear conditioning are related to PTSD symptoms among women with assault-related PTSD. While research has primarily been devoted to examining relatively simplistic learning processes, model-free and hybrid models cannot capture higher-level, context-related learning (e.g., learning when a stimulus is dangerous vs. safe), the latter of which can be captured by the model-based model and may be particularly important in the acquisition and revision of human fear. Although disruptions in model-based processes that scaled with PTSD symptoms occurred within brain regions or networks that were not anticipated (e.g., reduced encoding within visual processing regions during latent-state belief updates), our results provide preliminary evidence of model-based RL-related impairments in PTSD that are separate from other learning processes (e.g., value estimation). Our latent-state model also identified a pattern of value estimation encoding that is distinct from latent-state update encoding that may disrupt normative fear and safety learning among individuals with assault-related PTSD. Critically, exposure-based therapy, which is a "gold standard" treatment for PTSD, depends on learning processes to extinguish fear (Hermans et al., 2005), and RL deficits identified in this study may affect treatment response. Given that even the best available treatments for PTSD have limited efficacy, with remission occurring for approximately half of individuals who receive treatment (Morina et al., 2014; Resick et al., 2002; Schnurr et al., 2007), it is proposed that future research examine the role of RL impairment in treatment-related outcomes among individuals with PTSD (IPV and non-IPV) using a model-based framework.

Open Practices Statement None of the data or materials for the experiments reported here is available, and none of the experiments was preregistered.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13415-021-00943-4.

References

Alvarez, J. A., & Emory, E. (2006). Executive function and the frontal lobes: A meta-analytic review. *Neuropsychology Review, 16*(1), 17-42.

Bach, D.R. (2014). A head-to-head comparison of SCRalyze and Ledalab, two model-based methods for skin conductance analysis. *Biological Psychology*, *103*, 63-68.

Bach, D.R., & Friston, K.J. (2013). Model-based analysis of skin conductance responses: Towards causal models in psychophysiology. *Psychophysiology*, *50*(1), 15-22.

Bach, D.R., Flandin, G., Friston, K.J., & Dolan, R.J. (2010). Modelling event-related skin conductance responses. *International Journal of Psychophysiology*, *75*(3), 349-356.

Bach, D.R., Friston, K.J., & Dolan, R.J. (2013). An improved algorithm for model-based analysis of evoked skin conductance responses. *Biological Psychology*, *94*(3), 490-497.

Beierholm, U. R., Anen, C., Quartz, S., & Bossaerts, P. (2011). Separate encoding of model-based and model-free valuations in the human brain. *Neuroimage*, *58*(3), 955-962.

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., & Büchel, C. (2013). Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. *European Journal of Neuroscience*, *37*(5), 758-767.

Breslau, N., Kessler, R. C., Chilcoat, H. D., Schultz, L. R., Davis, G. C., & Andreski, P. (1998). Trauma and posttraumatic stress disorder in the community: The 1996 Detroit Area Survey of Trauma. *Archives of General Psychiatry, 55*(7), 626-632.

Brown, V. M., Zhu, L., Wang, J. M., Frueh, B. C., King-Casas, B., & Chiu, P. H. (2018). Associability-modulated loss learning is increased in posttraumatic stress disorder. *Elife*, *7*, e30150.

Brownell, R. (2000). Expressive and receptive one-word picture vocabulary tests (EOWPVT, ROWPVT). Psychological Corporation.

Calhoun, V. D., Adali, T., Pearlson, G. D., & Pekar, J. J. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, *14*(3), 140-151.

Cisler, J. M., Begle, A. M., Amstadter, A. B., Resnick, H. S., Danielson, C. K., Saunders, B. E., & Kilpatrick, D. G. (2012). Exposure to interpersonal violence and risk for PTSD, depression, delinquency, and binge drinking among adolescents: Data from the NSA-R. *Journal of Traumatic Stress, 25*(1), 33-40.

Cisler, J. M., Bush, K., Steele, J. S., Lenow, J. K., Smitherman, S., & Kilts, C. D. (2015). Brain and behavioral evidence for altered social learning mechanisms among women with assault-related posttraumatic stress disorder. *Journal of Psychiatric Research*, *63*, 75-83.

