Asymmetric representation of aversive prediction errors in Pavlovian threat conditioning

Karita E. Ojala , Athina Tzovara , Benedikt A. Poser ,
Antoine Lutti , Dominik R. Bach

PII: S1053-8119(22)00694-2
DOI: https://doi.org/10.1016/j.neuroimage.2022.119579
Reference: YNIMG 119579

To appear in: *NeuroImage*

Received date: 6 January 2022
Revised date: 16 August 2022
Accepted date: 17 August 2022

Please cite this article as: Karita E. Ojala , Athina Tzovara , Benedikt A. Poser , Antoine Lutti , Dominik R. Bach , Asymmetric representation of aversive prediction errors in Pavlovian threat conditioning, *NeuroImage* (2022), doi: https://doi.org/10.1016/j.neuroimage.2022.119579

Asymmetric representation of aversive prediction errors in Pavlovian threat conditioning

Karita E. Ojala[1, 2] * [CO], Athina Tzovara[1-3 CO], Benedikt A. Poser[4], Antoine Lutti[5] and Dominik R. Bach[1, 2, 6] *

[1] Computational Psychiatry Research, Department of Psychiatry, Psychotherapy and Psychosomatics, Psychiatric Hospital, University of Zurich, Lenggstrasse 31, 8032 Zurich, Switzerland

[2] Neuroscience Centre Zurich, University of Zurich, Winterthurerstrasse 190, 8057 Zürich, Switzerland

[3] Institute of Computer Science, University of Bern, Neubrückstrasse 10, 3012 Bern, Switzerland

[4] Department of Cognitive Neuroscience, Faculty of Psychology and Neuroscience, Maastricht University, Oxfordlaan 55, 6299 EV Maastricht, The Netherlands

[5] Laboratory for Research in Neuroimaging, Department of Clinical Neuroscience, Lausanne University Hospital and University of Lausanne, Chemin de Mont-Paisible 16, 1011 Lausanne, Switzerland

[6] Wellcome Centre for Human Neuroimaging and Max-Planck UCL Centre for Computational Psychiatry and Ageing Research, University College London, 10-12 Russell Square, WC1B 5EH London, United Kingdom

* Corresponding authors, CO = Equal contribution/co-first authors

Corresponding authors:

Karita E. Ojala, Institute of Systems Neuroscience, Center for Experimental Medicine, University Medical Center Hamburg-Eppendorf, Martinistrasse 52, 22046 Hamburg, Germany, k.ojala@uke.de

Dominik R. Bach, Hertz Chair for Artificial Intelligence and Neuroscience, Transdisciplinary Research Area "Life and Health", University of Bonn, Am Probsthof 49, 53121 Bonn, Germany, d.bach@uni-bonn.de

Competing interests: Authors report no competing interests.

Abstract

Survival in biological environments requires learning associations between predictive sensory cues and threatening outcomes. Such aversive learning may be implemented through reinforcement learning algorithms that are driven by the signed difference between expected and encountered outcomes, termed prediction errors (PEs). While PE-based learning is well established for reward learning, the role of putative PE signals in aversive learning is less clear. Here, we used functional magnetic resonance imaging in humans (21 healthy men and women) to investigate the neural representation of PEs during maintenance of learned aversive associations. Four visual cues, each with a different probability (0, 33, 66, 100%) of being followed by an aversive outcome (electric shock), were repeatedly presented to participants. We found that neural activity at omission (US-) but not occurrence of the aversive outcome (US+) encoded PEs in the medial prefrontal cortex. More expected omission of aversive outcome was associated with lower neural activity. No neural signals fulfilled axiomatic criteria, which specify necessary and sufficient components of PE signals, for signed PE representation in a whole-brain search or in a-priori regions of interest. Our results might

suggest that, different from reward learning, aversive learning does not involve signed PE signals that are represented within the same brain region for all conditions.

Key words: aversive prediction errors, threat learning, axiomatic conditions, reinforcement learning, fMRI

Introduction

Learning from aversive experiences benefits long-term survival by improving an organism's capacity to avoid threatening situations (Seymour, 2019). Reinforcement learning theory prescribes how violations of prior expectations, termed prediction errors (PE), might drive associative cue-outcome learning (Rescorla and Wagner, 1972). While PE signals in dopaminergic midbrain circuits are required for appetitive learning (Chang et al., 2017; Schultz and Dickinson, 2000; Steinberg et al., 2013), the same is not established for aversive learning. During Pavlovian threat conditioning, also termed fear conditioning, neurons in periaqueductal gray (PAG) and lateral amygdala (LA) reduce firing to a repeated unconditioned stimulus (US), possibly due to progressive inhibition from central amygdala (Groessl et al., 2018; Johansen et al., 2010; Ozawa et al., 2017). This neural firing could reflect positive PE signals for "more aversive than expected" outcomes, which correspond here to US occurence. However, it is less clear which neural populations signal positive aversive PEs once US probabilities are learned, as established for appetitive PE signals (Lak et al., 2016), which pathways convey putative PE signals from PAG to LA, and where and how negative aversive PE signals (i.e., responses to US omission) are expressed (Herry and Johansen, 2014). These gaps limit our computational understanding of the neural circuits that underlie aversive conditioning.

In a search for formal learning mechanisms, computational neuroimaging studies have often committed to specific learning models and assumed a linear mapping of positive and negative PEs to neural signals. They have then regressed model-derived PEs onto blood-oxygen-dependent level (BOLD) signal and found correlations in striatum, a target region of reward PE-expressing midbrain neurons (Boll et al., 2013; Li et al., 2011; Seymour et al., 2004; Zhang et al., 2016), but also in the insula, periaqueductal grey, substantia nigra/ventral tegmental area, ventromedial prefrontal cortex, dorsolateral prefrontal cortex, orbitofrontal cortex, anterior cingulate cortex, middle cingulate cortex, thalamus, and amygdala (Dunsmoor et al., 2008; Pauli et al., 2015; Roy et al., 2014; Seymour et al., 2005, 2004; Spoormaker et al., 2011; Zhang et al., 2016). Although a powerful tool if the learning model is correct and explains all data, this approach has two limitations: first, its sensitivity might be reduced if the a priori chosen learning model does not closely correspond to the true learning model. Second, significant correlation between PE and neural signal can be driven by a strong relation only in some experimental conditions and no relation in others, such that the neural signal may not comply with computational requirements of reinforcement learning theory.

In contrast to the positive prediction error signals at US occurrence possibly serving to drive aversive learning, it has been suggested that negative prediction error signals at US omission might engage inhibitory extinction learning (Li and McNally, 2014). Extinction learning is driven by omission of a reinforcer, which is, at least categorically, encoded in the activity of dopaminergic midbrain neurons (Luo et al., 2018; Salinas-Hernández et al., 2018), and in humans in dopamine-dependent activity patterns in ventromedial prefrontal cortex (Esser et al., 2021; Gerlicher et al., 2018). Since dopaminergic midbrain neurons are known to encode reward prediction errors, and the omission of punishment could be seen as a reward, there is a possibility

that these omission responses are parametrically expressed and formally correspond to negative prediction error signals.

In order to distinguish different forms of parametric US occurrence and US omission responses, previous work has identified three general criteria, or 'axioms', that must be fulfilled for a PE signal in any computational learning algorithm (Caplin and Dean, 2008): 1) Occurrence of an outcome results in higher signal than omission of an outcome (or: higher intensity outcomes result in higher signal than lower intensity outcomes), 2) Unexpected outcomes result in higher signal than expected outcomes, 3) Fully expected outcomes result in no/equal signal regardless of outcome type. Signals that adhere to these axioms have been observed in appetitive Pavlovian conditioning (Hart et al., 2014; Rutledge et al., 2010) as well as in aversive instrumental conditioning, and in learning to predict pain intensities (Roy et al., 2014). It remains unknown whether these criteria are also fulfilled within any brain region in Pavlovian threat conditioning.

Here, we investigated neural PE signals to US outcomes that had previously been associated with predictive CS in an Pavlovian threat conditioning procedure. We used two distinct outcomes (US+: US occurrence; US−: US omission) and 4 conditioned stimuli (CS) with distinct rates of receiving the US+ (0%, 33%, 66%, 100%). This design allowed us to analyse PE signals after US occurrence as well as omission, without commitment to any particular learning model.

Materials and Methods

Participants

Twenty-one participants (6 women and 15 men; mean age ± SD: 25.5±4.2) were recruited from the general and student population for an fMRI experiment, and 19 participants (14 women, 5 men, mean age 24.7±3.7 years) for a behavioral experiment with the same experimental design. One participant in the behavioral experiment was excluded due to pupil data quality (see details below). Participants reported that they had no history of neurological and psychiatric illnesses and gave written informed consent. The study protocol, including the form of written consent, was in accordance with the Declaration of Helsinki and approved by the governmental research ethics committee (Kantonale Ethikkommission Zürich, 2016-00097).

Experimental design

In both experiments, participants underwent delay threat conditioning with four visual CS, which were triangles of different color (Fig. 1A). Each CS was associated with a distinct US rate: 0%, 33%, 66%, or 100% (Fig. 1B). US was an aversive electric shock to the right forearm, ending concurrently with the CS. The assignment of CS color to US rate was randomly determined for each participant. US started 6 seconds after CS onset, lasted 0.5 seconds, and co-terminated with the CS. This CS-US interval was chosen such that the canonical BOLD response to CS, and to US, are approximately uncorrelated. The intertrial interval was randomly drawn from {5 s, 6 s, 6 s, 7 s}, i.e., 6 s was twice as likely as the other values. During CS

presentation, participants were instructed to indicate CS color with a button or key press (button box in the fMRI experiment, keyboard in the behavioral experiment), in order to maintain attention during the task. CS color and button or key association was randomized across participants. Before the experiment started, participants trained the CS color-key press mapping (for fMRI: inside the scanner) until 80% accuracy over at least two presentations of each CS was reached. Participants were explicitly informed that after training, all CS may be followed by US but received no information about CS-US contingencies. To exclude potential confounds for fMRI analysis, we ensured there was no evidence that reaction time or accuracy depended on CS condition (see Table 1).