Cisler, J. M., Esbensen, K., Sellnow, K., Ross, M., Weaver, S., Sartin-Tarm, A., Herringa, R. J., & Kilts, C. D. (2019). Differential roles of the salience network during prediction error encoding and facial emotion processing among female adolescent assault victims. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging, 4,* 371-380.

Cisler, J. M., Privratsky, A. A., Sartin-Tarm, A., Sellnow, K., Ross, M., Weaver, S., Hahn, E., Herringa, R. J., James, G .A., & Kilts, C. D.

(2020). l-DOPA and consolidation of fear extinction learning among women with posttraumatic stress disorder. *Translational Psychiatry*, *10*(1), 1-11.

Cochran, A. L. & Cisler, J. M. (2019). A flexible and generalizable model of online latent-state learning. *PLoS Computational Biology, 15,* e1007331.

Cools, R., Clark, L., & Robbins, T. W. (2004). Differential responses in human striatum and prefrontal cortex to changes in object and rule relevance. *Journal of Neuroscience*, *24*(5), 1129-1135.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29,* 162-173.

Culbreth, A. J., Westbrook, A., Daw, N. D., Botvinick, M., & Barch, D. M. (2016). Reduced model-based decision-making in schizophrenia. *Journal of Abnormal Psychology, 125*(6), 777.

Davis, T., Goldwater, M., & Giron, J. (2017). From concrete examples to abstract relations: The rostrolateral prefrontal cortex integrates novel examples into relational categories. *Cerebral Cortex*, *27*(4), 2652-2670.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, *8*, 1704-1711.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204-1215.

Doll, B. B., Bath, K. G., Daw, N. D., & Frank, M. J. (2016). Variability in dopamine genes dissociates model-based and model-free reinforcement learning. *Journal of Neuroscience, 36*(4), 1211-1222

Dunsmoor, J. E., & Murphy, G. L. (2015). Categories, concepts, and conditioning: how humans generalize fear. *Trends in Cognitive Sciences*, *19*(2), 73-77.

Erdeniz, B., Rohe, T., Done, J., & Seidler, R. (2013). A simple solution for model comparison in bold imaging: the special case of reward prediction error and reward outcomes. *Frontiers in Neuroscience*, *7*, 116.

Fani, N., Tone, E. B., Phifer, J., Norrholm, S. D., Bradley, B., Ressler, K. J., Kamkwalala, A., & Jovanovic, T. (2012). Attention bias toward threat is associated with exaggerated fear expression and impaired extinction in PTSD. *Psychological Medicine, 42*, 533–543.

Fogelson, N., Shah, M., Scabini, D., & Knight, R. T. (2009). Prefrontal cortex is critical for contextual processing: evidence from brain lesions. *Brain*, *132*(11), 3002-3010.

Frans, Ö., Rimmö, P. A., Åberg, L., & Fredrikson, M. (2005). Trauma exposure and post-traumatic stress disorder in the general population. *Acta Psychiatrica Scandinavica, 111*(4), 291-290.

Garfinkel, S. N., Abelson, J. L., King, A. P., Sripada, R. K., Wang, X., Gaines, L. M., & Liberzon, I. (2014). Impaired contextual modulation of memories in PTSD: an fMRI and psychophysiological study of extinction retention and fear renewal. *Journal of Neuroscience*, *34*(40), 13435-13443.

Gläscher, J., Daw, N., Dayan, P., & O'Doherty, J. P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron, 66,* 585-595.

Haaker, J., Gaburro, S., Sah, A., Gartmann, N., Lonsdorf, T. B., Meier, K., Singewald, N., Pape, H. C., Morellini, F., & Kalisch, R. (2013). Single dose of L-dopa makes extinction memories context-independent and prevents the return of fear. *Proceedings of the National Academy of Sciences*, *110*(26), E2428-E2436.