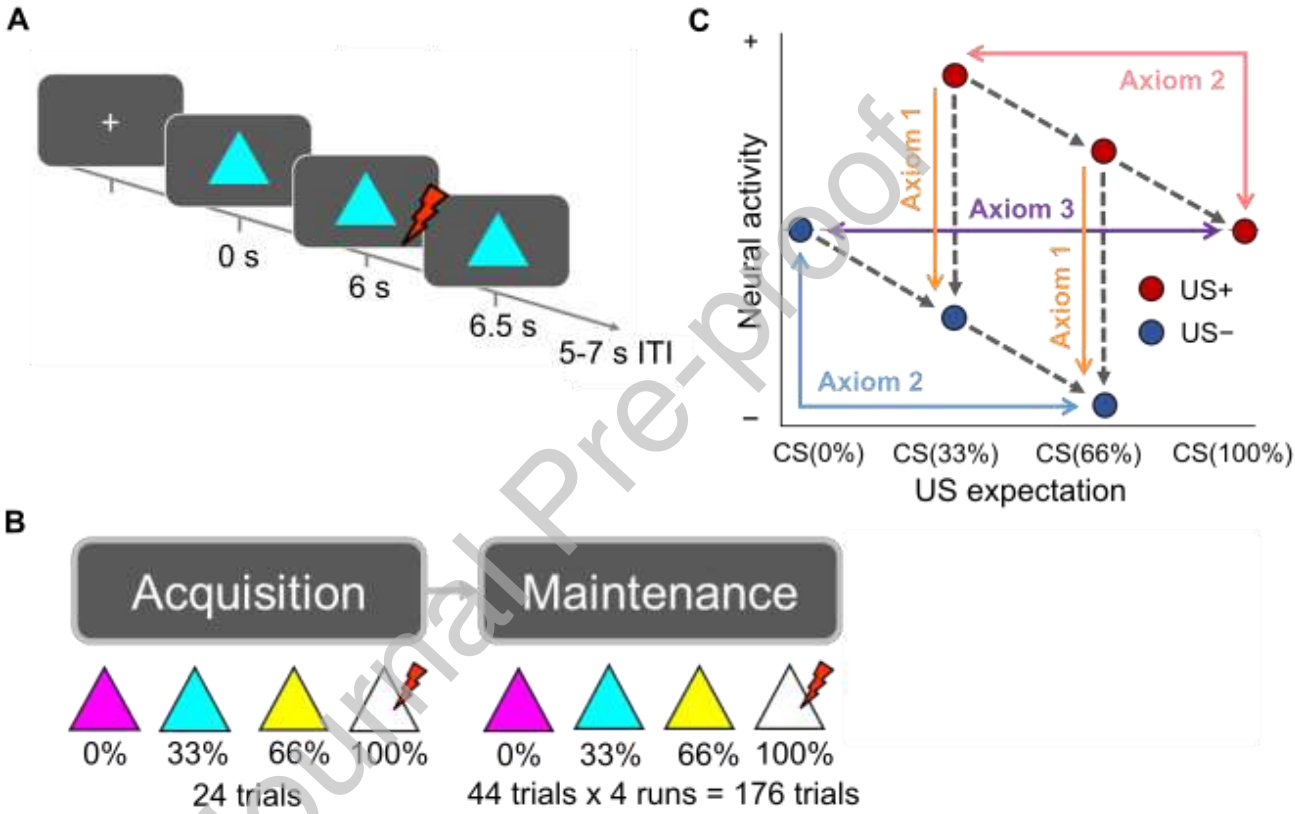

Figure 1. A, Experimental design. A classical delay threat conditioning paradigm was used with colored shapes as conditioned stimuli (CSs), presented for 6.5 s. The CSs predicted an aversive electric shock (US) with different rates (0%, 33%, 66%, 100%). If the US occurred (US+ trials), it started 6 s into CS presentation and lasted 0.5 s, co-terminating with the CS. The inter-trial interval was 5-7 s long. **B**, Experimental phases. In the acquisition phase, each CS (triangle) was presented 6 times in a row to facilitate learning. In the maintenance phase, each of these CSs was presented 44 times over four blocks in intermixed order. All reported analyses pertain to the maintenance phase, which could be analyzed without commitment to a particular learning model. **C**, The necessary and sufficient conditions for full signed PEs. Comparisons of conditions are theoretically possible in both directions (i.e., the positive and negative signs on the y-axis are arbitrary) but based on previous work we a priori expected higher neural activity for higher positive PE (positive values after US+, that is, US occurrence) and lower neural activity for lower negative PE (negative values after US−, that is, US omission). Grey dashed lines depict the tested contrasts, which were tested either all in direction of the arrows, or all into the opposite direction. Using the a priori expected direction of comparisons, axiom 1 states that shock outcomes are associated with higher activity than no shock

outcomes. Axiom 2 states that the more unexpected the outcome is, the higher the related BOLD activity regardless of outcome type (US+ or US−). Axiom 3 always states that activity is the same for fully expected outcomes regardless of outcome type.

Table 1. Reaction time and accuracy statistics for the fMRI experiment.

	CS(0%)	CS(33%)	CS(66%)	CS(100%)
Reaction time (Mean ± SD), ms	1046±212	1044±268	1086±269	1011±248
Accuracy (Mean ± SD), % correct	99.2±2.7	99.2±2.1	98.9±2.4	99.2±2.8
One-way repeated-measures ANOVA	*F*	*df*	*p*	
Reaction time ~ CS type	0.081	3, 76	0.97	
Accuracy ~ CS type	0.142	3, 76	0.935	

Reaction time and accuracy data from trials with reaction times shorter than 200 ms (0.2% of all trials over all participants) were excluded. Trials with incorrect or missed responses were excluded from reaction time analyses. Repeated-measures ANOVA was conducted with the 'aov' function in R.

Experimental phases. During the acquisition phase, participants were presented with 4 blocks of 6 consecutive trials of the same CS, in order to facilitate learning of the CS-US contingencies (24 trials in total). CS were triangles with different colors (RGB: 255, 0, 255; 0, 255, 255; 255, 255, 0; 255 255 255). CS-US pairings were balanced over these 6 trials per CS such that there were exactly 0, 2, 4, or 6 reinforced trials, respectively, for the four CS. Order of the blocks, and of the trials within blocks, was randomly determined for each participant. In the following maintenance phase, participants were presented with 176 trials (44 trials per CS) of the same CSs, now in pseudo-random intermixed order, reinforced randomly at constant rate per CS and divided into four blocks. The motivation for the blocked order in the acquisition phase was to facilitate learning, and the intermixed order for the maintenance phase was chosen for optimal elicitation of prediction errors after at least some learning had already taken place. The experiment was presented using Cogent 2000 (version 1.32, vislab.ucl.ac.uk) on Matlab. The visual presentation was projected onto a 42 cm x 33 cm size screen (1024 x 768 pixel resolution) at approximately 73 cm distance from the participants' eyes.

Delivery of the unconditioned stimuli. US was delivered with a constant current stimulator (Digitimer DS7A, Digitimer, Welvyn Garden City, UK) through a pin-cathode/ring-anode configuration on the right forearm. US intensity was individually calibrated for each participant (fMRI: outside the scanner) before the experiment. First, a clearly unpleasant intensity was determined with an ascending staircase procedure. After that, participants gave subjective ratings (0 = felt nothing to 100 = very unpleasant) for 14 random intensities below the initial threshold. The intensity corresponding to a rating of 85, acquired with linear interpolation, was chosen as the US intensity for the experiment (3.3±0.8 mA, range 1.5–5.5).

Data acquisition and statistical analyses

Subjective rating of US expectation. Participants rated their US expectation after each CS, reflecting their explicit knowledge of the CS-US contingencies after the maintenance phase, using a computerized visual analogue scale anchored with "0%" and "100%". The initial position of the slider was set to the middle of the scale. The US expectancy ratings were analyzed with a one-way repeated-measures ANOVA with the 'aov' function in R (version 3.6.1) (RCoreTeam, 2019) with RStudio (version 1.2.1335) (RStudioTeam, 2018), including CS type as a factor with four levels. Partial eta squared were computed with the 'etasq' function of R package heplots (version 1.3-5.) (Fox et al., 2018). Moreover, we computed pairwise one-sided paired t-tests for CS(100%) > CS(66%), CS(66%) > CS(33%), and CS(33%) > CS(0%) with Holm-Bonferroni multiple comparisons correction over the three comparisons. We did not exclude participants that did not show monotonic learning of the subjective ratings from further analyses, in line with standard practice in the study of human threat conditioning, as explicit knowledge of CS-US contingencies does not necessarily reflect the same learning mechanism as autonomic learning (Ojala and Bach, 2020).

Pupil size recording and analysis. Due to technical limitations of the particular scanner environment used to collect the fMRI data, no psychophysiological learning indices were available for the fMRI experiment. To ensure learning in this paradigm, we conducted a separate experiment beforehand (N = 19, 164 trials with 24 trials of acquisition and 140 trials of maintenance) with the same design, on an independent sample outside the MRI scanner. Gaze direction and pupil area were recorded with an EyeLink 1000 system (SR Research, Ottawa, ON, Canada) from both eyes of each participant at 500 Hz. For each participant, we used the eye with fewer missing data for analysis. The size of the visual presentation was 32 cm x 23 cm (1280 x 1024 pixel resolution). The center of the screen was at approximately 70 cm distance from the participants' eyes and the eye-tracking camera was at approximately the same distance. Calibration of gaze direction was done on a 3-by-3-point grid in the EyeLink software. EyeLink data files were converted and imported into the Psychophysiological Modelling (PsPM) toolbox (version 4.0.1, bachlab.github.io/PsPM/) in MATLAB2018a for further preprocessing and analysis. Blink and saccade periods were detected by the EyeLink online parsing algorithm and excluded from pupil data during import into PsPM. Data points for which gaze direction deviated more than 5° visual angle from the center of the screen were excluded (Korn et al., 2017; Korn and Bach, 2016). Raw pupil size data was filtered with a unidirectional first order Butterworth low pass filter with 25 Hz cut off frequency and downsampled to 50 Hz. Missing data were linearly interpolated for further analysis. One participant was excluded from further pupil size analysis based on a criterion of having more than 75% trials with more than 75% missing data points during 11 seconds following CS onset due to invalid fixations, saccades or blinks.

Pupil size has been suggested to relate to US prediction (Tzovara et al., 2018), but it is unclear how this relation evolves during CS presentation. A previous psychophysiological model for analysis of threat-

conditioned pupil size responses had been optimized for discriminative (one CS+ vs. one CS−) threat conditioning (Korn et al., 2017). This is why we here took a data-driven approach to analyze the relation between pupil size and US probability, using a cluster-level random permutation test (Maris and Oostenveld, 2007). This analysis was performed in R (version 3.5.2) (RCoreTeam, 2019) and RStudio (version 1.0.136) (RStudioTeam, 2018). First, we tested for a linear relation between CS type and pupil size by conducting a linear regression for every time point (in 0.1 s bins) during CS presentation until US onset, 6 s after CS onset, averaged over trials in the maintenance phase. The resulting coefficient and *p*-values were compared against values derived from 1000 regressions with randomly shuffled trial labels in a permutation test, under the null hypothesis that trial labels are exchangeable. To account for multiple comparison across time, we applied cluster-level correction for family-wise error (Maris and Oostenveld, 2007; Sassenhagen and Draschkow, 2019). This test controls the false positive rate for the statement that there is any effect somewhere within the correction window, and thus makes no a priori assumption about the location of an effect. Importantly, for this test, the temporal cluster extents are only descriptive and not controlled for the error rate. Next, we conducted post-hoc t-tests with permutation to investigate differences between the four CS conditions over the interval between CS and US onset.

fMRI data acquisition and preprocessing. Data were acquired using a 3 T Prisma MRI scanner (Siemens, Erlangen, Germany) with a 64-channel head coil. T_2^*-weighted multi-echo echo-planar images (EPI) were acquired using a custom-made 2D EPI sequence (Lutti et al., 2013). The in-plane resolution was 3 mm isotropic and the size of the acquisition matrix was 64 x 64 (FOV 192 mm). 40 axial slices were acquired in ascending order, with a nominal thickness of 2.5 mm and inter-slice gap of 0.5 mm (effective thickness 3 mm). The volume TR was 3.2 s and the flip angle 90°. Parallel imaging was used with an acceleration factor of 2 along the phase-encoding direction and images were reconstructed using GRAPPA (Griswold et al., 2002). In order to avoid signal dropouts in the EPI images and achieve maximal BOLD sensitivity in all brain areas, a multi-echo EPI acquisition was used (Poser et al., 2006) with the following echo times: TE = 17.4/35/53 ms . There were 6 fMRI runs in the experiment, with 24 trials in the first run, which are not analyzed here, and 44 trials in each of runs 2–5, summing up to a total of 200 trials (176 in the analyzed runs). The last run was another 24 trials of another acquisition that are also not presented here. Phase and magnitude B0 field maps were acquired at the beginning of the experiment (TE 10 and 12.46 ms, TR 1020 ms, FOV 192 mm, 64 transversal slices of 2 mm thickness). A high-resolution structural scan was obtained at the end of the scan session (MP-RAGE; TR 2000 ms, TE 2.39 ms, inversion time 920 ms, 1 x 1 x 1 mm voxel size, flip angle 9°, FOV 256 mm, 176 sagittal slices). During fMRI, we collected respiratory and cardiac data to correct for physiological noise in the fMRI analysis, using the scanner's in-built breathing belt and a strapped photoplethysmograph on the left index finger. Data were recorded with a PPG100C MRI amplifier and a BIOPAC MP150 system.

We used SPM12b (Wellcome Trust Centre for Neuroimaging, London) and MATLAB2016a (Mathworks, Sherborn, MA, USA) to preprocess and analyze fMRI data. Preprocessing of the structural imaging data included field inhomogeneity correction and segmentation. Preprocessing of the functional images started with the combination, for each volume, of the EPI images acquired at different echo times using a simple summation. Because the first echo has very good sensitivity for high-dropout regions and the two others give better sensitivity for other regions, this process leads to maximal BOLD sensitivity to all brain areas (Poser et al., 2006). This was followed by correction of image distortions using the SPM FieldMap toolbox (Hutton et al., 2002) and the B0 field map data, slice-time correction, motion correction (realignment), as well as co-registration with the T_1-weighted structural images, spatial normalization to the Montreal Neurological Institute (MNI) template, and spatial smoothing with an 8 x 8 x 8 mm FWHM Gaussian filter. Serial autocorrelations were estimated using SPM 12's FAST model (Corbin et al., 2018). Cardiac and respiratory signals were used for physiological noise correction with the RETROICOR method (Glover et al., 2000) as implemented in the PhysIO toolbox for SPM (Kasper et al., 2017). In total, 18 physiological noise regressors (cardiac: 3 orders, respiratory: 4 orders, interaction: 1 order) and 6 head motion regressors from the realignment were used as nuisance parameters in the analyses. The third run of one participant was excluded from the fMRI analyses due to head motion in the beginning of the run leading to a severe artefact affecting all volumes within the run.

In all analyses, we performed standard random effects analyses at the group level. First-level contrast images from each participant were entered into one-sample *t*-tests against zero and statistical parametric maps were created with cluster-level family-wise error (FWE) correction at $p < 0.05$ with initial cluster-forming threshold $p < 0.001$ (Eklund et al., 2016). For illustration, functional results were overlaid on a normalized mean anatomical (grey and white matter only) image of our sample of participants. Anatomical location of clusters was defined based on the Neuromorphometrics labels in SPM12. Importantly, there is no anatomical specificity for activity within any of the clusters due to the cluster-level correction. The anatomical labels are included to give the reader an approximation of the location of the entire cluster.

Mass univariate whole-brain analysis of PE signals. The first level GLMs for each participant modelled cue (CS) and outcome (US) time points as stick functions and included serially orthogonalized parametric modulators of these events as well as nuisance regressors. The CS-US interval of 6 seconds was chosen to reduce design matrix collinearity: the correlation of them modelled hemodynamic responses to CS and US event was Pearson's $r = -0.06$. As parametric modulators, we included expectation of the US outcome for CS time point, and US outcome (delivered/omitted) as well as PE for US time point. US expectation was formalized in the primary analysis as the overall US rate (0%, 33%, 66%, or 100%) for the CS presented on that trial (primary analysis), and in a supporting analysis as the prior expectation of the US+ probability from a normative Bayesian learning model, which in a previous study provided the best description of trial-by-trial

conditioned skin conductance and pupil size responses across several samples (Tzovara et al., 2018). We did not base our PE on declarative knowledge of the CS-US contingencies as there is evidence that it represents a learning process distinct from that reflected in autonomic indices, or in non-human animal behavior (Lovibond and Shanks, 2002; Ojala and Bach, 2020).

Notably, US expectation from these two approaches is almost identical during the maintenance phase. The US outcome was defined as either 1 (US+) or 0 (US−). For primary and exploratory follow-up analyses, we constructed four separate GLMs: Parametric GLM 1: full signed PE (outcome−expectation for both US+ and US− trials, primary analysis). Parametric GLM 2: difference between negative and positive PE for all trials (+(outcome−expectation) for US+ trials, −(outcome−expectation) for US- trials). Parametric GLM 2 can also be interpreted as a test for unsigned prediction errors (|outcome−expectation| for all trials). Parametric GLM 3: positive PE (outcome−expectation for US+ trials only). Parametric GLM 4: negative PE (outcome−expectation for US− trials only). These four different PEs were calculated with both definitions of expectation. As we used a parametric modulator for PE across all trials, this was fixed to the mean of all other trials for the US− trials in parametric GLM 3 and for the US+ trials in parametric GLM 4. For clarity, all analyses included the zero-PE conditions (0% and 100% reinforcement) for comparison.

For each contrast, we examined correlated BOLD activity with a one-tailed one-sample *t*-test against zero. Our a priori expectation was that larger positive PEs (positive values after US+) would relate to higher BOLD signal and larger negative PEs (negative values after US−) to lower BOLD signal, based on previous work (Roy et al., 2014).

Next, we conducted follow-up analyses of the averaged signal from significant clusters and a-priori anatomical regions (see section on region-of-interest analysis), as well as a follow-up whole-brain analysis, to determine whether BOLD signal in any detected cluster, or in any voxel, would fulfill the necessary and sufficient conditions for representing PEs (Fig. 1C) (Caplin and Dean, 2008). To this end, we computed an additional "categorical" GLM agnostic to the parametric values of PE, where we modelled the 4 different CS, and the 6 different US types (one for each possible CS-US pairing), in separate conditions. For the voxel-wise whole-brain analysis, we conducted a conjunction null test (logical “AND”) on the significance of all relevant condition contrasts in both directions for the outcome and expectancy conditions (Fig. 1C, axiom 1 and 2). We defined conjunctions separately for the full PE model (all 6 possible contrasts), PE difference/unsigned PE model (both US+ and US− trials but flipped for US−: CS(0%) < CS(33%) < CS(66%)), positive PE (US+ trials only), and negative PE (US− trials only). We did not explicitly test for the condition that fully expected outcomes should elicit similar BOLD activity (Fig. 1C, axiom 3). This would have required a test of equivalence, which was not necessary since the other axioms were already found to be not supported by the data.

Mass univariate region-of-interest analysis for PEs. We next analyzed whether BOLD signal in the significant cluster from our primary analysis, and in different anatomical regions-of-interest (ROI), fulfilled necessary and sufficient criteria to represent PEs. Anatomical masks for thalamus, anterior and posterior insula, and anterior cingulate cortex were created from the WFU PickAtlas AAL library (Maldjian et al., 2003; Tzourio-Mazoyer et al., 2002). Frontal cortex ROI masks were created separately for Brodmann Areas 8–11 and 44–47 (dilation level 1 in 2D). For amygdala, we binarized probabilistic masks from Abivardi and Bach (2017) (combined basolateral and centrocortical divisions) which are based on manual segmentation of N = 50 datasets from the Human Connectome Project (Van Essen et al., 2012). The binarization threshold was set at 0.5 to obtain mask volumes (mm^3, in final normalized functional space) within 1 SD of the mean native space volumes reported in Abivardi and Bach (2017). For periaqueductal grey (PAG), we used the high-resolution probabilistic anatomical mask for young people (linear option) from the ATAG atlas (Keuken et al., 2017). The probabilistic PAG mask was binarized at a threshold of 0.13, which best retained the anatomical shape of the PAG when inspected qualitatively with respect to a normalized mean image of the participants' anatomical scans. We used high-resolution anatomical masks from the recent Reinforcement Learning Atlas (Pauli et al., 2018) for ventral striatum (nucleus accumbens), dorsal striatum (caudate nucleus and putamen), and dopaminergic midbrain (substantia nigra pars reticulata/compacta and ventral tegmental area). The anatomical ROIs were defined in the MNI space, co-registered to the functional space, and used in the analyses at the group level. Moreover, to explore the results from the parametric GLMs, we extracted parameter estimates from clusters with significant activity associated with each different type of PE (cluster-level corrected FWE $p < 0.05$ with $p < 0.001$ initial threshold, see Table 3 for the clusters and their statistics).

For each anatomical ROI and significant functional cluster, we extracted the average BOLD amplitude estimates from the categorical GLM for the six US outcome conditions in the maintenance trials. For the a priori anatomical ROIs, we investigated whether the average BOLD signals fulfilled the axioms by conducting paired Bayesian t-tests in JASP (version 0.16.3, JASP Team, 2022) for four reduced comparisons: Axiom 1) US+ > US− over US expectation conditions CS(33%) and CS(66%), Axiom 2) different levels of US+ expectation: CS(0%) > CS(66%) for US−, and CS(33%) > CS(100%) for US+ trials, and Axiom 3) for CS(100%) = CS(0%). For those ROIs that were included in at least two out of three previous axiomatic studies (Fazeli and Büchel, 2018; Geuter et al., 2017; Roy et al., 2014; ACC, amygdala, thalamus and PAG), we used an informed prior defined as a normal distribution with mean and standard deviation set as the mean and standard deviation over reported values in the previous studies for each ROI and comparison (see Supplementary Table 2 for the values of individual studies). For the other ROIs, we set the prior as the default of JASP, which is a Cauchy distribution with scale 0.707. We also computed frequentist paired Cohen's *d* effect sizes ('cohensD' function of lsr package in R) (Navarro, 2015) for the full axiomatic comparisons: Axiom 1): US+ > US− for US expectation conditions CS(33%) and CS(66%), (2) Axiom 2): different levels of US+ expectation: CS(0%) > CS(33%) and CS(33%) > CS(66%) for US−, and CS(33%) > CS(66%) and CS(66%) > CS(100%) for US+

trials, and Axiom 3) CS(100%) > CS(0%) (see Fig. 1C; 7 effect size computations in total). Moreover, we created linear mixed effects models ('lme' function in the nlme package in R) (Pinheiro et al., 2020) on the BOLD amplitude estimates for (1) full signed PEs, (2) positive PEs, (3) negative PEs, (4) PE difference/unsigned PEs, (5) US+/US− outcome, and (6) null model. Each model included PE or outcome values as the fixed effect. To account for potential asymmetry between positive and negative PEs, we also included a full PE model with separate fixed effects for positive and negative PEs, allowing different intercepts and slopes. The null model only contained a constant value 1 as the intercept. Each model included a participant intercept as a random factor, allowing for a different intercept but not slope for each participant (1|Participants). All models were estimated using the maximum likelihood (ML) method to allow extraction of model evidence metrics. To formally compare the different models, we computed Bayes factors with Bayesian Information Criterion approximation for frequentist linear regression models with R package bayestestR (Makowski et al., 2019; Wagenmakers, 2007). For the functional clusters, we conducted Bayesian t-tests and post-hoc effect size computations for the axioms with Cohen's *d* for paired observations similarly to the tests for the anatomical ROIs (Fig. 1C).

Code and data availability. The code for the experiment, data analysis and figures are available in a public repository gitlab.com/kojala/threatlearning_fmri. Group-level unthresholded Statistical Parametric Maps, ROI masks and mean beta values relevant to the analyses are available in a public repository with DOI 10.5281/zenodo.6983543. Data from the behavioral experiment outside the scanner are available in a public repository with DOI 10.5281/zenodo.3872055. Due to data protection regulations and given a risk of statistical identification of brain anatomy, individual-level MRI data are available from the authors for scientific purposes under a data protection agreement.

Results

Declarative knowledge of CS-US contingencies

Participants reported explicit declarative knowledge of the CS-US contingencies by rating their US expectation for each CS after the maintenance phase of the fMRI experiment (200 trials, Fig. 1B, 2A). There was a significant linear effect of CS type on US expectation ratings, and pairwise differences for CS(100%) > CS(66%), CS(66%) > CS(33%), and for CS(33%) > CS(0%) (Table 2). Results were similar in the behavioral experiment outside the scanner (164 trials, Table 2).

Table 2. Statistics for ratings of US expectation.