Hermans, D., Dirikx, T., Vansteenwegenin, D., Baeyens, F., Van den Bergh, O., & Eelen, P. (2005). Reinstatement of fear responses in human aversive conditioning. *Behaviour Research and Therapy, 43*(4), 533-551.

Homan, P., Levy, I., Feltham, E., Gordon, C., Hu, J., Li, J., Pietrzak, R. H., Southwick, S., Krystal, J. H., Harpaz-Rotem, I., & Schiller, D. (2019). Neural computations of threat in the aftermath of combat trauma. *Nature Neuroscience*, *22*(3), 470-476.

Huys, Q. J., Maia, T. V., & Frank, M. J. (2016). Computational psychiatry as a bridge from neuroscience to clinical applications. *Nature Neuroscience*, *19*(3), 404-413.

Iverson, K. M., Dick, A., McLaughlin, K. A., Smith, B. N., Bell, M. E., Gerber, M. R., Cook, N., & Mitchell, K. S. (2013). Exposure to interpersonal violence and its associations with psychiatric morbidity in a U.S. national sample: A gender comparison. *Psychology of Violence, 3*(3), 273–287.

Jiang, J., Beck, J., Heller, K., & Egner, T. (2015). An insula-frontostriatal network mediates flexible cognitive control by adaptively predicting changing control demands. *Nature Communications*, *6*(1), 1-11.

Jovanovic, T., Norrholm, S. D., Fennell, J. E., Keyes, M., Fiallos, A. M., Myers, K. M., Davis, M., & Duncan, E. J. (2009). Posttraumatic stress disorder may be associated with impaired fear inhibition: Relation to symptom severity. *Psychiatry Research*, *167*(1-2), 151-160.

Jovanovic, T., Norrholm, S. D., Blanding, N. Q., Davis, M., Duncan, E., Bradley, B., & Ressler, K. J. (2010). Impaired fear inhibition is a biomarker of PTSD but not depression. *Depression and Anxiety*, *27*(3), 244-251.

Jovanovic, T., Kazama, A., Bachevalier, J., & Davis, M. (2012). Impaired safety signal learning may be a biomarker of PTSD. *Neuropharmacology*, *62*(2), 695-704.

Kaczkurkin, A. N., Burton, P. C., Chazin, S. M., Manbeck, A. B., Espensen-Sturges, T., Cooper, S. E., Sponheim, S. R., & Lissek, S. (2017). Neural substrates of overgeneralized conditioned fear in PTSD. *American Journal of Psychiatry*, *174*(2), 125-134.

Kelley, L. P., Weathers, F. W., McDevitt-Murphy, M. E., Eakin, D. E., & Flood, A. M. (2009). A comparison of PTSD symptom patterns in three types of civilian trauma. *Journal of Traumatic Stress, 22*(3), 227-235.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. *Archives of General Psychiatry, 62*(6), 617-627.

Kessler, R. C., Aguilar-Gaxiola, S., Alonso, J., Benjet, C., Bromet, E. J., Cardoso, G., Degenhardt, L., de Girolamo, G., Dinolova, R. V., Ferry, F., Florescu, S., Gureje, O., Haro, J. M., Huang, Y., Karam, E. G., Kawakami, N., Lee, S., Lepine, J. P., Levinson, D., … Koenen, K. C. (2017). Trauma and PTSD in the WHO world mental health surveys. *European Journal of Psychotraumatology, 8*(sup5), 1353383.

Kilpatrick, D. G., Resnick, H. S., Milanak, M. E., Miller, M. W., Keyes, K. M., & Friedman, M. J. (2013). National estimates of exposure to traumatic events and PTSD prevalence using DSM-IV and DSM-5 criteria. *Journal of Traumatic Stress, 26*(5), 537-547.

Klein, I., Paradis, A. L., Poline, J. B., Kosslyn, S. M., & Le Bihan, D. (2000). Transient activity in the human calcarine cortex during visual-mental imagery: An event-related fMRI study. *Journal of Cognitive Neuroscience*, *12*, 15-23.

Le Pelley, M. E. (2004). The role of associative history in models of associative learning: A selective review and a hybrid model. *The Quarterly Journal of Experimental Psychology Section B*, *57*(3b), 193-243.