Ratings for fMRI experiment (N = 21, after 200 trials)				
	CS(0%)	CS(33%)	CS(66%)	CS(100%)
Mean ± SD	14.8 ± 24.1	44.3 ± 17.7	55.4 ± 19.3	78.6 ± 31.7
Repeated-measures ANOVA	F	df	p	η^2_p
Subjective rating ~ CS type	25.99	3, 80	$7.78e^{-12}$	0.49
Linear contrast	75.88	1, 80	$3.25e^{-13}$	
Paired t-test, one-sided	T	df	p	$\|d\|$
CS(100%) > CS(66%)	4.06	20	0.0003*	0.44
CS(66%) > CS(33%)	2.02	20	0.028*	0.22
CS(33%) > CS(0%)	6.09	20	0.00003*	0.66
Ratings for behavioral outside-scanner experiment (N = 18, after 164 trials)				
	CS(0%)	CS(33%)	CS(66%)	CS(100%)
Mean ± SD	7.6 ± 13.1	40.7 ± 25.4	67.5 ± 22.5	85.6 ± 26.5
Repeated-measures ANOVA	F	df	p	η^2_p
Subjective rating ~ CS type	44.03	3, 72	$2.84e^{-16}$	0.65
Linear contrast	129.07	1, 72	$2.00e^{-16}$	
Paired t-test, one-sided	T	df	p	$\|d\|$
CS(100%) > CS(66%)	2.30	17	0.0167*	0.26
CS(66%) > CS(33%)	4.16	17	0.0003*	0.48
CS(33%) > CS(0%)	4.67	17	0.00009*	0.54

For paired t-tests, Holm-Bonferroni correction was applied over the three comparisons within each experiment. * $p < 0.05$ with corrected α-level.

Pupil size responses

To ensure implicit learning in this paradigm, we analyzed pupil data from a behavioral experiment outside the scanner. We were interested in how US expectation, while seeing one of four CSs with different US rates, was reflected in pupil size. Across the entire experiment, we found a significant linear effect ($p < .05$) of US expectation (Fig. 2B) with greater pupil dilation for higher US expectation between about 1-6 s after CS onset. Post-hoc pairwise comparisons further showed that the response to CS(100%) was larger than for CS(66%) from around 4-6 s after CS onset, CS(66%) was most of the time more pronounced than for CS(33%) between about 0.5-6 s after CS onset, and greater for CS(33%) than for CS(0%) around 4-5 s after CS onset (Fig. 2B).

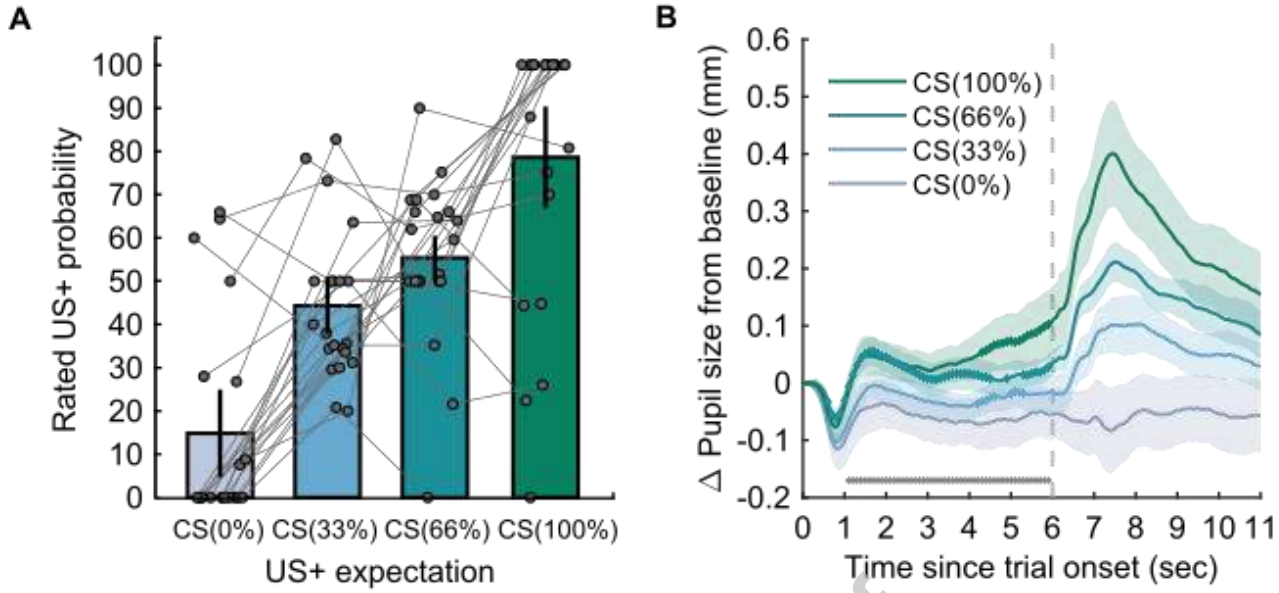

Figure 2. Ratings of US expectation, and threat-conditioned pupil size responses, for each CS. **A**, Subjective US expectancy ratings after the maintenance phase of the experiment in the fMRI sample only. The plot shows mean and standard errors of the mean as well as individual ratings (connected lines refer to individual participants). **B**, Average pupil size change from baseline in the outside-scanner sample, over trial time during maintenance phase. Shaded areas depict the standard error of the mean. Grey horizontal markers below the time courses show the significant effect of CS type on pupil size, based on a cluster-based correction for multiple comparison across the entire CS-US interval. Markers on CS time courses show the significant clusters for the comparison of each CS type in relation to the previous one (CS(100%) > CS(66%), CS(66%) > CS(33%), CS(33%) > CS(0%)). There was one significant cluster in the last third of the CS period right before the us for CS(100%) > CS(66%), a significant cluster covering most of the CS-US interval and two smaller later clusters for CS(66%) > CS(33%), and two significant clusters at around 4-5 seconds after CS onset for CS(33%) > CS(0%). Location of the clusters is shown for illustration only and is not part of the statistical test.

Neural representation of PEs: whole-brain analysis

As a positive control, we observed an effect of US type (US+ > US−) on BOLD fMRI activity in the bilateral anterior and posterior insula, bilateral temporal, parietal and central operculum, right supramarginal gyrus, right superior temporal gyrus and left transverse temporal gyrus (voxel-wise FWE $p < .05$).

In our primary analysis (parametric GLM 1), we investigated the relation between BOLD signal at the US time point, and PE across all trials during the maintenance phase of the experiment, using a GLM that included separate parametric modulators for US presence/absence and for PE. We found that BOLD responses were correlated with full signed PEs in two clusters approximately in the bilateral superior medial prefrontal cortex, and right middle-superior occipital gyrus and superior parietal lobule ($p < .05$ cluster-level FWE; Fig. 3A, Table 3). That is, more unexpected US+ outcomes were associated with higher BOLD activity, and more unexpected US− outcomes, i.e., omission of US, were associated with lower BOLD activity in these clusters (in accordance with Fig. 1C).

However, examination of BOLD amplitude estimates in individual conditions in these clusters suggested that this effect was driven by the influence of negative PEs, whereas for positive PEs, condition averages did not show a linear relation between US+ expectation and BOLD signals (Fig. 3A). To allow for a possibility that the brain represents positive and negative PEs in partly different regions, we tested for a difference between negative and positive PEs, and then analyzed them individually. We found a cluster in which slope of a BOLD activity relation with negative PEs was steeper (more negative) than for positive PE, located approximately around left superior frontal and bilateral medial frontal regions (Fig. 3C), and partly overlapping with the ventromedial part of the negative PE frontal cluster but not with the dorsomedial full signed PE cluster (Fig. 3C,D, Table 3). An alternative interpretation for this cluster is a negative correlation between unsigned PEs and BOLD activity in this region. However, investigation of the extracted parameter estimates from the categorical GLM was in favor of the former interpretation: the slope of BOLD activity relation with positive PEs was flat and not positive, as would be expected for an unsigned PE representation.

In keeping with this, more unexpected US− outcomes were associated with lower BOLD activity in clusters approximately located around bilateral superior frontal gyrus, left angular gyrus and left posterior cingulate gyrus, partly overlapping with the smaller frontal cluster of the full PE model (Fig. 3B,D). Extracted condition averages from our categorical GLM showed a linear gradient of negative PEs, as expected. On the other hand, we found no evidence of BOLD activity association with positive PEs.

In these PE models, we used the overall US rate to compute PEs, but participants would not have perfectly learned these at the start of the maintenance phase. To ensure this did not obscure representation of PEs, we computed PEs with a normative (statistically optimal) learning model. We found very similar results to the full signed PE model, that is, larger PEs were associated with increased BOLD activity in a cluster approximately located around left medial superior frontal gyrus (cluster-level FWE-corrected p = 0.014, cluster size 366 voxels).

Table 3. PE related BOLD activity during maintenance of threat associations.

Regressor	Approximate cluster anatomical region	Cluster size	Cluster *p*
Full signed PE (parametric GLM 1)	1. Superior frontal gyrus medial L, Superior frontal gyrus R	356	0.014
	2. Middle & superior occipital gyrus R, Superior parietal lobule R	266	0.044
Difference positive vs. negative PE (parametric GLM 2) *	1. Superior frontal gyrus L	404	0.007
	2. Subcallosal area L, Superior frontal gyrus medial L, Medial frontal cortex R	1,636	$1.19e^{-07}$
Positive PE (parametric GLM 3)	No significant clusters	–	–
Negative PE (parametric GLM 4)	1. Superior frontal gyrus L, R	3,001	4.23^{-11}
	2. Angular gyrus L	418	0.008
	3. Posterior cingulate gyrus L	350	0.016

MNI, Montreal Neurological Institute. Statistical parametric maps were cluster-corrected at FWE $p < 0.05$, with initial threshold of $p < 0.001$ uncorrected. Cluster p: corrected p-value. For full signed and positive PE models, the reported contrasts reflect higher BOLD activity related to larger PE (positive for US+, and larger for less expected US+) and lower BOLD activity for larger negative PE. For the difference, the reported contrast represents higher BOLD activity for larger negative PE than positive PE. This contrast would also reflect larger BOLD activity for smaller unsigned PE (see Fig. 1C). Opposite directions were tested for all models but there were no further significant findings. Anatomical labels (Neuromorphometrics, SPM12) are reported for each cluster for approximate localization. Peak statistics are not included as they are not relevant for our inferences and are by their nature biased (Davenport and Nichols, 2020).

Neural representation of PEs: region-of-interest analysis

Whole-brain search may provide limited statistical power if full signed PE representations occurred in small regions. Hence, we investigated PE representations in a priori defined anatomical regions of interest. We used a formal Bayesian model selection approach to avoid multiple null hypothesis tests. Distinct from some of our previous analysis, this approach seeks to simultaneously explain responses to US occurrence and US omission. Our analysis revealed that the symmetric full PE model was the best model (log BF > 3) for BA 9 and ACC. The outcome-only (US+ vs. US−) model best explained the data (log BF > 3) for BA 44, BA 47, anterior insula and posterior insula (Fig. 4). There was no decisive evidence in any of the other regions.

We applied the same analysis to the significant clusters from our whole-brain analysis, to facilitate interpretation (Fig. 5). The full signed PE cluster in superior frontal gyrus was best explained by a model including negative PE only (i.e., no expression of positive PE), and the full signed PE cluster in occipital and parietal areas was best explained by an asymmetric full PE model, which implies an encoding of positive PE but with different slope than negative PEs. One PE difference cluster was best explained by a negative PE model, and the other by an unsigned PE model (that is, opposite representation of negative and positive PE but with the same slope).

In a supplementary analysis, we found that the model-free BOLD timecourses show that the US outcome response does not always align well with the canonical haemodynamic response function for all ROIs and/or conditions (Supplementary Figures 1-4).

Necessary and sufficient conditions for full signed PE model

We next evaluated whether BOLD responses in any brain region fulfill three criteria, or 'axioms' (Fig. 1C), to represent PE signals in a learning-theoretic sense. In a whole-brain analysis, there were no significant clusters fulfilling the conjunction of axioms 1 (i.e., higher activity for US+ than US− outcome) and 2 (i.e., higher activity for more unexpected US+ outcomes and for more expected US− outcomes). Axiom 1 was fulfilled in four large clusters approximately in the left central operculum/posterior insula, right parietal operculum/superior frontal gyrus, and bilateral middle cingulate gyrus/left superior frontal gyrus, and right cuneus (Supplementary Table 1). However, axiom 2 was not fulfilled in any region at the whole-brain level. In an exploratory analysis, we also verified that axiom 2 was not fulfilled in the other direction (i.e., higher activity for more unexpected US− outcomes and for more expected US+ outcomes).

For region-of-interest analysis, we extracted parameter estimates, conducted Bayesian t-tests, and also calculated corresponding frequentist effect sizes for axiomatic comparisons. We report here the results on regions that showed significance or decisive model evidence in favor of full signed prediction errors in our previous analyses; full results are found in Figure 6, Table 4 and Supplementary Table 3. In the first significant full signed PE cluster from our whole-brain search, as well as in anatomical BA 9 and in anatomical ACC, there was no conclusive evidence (|log Bayes Factor| < 3) when looking at the difference between CS(33%) and CS(100%) when US occurred or between CS(0%) and CS(100%); thus signed axiom 2 (expectation effect) and axiom 3 (equivalence of fully expected outcomes, Fig. 1C) could not be accepted or rejected based on this data. . Moreover, BA 9 and ACC also did not have conclusive evidence for US omission expectation effect. The second significant full signed PE cluster from our whole-brain search showed moderate evidence for a difference between CS(66%) and CS(33%) at US occurrence (log BF = 3.71), and somewhat did not fulfill axiom 3 (log BF = −2.93) or axiom 2 for US omission (log BF = 1.45). Overall, no region had at least moderate Bayesian evidence (Table 4) or small-to-medium frequentist effect sizes (d > 0.20) for all axiomatic tests (Supplementary Table 3).

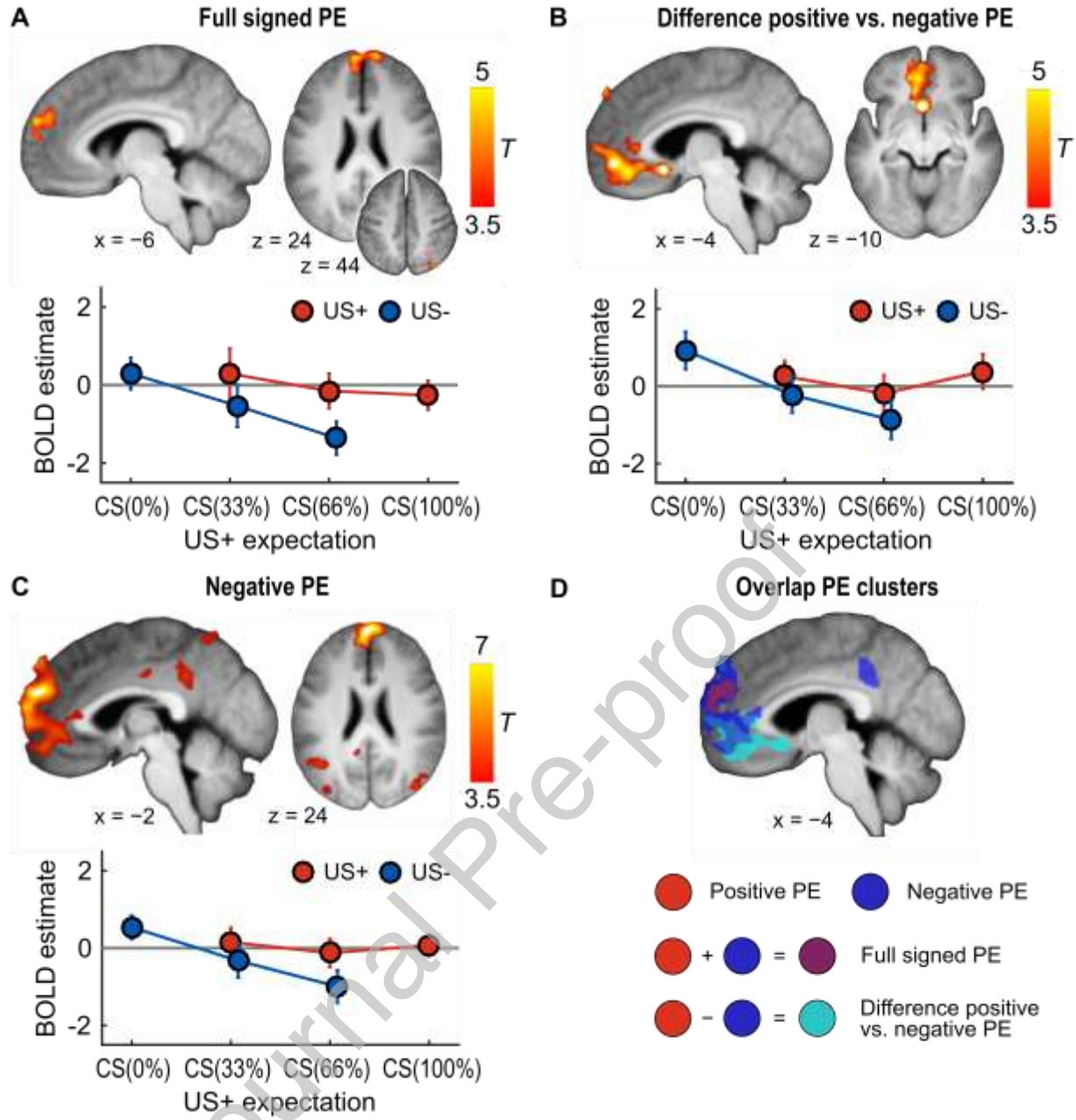

Figure 3. Whole-brain PE fMRI results. **A**, Full signed PEs correlated with BOLD activity in the dorsomedial prefrontal cortex (dmPFC) and superior parieto-occipital cortex. Average responses for each condition from the frontal cluster show a clear linear relationship with US expectation only for US− conditions (parametric GLM 1). However, this cluster was entirely overlapped by the negative PE (US omission) cluster (panel D) and was equally well explained by negative PE as full signed PE models (Fig 5.). **B**, Interaction of PE with outcome (US) type in BOLD activity in vmPFC and rostral anterior cingulate cortex (rACC), indicating a steeper (negative) BOLD relation for negative (US omission) than positive (US occurrence) PE, or generally a representation of less expected outcomes in lower BOLD signal (parametric GLM 2). **C**, Negative PEs (US omission) correlated with BOLD activity in the dmPFC and ventromedial PFC (vmPFC), angular gyrus and posterior cingulate cortex (PCC) (parametric GLM 3). **A**-**C**, BOLD amplitude estimates are shown as mean and standard error of the mean. Statistical parametric maps were thresholded at $p < 0.05$ cluster-level FWE with initial threshold $p < 0.001$. Unthresholded SPMs are available online. BOLD estimates are shown for the cluster with the lowest corrected p-value for each PE model. **D**, Significant PE clusters and their overlap. The negative PE (US omission) PFC cluster almost entirely overlaps with or encompasses the signed PE PFC

cluster, whereas the PE interaction cluster extends also beyond the negative PE cluster. There were no significant positive PE clusters.

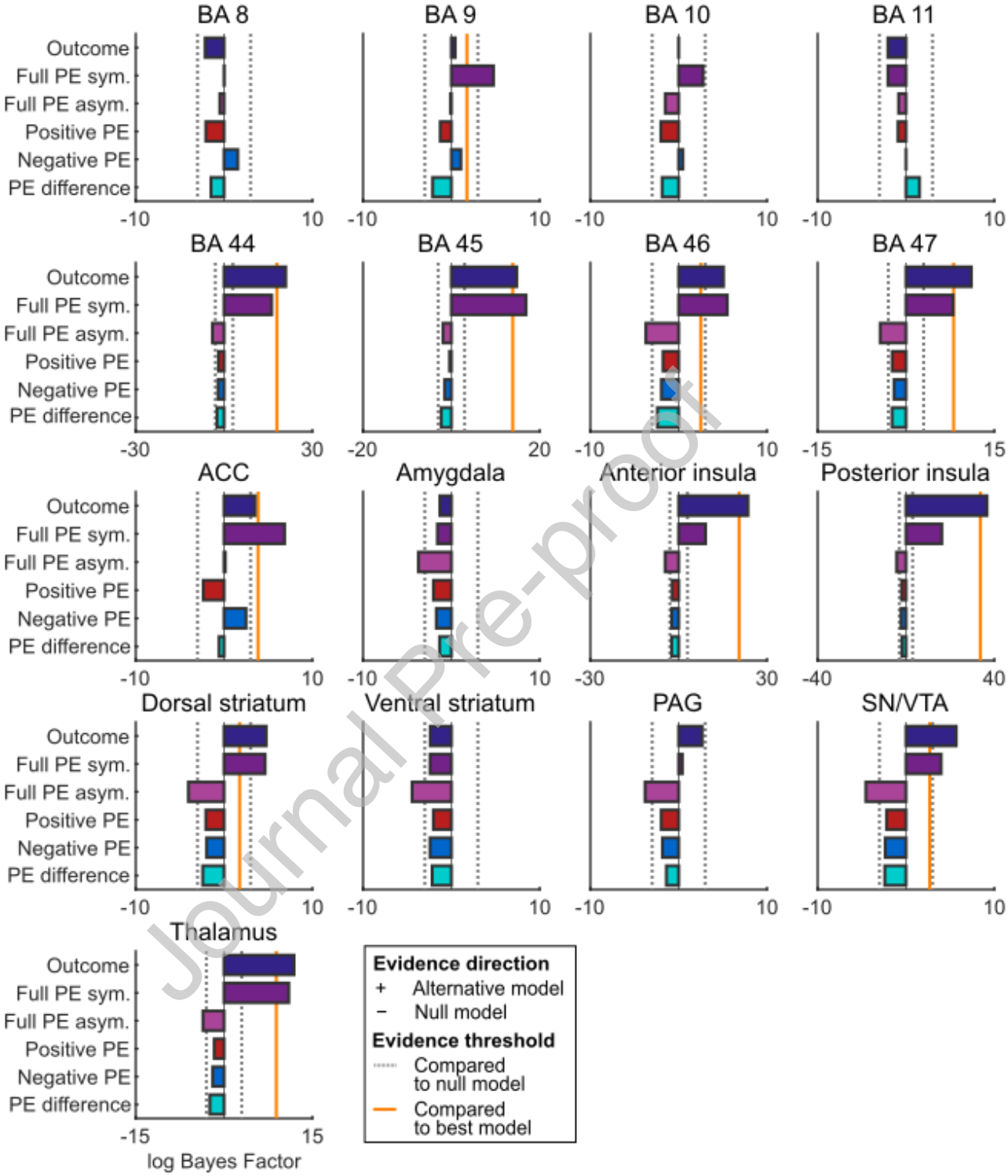

Figure 4. Model comparison of PE and outcome-only models for BOLD signals from each anatomical region-of-interest. Log Bayes Factors (BF) > 3 (dotted grey line) indicate moderate support for a model over the null model, whereas log BF < −3 denote moderate evidence for the null model, with values in between representing inconclusive evidence for any model. The orange line marks the evidence threshold (log BF 3) for moderate difference between the best model and other models. Full PE sym. = one intercept and slope parameter for both positive (US occurrence) and negative (US omission) PE; Full PE asym. = separate intercepts and slopes for positive and negative PE.