Lee, S. W., Shimojo, S., & O'Doherty, J. P. (2014). Neural computations underlying arbitration between model-based and model-free learning. *Neuron*, *81*(3), 687-699.

Leech, R., & Sharp, D. J. (2014). The role of the posterior cingulate cortex in cognition and disease. *Brain*, *137*(1), 12-32.

Leskin, L. P., & White, P. M. (2007). Attentional networks reveal executive function deficits in posttraumatic stress disorder. *Neuropsychology, 21*, 275–284.

Letkiewicz, A. M., Cochran, A. L., & Cisler, J. M. (2020). Frontoparietal network activity during model-based reinforcement learning updates

is reduced among adolescents with severe sexual abuse. *Journal of Psychiatric Research* S0022-3956(20)31067-0

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience*, *14*(10), 1250-1252.

Lissek, S., & van Meurs, B. (2015). Learning models of PTSD: Theoretical accounts and psychobiological evidence. *International Journal of Psychophysiology*, *98*(3), 594-605.

Lopresto, D., Schipper, P., & Homberg, J. R. (2016). Neural circuits and mechanisms involved in fear generalization: implications for the pathophysiology and treatment of posttraumatic stress disorder. *Neuroscience & Biobehavioral Reviews*, *60*, 31-42.

Martin, N. A., & Brownell, R. (2011). Receptive One-Word Picture Vocabulary Test (4th ed.). (ROWPVT-4). Academic Therapy Publications.

McDannald, M. A., Lucantonio, F., Burke, K. A., Niv, Y., & Schoenbaum, G. (2011). Ventral striatum and orbitofrontal cortex are both required for model-based, but not model-free, reinforcement learning. *Journal of Neuroscience*, *31*(7), 2700-2705.

Mihatsch, O., & Neuneier, R. (2002). Risk-sensitive reinforcement learning. *Machine Learning*, *49*(2-3), 267-290.

Morina, N., Wicherts, J. M., Lobbrecht, J., & Priebe, S. (2014). Remission from post-traumatic stress disorder in adults: a systematic review and meta-analysis of long-term outcome studies. *Clinical Psychology Review*, *34*(3), 249-255.

Pacella, M. L., Hruska, B., & Delahanty, D. L. (2013). The physical health consequences of PTSD and PTSD symptoms: A meta-analytic review. *Journal of Anxiety Disorders*, *27*(1), 33-46.

Palminteri, S., Justo, D., Jauffret, C., Pavlicek, B., Dauta, A., Delmaire, C., Czernecki, V., Karachi, C., Capelle, L., Durr, A., & Pessiglione, M. (2012). Critical roles for anterior insula and dorsal striatum in punishment-based avoidance learning. *Neuron*, *76*(5), 998-1009.

Polak, A. R., Witteveen, A. B., Reitsma, J. B., & Olff, M. (2012). The role of executive function in posttraumatic stress disorder: A systematic review. *Journal of Affective Disorders*, *141*(1), 11-21.

Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*(11), 2745-2752.

Price, R. B., Brown, V., & Siegle, G. J. (2019). Computational modeling applied to the dot-probe task yields improved reliability and mechanistic insights. *Biological Psychiatry*, *85*(7), 606-612.

Privratsky, A.A., Bush, K.A., Bach, D.R., Hahn, E.M., & Cisler, J.M. (2020). Filtering and model-based analysis independently improve skin-conductance response measures in the fMRI environment: Validation in a sample of women with PTSD. *International Journal of Psychophysiology*, *158*, 86-95.

Raij, T., Nummenmaa, A., Marin, M. F., Porter, D., Furtak, S., Setsompop, K., & Milad, M. R. (2018). Prefrontal cortex stimulation enhances fear extinction memory in humans. *Biological Psychiatry*, *84*(2), 129-137.

Redish, A. D., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*(3), 784.