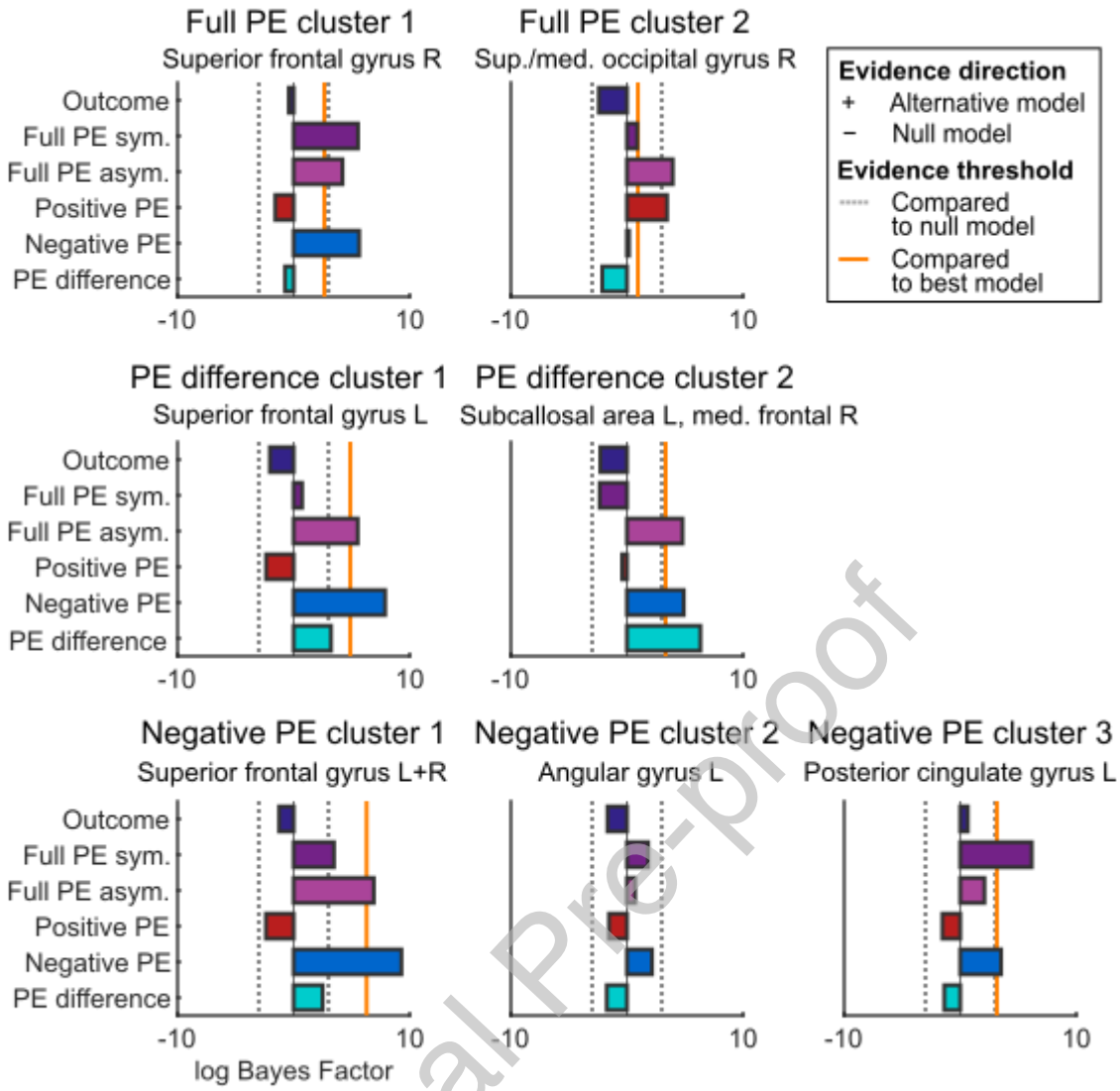

Figure 5. Model comparison of PE and outcome-only models for BOLD signals from the significant clusters for full signed PE (parametric GLM 1), difference positive (US occurrence) vs. negative (US omission) PE (parametric GLM 2) and negative PE (parametric GLM 3). Log Bayes Factors (BF) > 3 (dotted grey line) indicate moderate support for a model over the null model, whereas log BF < −3 denote moderate evidence for the null model, with values in between representing inconclusive evidence for any model. The orange line marks the evidence threshold (log BF 3) for moderate difference between the best model and other models. Full PE sym. = one intercept and slope parameter for both positive and negative PE; Full PE asym. = separate intercepts and slopes for positive and negative PE; Sup. = superior; Med. = medial. R = right hemisphere; L = left hemisphere. Note that this model comparison is meant for post-hoc illustrative purposes only, as the comparison is conducted on data that was already selected based on an association with one of the PE models in the whole-brain analysis.

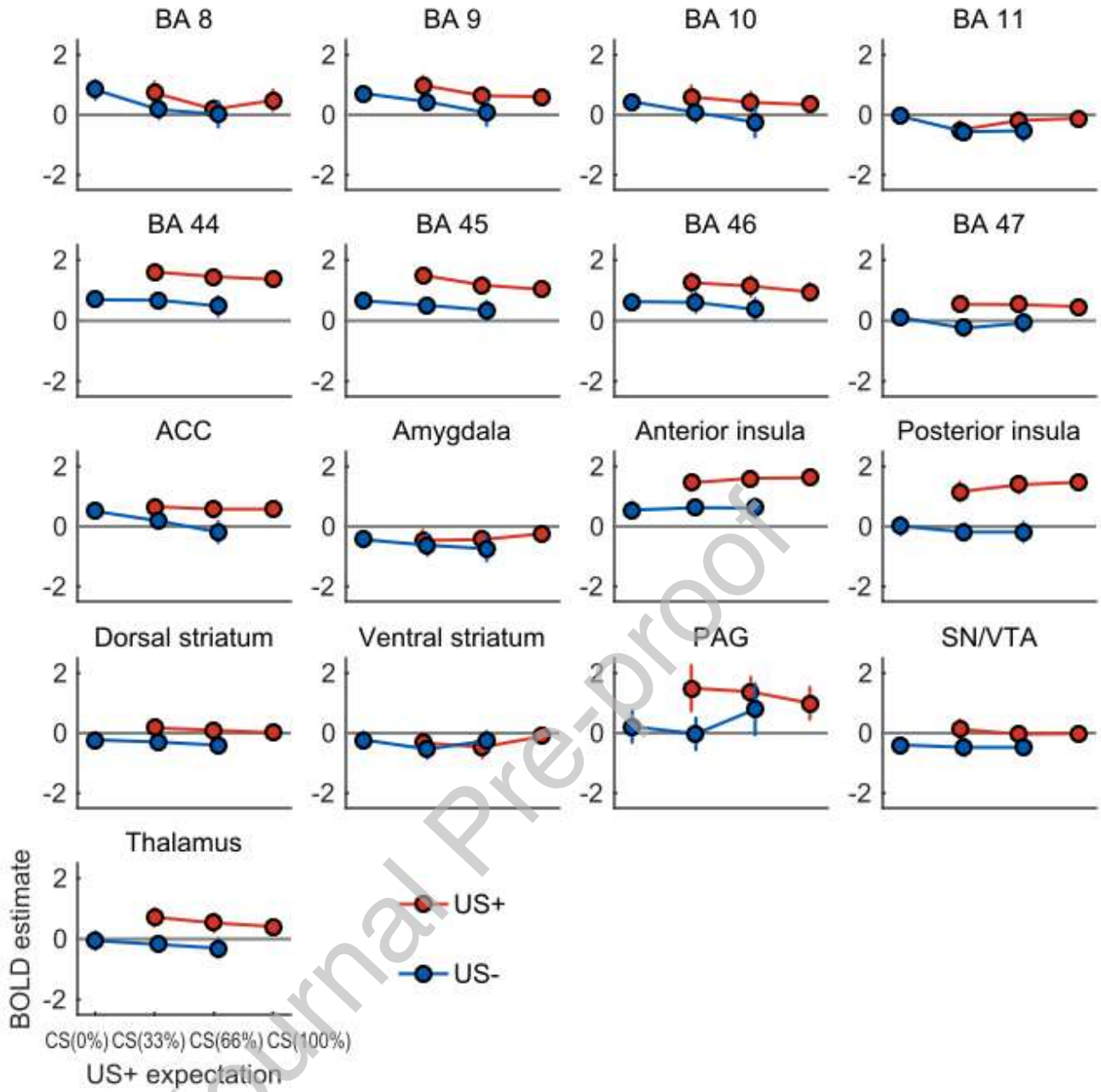

Figure 6. Average BOLD amplitude estimates during maintenance for each experimental condition extracted from the anatomical ROIs. Left and right hemispheres are combined. BA = Brodmann Area. ACC = Anterior Cingulate Cortex. PAG = Periaqueductal Grey. SN = Substantia Nigra. VTA = Ventral Tegmental Area. Error bars are within-subject standard errors of the mean. See Table 4 for Bayesian evidence and Supplementary Table 3 for frequentist effect sizes of the axiomatic comparisons for these ROIs.

Table 4. Bayesian evidence for axiomatic comparisons for anatomical regions-of-interest and significant functional clusters during maintenance of threat associations.

	Axiom 1 (reduced)	Axiom 2 (reduced)		Axiom 3
	US+ > US−	US+	US−	US+ > US−
ROI	Mean US+ > mean US−	CS(33%) > CS(100%)	CS(0%) > CS(66%)	CS(100%) = CS(0%)
	log BF	log BF	log BF	log BF
BA 8	0.34	−0.76	1.88	−0.85
BA 9	3.14 *	0.16	1.73	1.32
BA 10	1.86	−0.64	1.61	1.37
BA 11	−0.47	−2.31	1.32	1.41
BA 44	10.56 *	−0.20	−0.69	−4.93
BA 45	13.67 *	1.44	−0.12	−2.09
BA 46	5.56 *	−0.47	−0.48	0.34
BA 47	9.22 *	−1.08	−0.78	−0.96
ACC °	4.26 *	−0.22	2.03	2.57
Amygdala °	1.55	−1.30	0.50	2.24
Anterior insula	10.46 *	−1.99	−1.67	−6.73
Posterior insula	12.03 *	−2.20	−0.74	−8.90
Dorsal striatum	3.49 *	−0.48	−0.64	0.72
Ventral striatum	−1.49	−2.01	−1.40	1.32
PAG °	2.84	0.33	−1.99	0.32
SN/VTA	1.96	−0.81	−1.09	−1.03
Thalamus °	6.21 *	0.59	0.21	6.35 *
Full PE cluster 1	3.81 *	−0.30	12.24 *	0.16
Full PE cluster 2	1.22	3.71 *	1.45	−2.93
PE difference cluster 1	1.45	−1.74	5.80 *	0.78
PE difference cluster 2	−1.20	−2.45	6.06 *	−0.05
Negative PE cluster 1	2.52	−1.24	9.18 *	0.24
Negative PE cluster 2	1.35	−0.19	4.59 *	−0.56
Negative PE cluster 3	1.59	−0.46	6.23 *	0.57

Log BF = logarithm of Bayes Factor (BF). BF > 3 are commonly interpreted as moderate support for the alternative model over the null model, whereas log BF < −3 are interpreted as moderate evidence for the null model, with values in between representing inconclusive evidence for either model. The alternative model is indicated on the third row of the table, whereas the null model is its opposite (no difference between conditions for axiom 1 and 2, and a difference for axiom 3). * At least moderate evidence for the alternative hypothesis. ◦ ROIs with informed prior based on previous literature, the others have the default uninformative prior.

Discussion

Survival in biological environments requires learning associations between predictive cues and potential threatening outcomes. It has been suggested that such aversive learning is driven by prediction error (PE) signals, similarly to reward learning (Yau and McNally, 2018). Here, we used human BOLD fMRI to investigate neural representation of PEs after Pavlovian threat conditioning and under continuing CS-US pairings. We found no systematic evidence for theoretically expected neural PE signals. Instead, we identified regions that express PE signals more strongly when US was omitted as opposed to when US occurred. Such asymmetric PE representations cannot on their own be used to learn unbiased estimates of US probability (Dabney et al., 2020), although they might be combined, and could serve important biological functions in the context of a switch to extinction (Kim et al., 2020).

Neural representations of prediction errors

Our primary analysis revealed that BOLD activity in dorsomedial PFC and posterior parietal cortex correlated with signed PE. However, secondary analyses provided several arguments why these BOLD signals are unlikely to represent full signed PEs. First, average BOLD estimates from significant PE clusters did not fulfill the axiomatic criteria for PE representation (Caplin and Dean, 2008; Roy et al., 2014; Rutledge et al., 2010). Specifically, BOLD estimates showed differences between US expectation levels (axiom 2) only for US omission, but not for US occurrence. Bayesian model comparison (Fig. 5) suggested these BOLD signals were better or equally well explained by models that included asymmetric BOLD responses for unexpected US omission (negative PE) and US occurrence (positive PE). Second, a whole-brain search for negative PEs revealed significant BOLD activity in the dorsomedial and ventromedial PFC as well as rostral ACC that entirely encompassed, as well as extended beyond, the prefrontal full signed PE-encoding cluster. Meanwhile, no significant BOLD activity was associated with positive PEs only, over and above a constant representation of the US. Third, in a cluster in the vmPFC and rostral ACC, the encoding of positive and negative PEs was significantly different. This cluster expressed negative PEs more strongly than positive PEs.

Next, we explored whether any a priori anatomical regions of interest expressed PE signals. Formal model comparison revealed decisive evidence that averaged BOLD signals in BA 9 and ACC were better explained by full signed PE-encoding than alternative models, including some asymmetric models. In other areas, including PAG, Bayesian model comparison either supported outcome-encoding only, or the evidence was inconclusive or weak. Despite the full signed PE model winning the model comparison for two regions, there was no conclusive evidence that extracted BOLD signals from these, or any other region, fulfilled all of the axiomatic criteria for full signed PE-encoding.

Notably, some formal reinforcement learning models build on unsigned (absolute) rather than signed PEs (Li et al., 2011; Pearce and Hall, 1980). In our design, testing for the negative association of unsigned PEs to BOLD signal was formally equivalent to testing the slope difference between positive and

negative PEs ("PE difference" model). We did not observe a relation of unsigned PE signals with increased BOLD signal for any unexpected outcome. The opposite contrast (larger BOLD signal with smaller unsigned PE) showed two significant clusters. Extracted BOLD estimates from these cluster however appeared more consistent with negative PE signals only, although we note that a Bayesian analysis revealed some evidence of negative association with unsigned PEs for one of the two clusters.

Asymmetry of positive and negative threat prediction errors

Using designs different from ours, previous human neuroimaging studies have reported both positive and negative PEs in aversive learning to be represented in the same or in different brain regions (Seymour et al., 2005; Spoormaker et al., 2011; Roy et al., 2014; Shih et al., 2019 ; see also studies on learning-based US diminution, i.e. positive PE: Dunsmoor et al., 2008; Knight et al., 2010; Wood et al., 2013). For example during instrumental and pain intensity conditioning, Roy et al. (2014) found that BOLD activity in PAG fulfilled all of the axiomatic criteria for full signed PE signals. They also found that US expectation, but not axiomatic PE, was represented in the vmPFC, and positive PEs in the dmPFC. While instrumental and Pavlovian conditioning may engage distinct learning algorithms (Maia, 2010), there are also important differences between the Pavlovian conditioning experiments by Roy et al. (2014), and the present study. Specifically, Roy et al. used cues predicting different heat pain intensity, rather than different probability of presenting the same shock stimulus as in the present study; there were no fully predicted outcomes, and to derive PE they fitted a temporal difference learning model to participants' choices, which commits a priori to a specific learning model.

What could underlie the differential expression of positive and negative PE in our study? A first possible reason is to be found in asymmetric neural firing. Negative PEs in our study correspond to better-than-expected outcomes. Many dopaminergic midbrain neurons encode better-than-expected outcomes in increased firing rates, and worse-than-expected outcomes in reduced firing rates, but this reduction is often less pronounced than the increase (Schultz, 2016), despite variability between individual neurons (Dabney et al., 2020). Assuming an asymmetry in neural firing changes, and a constant noise level in the fMRI measurement, it might be more difficult to detect the smaller firing reduction than the larger firing increase. However, different from reward learning, there is currently no electrophysiological or voltammetric evidence for differential encoding of aversive PE in firing rates of the same neurons: those populations that respond to US occurrence have not been shown to be responsive to US omission (Groessl et al., 2018; Walker et al., 2020); as such this remains a speculative interpretation. Similarly, the time course of neural firing might differ between US occurrence and US omission, and this might make the hemodynamic model implicit in the fMRI analysis more appropriate for one or the other condition, thus hampering an unbiased estimation of the underlying changes in neural firing (see Supplementary Figures 1-4). Moreover, aversive positive PEs may be influenced to larger extent by baseline US responses than reward PEs, due to the high salience of

nociceptive stimuli. Accordingly, it has been shown in some rodent studies that neurons signaling aversive positive PEs also somewhat respond to the US alone (e.g., Walker et al., 2020). However, there is no clear pattern of US response with a (positive) PE-like response stacked on top in any same region in our data (but see e.g. PAG in Figure 6).

As a second possible reason, biased PE encoding in individual neurons can, when integrated on the population level, afford probabilistic learning (Dabney et al., 2020). This study addressed variability of reward PE encoding bias in neurons within one region, but the same mechanism could also act across regions. Indeed, opponent learning systems for reward and punishment have been suggested, possibly separated in different neurotransmitter systems and/or topographically distinct regions, and there are instances where these opponent learning systems might be integrated to solve a single task (Janssen et al., 2015; Skvortsova et al., 2014). Reward and punishment learning may also converge, for example in the case where avoiding punishment becomes rewarding (Palminteri et al., 2015). Consequently, the neural representations of PE may be different depending on the context and therefore the precise experimental design. Hence, it appears possible that positive PE are expressed in other brain areas that our present study was not optimally designed to detect. For example, positive PE have been suggested in PAG during aversive learning (Herry and Johansen, 2014; Walker et al., 2020), and our fMRI sequence was not specifically optimized for PAG coverage. It is also possible that aversive PEs, even for simple forms of associative learning, are not expressed in the same brain regions across mammal species, and therefore we cannot directly derive hypotheses for aversive PE representation from rodent studies.

As a final reason, some learning algorithms use teaching signals that are distinct from PE signals. For example, a normative Bayesian learner used previous work (Tzovara et al., 2018) requires only a categorical representation of the US to update its predictions. This raises the question whether the negative PE-encoding regions identified here are truly part of a learning system, or whether they encode an output signal that drives behavior after US omission. For example, mPFC has an important role in fear and extinction memory consolidation (Marek et al., 2013) and in signaling safety to the amygdala to diminish fear responses (Likhtik et al., 2014). The negative PE signals in the vmPFC in our study could reflect phasic safety signals in response to upward changes in environmental circumstances, consistent with previous studies (Fullana et al., 2016; Harrison et al., 2017), and could possibly engage inhibitory extinction learning. We note that previous studies have demonstrated a categorical representation of punishment omission in increased midbrain dopaminergic firing (Luo et al., 2018; Salinas-Hernández et al., 2018), whereas here we report a parametric PE-like representation in the opposite direction (lower BOLD activity with more unexpected omission). Previous research in humans has shown categorical dopamine-dependent activity patterns in a similar region in the ventromedial prefrontal cortex (Gerlicher et al., 2018).

Limitations

As a general limitation of the mass-univariate fMRI approach used here and in previous work, it is possible that PEs are represented by neural populations that are sparse (Reijmers et al., 2007), or that differ in sign and have an interleaved spatial organization, as has for example been shown for reward value representation in orbitofrontal cortex (Kahnt et al., 2014), CS+ representations in amygdala (Ciocchi et al., 2010; Haubensak et al., 2010), or biased PE signals in dopaminergic midbrain (Dabney et al., 2020). Multivariate analysis of high-resolution fMRI might be more appropriate to delineate such representations (Bach et al., 2011; Staib et al., 2020; Staib and Bach, 2018). Finally, the sample size of this study is small due to practical constraints. After the data acquisition for this study was finished, two further studies using the axiomatic approach with different paradigms were published (Fazeli and Büchel, 2018; Geuter et al., 2017). When taking the information from previous studies and the current study together, it is clear that much larger sample sizes will be required to robustly investigate US expectation effects that are much smaller than US outcome effects (i.e., Cohen's *d* = 0.15 to 0.4 on average, corresponding to sample sizes of N = 40 to hundreds of participants to reach statistical power of 80% with alpha level 0.05, see Supplementary Table 2). Next to increasing sample sizes to detect smaller effects, it would also be important to consider what size effects are considered theoretically or practically meaningful, and how to optimize our experimental designs to detect these effects (Melinscak and Bach, 2020). Finally, as we were unable to acquire valid psychophysiological signals in the MRI scanner environment and had opted to not measure US expectation on each trial due to potential effects on the learning process, we had no online index of learning, and as such cannot confirm whether participants' US expectations changed during the maintenance phase.

Conclusions

We found no evidence of full signed PE signals in any brain region but show that BOLD signals in a ventromedial prefrontal region encode negative PEs more strongly than positive PE. We speculate this may be due to biophysical asymmetries, integration of biased PE signals across regions, or reflect biological functions outside simple learning algorithms, such as engaging extinction learning. However, many open research questions remain regarding aversive PE signaling, which has been studied far less extensively than reward PEs. Future studies may shed light on the commonalities and differences of Pavlovian conditioning to different intensities or probabilities of aversive outcomes, further investigate the causal roles of the putative aversive PE encoding regions for learning, utilize optimized sequences to search for aversive PE signals in the midbrain and brainstem nuclei, and refine our understanding of possible sparse or multivariate encoding of aversive PEs in both humans and animal models.

Acknowledgements

We thank Samuel Gerster and Jules Brochard for technical assistance and Bogdan Draganski for continuing support. MRI data was acquired on the MRI platform of the Clinical Neuroscience Department, Lausanne University Hospital. This project was supported by Olga Mayenfisch Foundation and Swiss National Science Foundation (320030_149586 to DRB, 320030_188737 to AT, and 320030_184784 to AL). The Wellcome Centre for Human Neuroimaging receives core funding from the Wellcome Trust (091593/Z/10/Z). DRB is supported by funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. ERC-2018 CoG-816564 ActionContraThreat). AT is supported by the Interfaculty Research Cooperation "Decoding Sleep: From Neurons to Health & Mind" of the University of Bern and the Fondation Pierre Mercier pour la science. BAP is supported by Netherlands Organization for Scientific Research (NWO) VIDI 016.178.052 and by partial funding from R01 MH111444/MH/NIMH NIH. AL is supported by the ROGER DE SPOELBERCH Foundation.

Author contributions

Conceptualization DRB, AT; Data curation KEO, AT, DRB; Formal analysis KEO, DRB; Funding acquisition DRB; Investigation AT, DRB, AL; Methodology AT, BP, AL, DRB; Project administration KEO, AT, DRB; Resources BP, AT, DRB; Software KEO, AT, DRB; Supervision DRB; Validation KEO, DRB; Visualization KEO; Roles/Writing - original draft KEO, AT, DRB; Writing - review & editing BP, AL

Data and Code Availability Statement

The code for the experiment, data analysis and figures are available in a public repository gitlab.com/kojala/threatlearning_fmri. Group-level unthresholded Statistical Parametric Maps, ROI masks and mean beta values relevant to the analyses are available in a public repository with DOI 10.5281/zenodo.6983543. Data from the behavioral experiment outside the scanner are available in a public repository with DOI 10.5281/zenodo.3872055. Due to data protection regulations and given a risk of statistical identification of brain anatomy, individual-level MRI data are available from the authors for scientific purposes under a data protection agreement.