Resnick, H. S., Kilpatrick, D. G., Dansky, B. S., Saunders, B. E., & Best, C. L. (1993). Prevalence of civilian trauma and posttraumatic stress disorder in a representative national sample of women. *Journal of Consulting and Clinical Psychology, 61*, 984–991.

Resick, P. A., Nishith, P., Weaver, T. L., Astin, M. C., & Feuer, C. A. (2002). A comparison of cognitive-processing therapy with prolonged exposure and a waiting condition for the treatment of chronic posttraumatic stress disorder in female rape victims. *Journal of Consulting and Clinical Psychology*, *70*(4), 867.

Ross, M. C., Lenow, J. K., Kilts, C. D., & Cisler, J. M. (2018). Altered neural encoding of prediction errors in assault-related posttraumatic stress disorder. *Journal of Psychiatric Research, 103*, 83-90.

Schnurr, P. P., Friedman, M. J., Engel, C. C., Foa, E. B., Shea, M. T., Chow, B. K., Resick, P.A., Thurston, V., Orsillo, S.M., Haug, R. and Turner, C., & Bernardy, N. (2007). Cognitive behavioral therapy for posttraumatic stress disorder in women: A randomized controlled trial. *Jama, 297*, 820–830.

Schönberg, T., Daw, N. D., Joel, D., & O'Doherty, J. P. (2007). Reinforcement learning signals in the human striatum distinguish learners from nonlearners during reward-based decision making. *Journal of Neuroscience*, *27*(47), 12860-12867.

Schönberg, T., O'Doherty, J. P., Joel, D., Inzelberg, R., Segev, Y., & Daw, N. D. (2010). Selective impairment of prediction error signaling in human dorsolateral but not ventral striatum in Parkinson's disease patients: evidence from a model-based fMRI study. *Neuroimage*, *49*(1), 772-781.

Steiger, F., Nees, F., Wicking, M., Lang, S., & Flor, H. (2015). Behavioral and central correlates of contextual fear learning and contextual modulation of cued fear in posttraumatic stress disorder. *International Journal of Psychophysiology, 98(3)*, 584-593.

Stein, M. B., Kennedy, C. M., & Twamley, E. W. (2002). Neuropsychological function in female victims of intimate partner violence with and without posttraumatic stress disorder. *Biological Psychiatry*, *52*(11), 1079-1088.

Stephan, K. E., & Mathys, C. (2014). Computational approaches to psychiatry. *Current Opinion in Neurobiology*, *25*, 85-92.

Suarez-Jimenez, B., Albajes-Eizagirre, A., Lazarov, A., Zhu, X., Harrison, B. J., Radua, J., Neria, Y., & Fullana, M. A. (2020). Neural signatures of conditioning, extinction learning, and extinction recall in posttraumatic stress disorder: a meta-analysis of functional magnetic resonance imaging studies. *Psychological Medicine, 50*(9), 1442-1451.

Van Otterlo, M., & Wiering, M. (2012). Reinforcement learning and Markov decision processes. In: *Reinforcement Learning* (pp. 3-42). Springer.

Wang, Q., Luo, S., Monterosso, J., Zhang, J., Fang, X., Dong, Q., & Xue, G. (2014). Distributed value representation in the medial prefrontal cortex during intertemporal choices. *Journal of Neuroscience*, *34*(22), 7522-7530.

Weathers, F. W., Bovin, M. J., Lee, D. J., Sloan, D. M., Schnurr, P. P., Kaloupek, D. G., Keane, T. M., & Marx, B. P. (2018). The Clinician-Administered PTSD Scale for DSM–5 (CAPS-5): Development and initial psychometric evaluation in military veterans. *Psychological Assessment*, *30*(3), 383-395.

Woon, F. L., Farrer, T. J., Braman, C. R., Mabey, J. K., & Hedges, D. W. (2017). A meta-analysis of the relationship between symptom severity of posttraumatic stress disorder and executive function. *Cognitive Neuropsychiatry*, *22*(1), 1-16.

Wunderlich, K., Dayan, P., & Dolan, R. J. (2012). Mapping value based planning and extensively trained choice in the human brain. *Nature Neuroscience*, *15*(5), 786-791.