References

Abivardi, A., Bach, D.R., 2017. Deconstructing white matter connectivity of human amygdala nuclei with thalamus and cortex subdivisions in vivo. Hum. Brain Mapp. 38, 3927–3940. https://doi.org/10.1002/hbm.23639

Bach, D.R., Weiskopf, N., Dolan, R.J., 2011. A Stable Sparse Fear Memory Trace in Human Amygdala. J. Neurosci. 31, 9383–9389. https://doi.org/10.1523/JNEUROSCI.1524-11.2011

Boll, S., Gamer, M., Gluth, S., Finsterbusch, J., Büchel, C., 2013. Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. Eur. J. Neurosci. 37, 758–767. https://doi.org/10.1111/ejn.12094

Caplin, A., Dean, M., 2008. Axiomatic methods, dopamine and reward prediction error. Curr. Opin. Neurobiol. 18, 197–202. https://doi.org/10.1016/j.conb.2008.07.007

Chang, C.Y., Gardner, M., Di Tillio, M.G., Schoenbaum, G., 2017. Optogenetic Blockade of Dopamine Transients Prevents Learning Induced by Changes in Reward Features. Curr. Biol. 27, 3480-3486.e3. https://doi.org/10.1016/j.cub.2017.09.049

Ciocchi, S., Herry, C., Grenier, F., Wolff, S.B.E., Letzkus, J.J., Vlachos, I., Ehrlich, I., Sprengel, R., Deisseroth, K., Stadler, M.B., Müller, C., Lüthi, A., 2010. Encoding of conditioned fear in central amygdala inhibitory circuits. Nature 468, 277–282. https://doi.org/10.1038/nature09559

Corbin, N., Todd, N., Friston, K.J., Callaghan, M.F., 2018. Accurate modeling of temporal correlations in rapidly sampled fMRI time series. Hum. Brain Mapp. 39, 3884–3897. https://doi.org/10.1002/hbm.24218

Dabney, W., Kurth-Nelson, Z., Uchida, N., Starkweather, C.K., Hassabis, D., Munos, R., Botvinick, M., 2020. A distributional code for value in dopamine-based reinforcement learning. Nature 577, 671–675. https://doi.org/10.1038/s41586-019-1924-6

Davenport, S., Nichols, T.E., 2020. Selective peak inference: Unbiased estimation of raw and standardized effect size at local maxima. Neuroimage 209, 116375. https://doi.org/10.1016/j.neuroimage.2019.116375

Dunsmoor, J.E., Bandettini, P.A., Knight, D.C., 2008. Neural correlates of unconditioned response diminution during Pavlovian conditioning. Neuroimage 40, 811–817. https://doi.org/10.1016/j.neuroimage.2007.11.042

Eklund, A., Nichols, T.E., Knutsson, H., 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. Proc. Natl. Acad. Sci. 113, 7900–7905. https://doi.org/10.1073/pnas.1602413113

Esser, R., Korn, C.W., Ganzer, F., Haaker, J., 2021. L-DOPA modulates activity in the vmPFC, nucleus accumbens, and VTA during threat extinction learning in humans. Elife 10, 1–21. https://doi.org/10.7554/eLife.65280

Fazeli, S., Büchel, C., 2018. Pain-Related Expectation and Prediction Error Signals in the Anterior Insula Are Not Related to Aversiveness. J. Neurosci. 38, 6461–6474. https://doi.org/10.1523/JNEUROSCI.0671-18.2018

Fox, J., Friendly, M., Monetter, G., 2018. heplots: Visualizing Tests in Multivariate Linear Models.

Fullana, M.A., Harrison, B.J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., Radua, J., 2016. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. Mol. Psychiatry 2188, 500–508. https://doi.org/10.1038/mp.2015.88

Gerlicher, A.M.V., Tüscher, O., Kalisch, R., 2018. Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. Nat. Commun. 9. https://doi.org/10.1038/s41467-018-06785-y

Geuter, S., Boll, S., Eippert, F., Büchel, C., 2017. Functional dissociation of stimulus intensity encoding and predictive coding of pain in the insula. Elife 6, e24770. https://doi.org/10.7554/eLife.24770

Glover, G.H., Li, T.Q., Ress, D., 2000. Image-based method for retrospective correction of physiological motion effects in fMRI: RETROICOR. Magn. Reson. Med. 44, 162–167. https://doi.org/10.1002/1522-2594(200007)44:1<162::AID-MRM23>3.0.CO;2-E

Griswold, M.A., Jakob, P.M., Heidemann, R.M., Nittka, M., Jellus, V., Wang, J., Kiefer, B., Haase, A., 2002. Generalized Autocalibrating Partially Parallel Acquisitions (GRAPPA). Magn. Reson. Med. 47, 1202–1210. https://doi.org/10.1002/mrm.10171

Groessl, F., Munsch, T., Meis, S., Griessner, J., Kaczanowska, J., Pliota, P., Kargl, D., Badurek, S., Kraitsy, K., Rassoulpour, A., Zuber, J., Lessmann, V., Haubensak, W., 2018. Dorsal tegmental dopamine neurons gate associative learning of fear. Nat. Neurosci. 21, 952–962. https://doi.org/10.1038/s41593-018-0174-5

Harrison, B.J., Fullana, M.A., Via, E., Soriano-Mas, C., Vervliet, B., Martínez-Zalacaín, I., Pujol, J., Davey, C.G., Kircher, T., Straube, B., Cardoner, N., 2017. Human ventromedial prefrontal cortex and the positive affective processing of safety signals. Neuroimage 152, 12–18. https://doi.org/10.1016/j.neuroimage.2017.02.080

Hart, A.S., Rutledge, R.B., Glimcher, P.W., Phillips, P.E.M., 2014. Phasic dopamine release in the rat nucleus accumbens symmetrically encodes a reward prediction error term. J. Neurosci. 34, 698–704. https://doi.org/10.1523/JNEUROSCI.2489-13.2014

Haubensak, W., Kunwar, P.S., Cai, H., Ciocchi, S., Wall, N.R., Ponnusamy, R., Biag, J., Dong, H.-W., Deisseroth, K., Callaway, E.M., Fanselow, M.S., Lüthi, A., Anderson, D.J., 2010. Genetic dissection of an amygdala microcircuit that gates conditioned fear. Nature 468, 270–276. https://doi.org/10.1038/nature09553

Herry, C., Johansen, J.P., 2014. Encoding of fear learning and memory in distributed neuronal circuits. Nat. Neurosci. 17, 1644–1654. https://doi.org/10.1038/nn.3869

Hutton, C., Bork, A., Josephs, O., Deichmann, R., Ashburner, J., Turner, R., 2002. Image distortion correction in fMRI: A quantitative evaluation. Neuroimage 16, 217–240. https://doi.org/10.1006/nimg.2001.1054

Janssen, L.K., Sescousse, G., Hashemi, M.M., Timmer, M.H.M., Ter Huurne, N.P., Geurts, D.E.M., Cools, R., 2015. Abnormal modulation of reward versus punishment learning by a dopamine D2-receptor antagonist in pathological gamblers. Psychopharmacology (Berl). 232, 3345–3353. https://doi.org/10.1007/s00213-015-3986-y

JASP Team, 2022. JASP.

Johansen, J.P., Tarpley, J.W., LeDoux, J.E., Blair, H.T., 2010. Neural substrates for expectation-modulated fear learning in

the amygdala and periaqueductal gray. Nat. Neurosci. 13, 979–86. https://doi.org/10.1038/nn.2594
Kahnt, T., Park, S.Q., Haynes, J.-D., Tobler, P.N., 2014. Disentangling neural representations of value and salience in the human brain. Proc. Natl. Acad. Sci. U. S. A. 111, 5000–5. https://doi.org/10.1073/pnas.1320189111
Kasper, L., Bollmann, S., Diaconescu, A.O., Hutton, C., Heinzle, J., Iglesias, S., Hauser, T.U., Sebold, M., Manjaly, Z.M., Pruessmann, K.P., Stephan, K.E., 2017. The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. J. Neurosci. Methods 276, 56–72. https://doi.org/10.1016/j.jneumeth.2016.10.019
Keuken, M.C., Bazin, P.L., Backhouse, K., Beekhuizen, S., Himmer, L., Kandola, A., Lafeber, J.J., Prochazkova, L., Trutti, A., Schäfer, A., Turner, R., Forstmann, B.U., 2017. Effects of aging on T1 , T2∗ , and QSM MRI values in the subcortex. Brain Struct. Funct. 222, 2487–2505. https://doi.org/10.1007/s00429-016-1352-4
Kim, O.A., Ohmae, S., Medina, J.F., 2020. A cerebello-olivary signal for negative prediction error is sufficient to cause extinction of associative motor learning. Nat. Neurosci. 23, 1550–1554. https://doi.org/10.1038/s41593-020-00732-1
Knight, D.C., Waters, N.S., King, M.K., Bandettini, P.A., 2010. Learning-related diminution of unconditioned SCR and fMRI signal responses. Neuroimage 49, 843–848. https://doi.org/10.1016/j.neuroimage.2009.07.012
Korn, C.W., Bach, D.R., 2016. A solid frame for the window on cognition: Modeling event-related pupil responses. J. Vis. 16, 28. https://doi.org/10.1167/16.3.28
Korn, C.W., Staib, M., Tzovara, A., Castegnetti, G., Bach, D.R., 2017. A pupil size response model to assess fear learning. Psychophysiology 54, 330–343. https://doi.org/10.1111/psyp.12801
Lak, A., Stauffer, W.R., Schultz, W., 2016. Dopamine neurons learn relative chosen value from probabilistic rewards. Elife 5, 1–19. https://doi.org/10.7554/eLife.18044
Li, J., Schiller, D., Schoenbaum, G., Phelps, E.A., Daw, N.D., 2011. Differential roles of human striatum and amygdala in associative learning. Nat. Neurosci. 14, 1250–1252. https://doi.org/10.1038/nn.2904
Li, S.S.Y., McNally, G.P., 2014. The conditions that promote fear learning: Prediction error and Pavlovian fear conditioning. Neurobiol. Learn. Mem. 108, 14–21. https://doi.org/10.1016/j.nlm.2013.05.002
Likhtik, E., Stujenske, J.M., Topiwala, M.A., Harris, A.Z., Gordon, J.A., 2014. Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety. Nat. Neurosci. 17, 106–113. https://doi.org/10.1038/nn.3582
Lovibond, P.F., Shanks, D.R., 2002. The role of awareness in Pavlovian conditioning: Empirical evidence and theoretical implications. J. Exp. Psychol. Anim. Behav. Process. 28, 3–26. https://doi.org/10.1037/0097-7403.28.1.3
Luo, R., Uematsu, A., Weitemier, A., Aquili, L., Koivumaa, J., McHugh, T.J., Johansen, J.P., 2018. A dopaminergic switch for fear to safety transitions. Nat. Commun. 9, 1–11. https://doi.org/10.1038/s41467-018-04784-7
Lutti, A., Thomas, D.L., Hutton, C., Weiskopf, N., 2013. High-resolution functional MRI at 3 T: 3D/2D echo-planar imaging with optimized physiological noise correction. Magn. Reson. Med. 69, 1657–1664. https://doi.org/10.1002/mrm.24398
Maia, T. V., 2010. Two-factor theory, the actor-critic model, and conditioned avoidance. Learn. Behav. 38, 50–67. https://doi.org/10.3758/LB.38.1.50
Makowski, D., Ben-Shachar, M., Lüdecke, D., 2019. bayestestR: Describing Effects and their Uncertainty, Existence and Significance within the Bayesian Framework. J. Open Source Softw. 4, 1541. https://doi.org/10.21105/joss.01541
Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H., 2003. An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19, 1233–1239. https://doi.org/10.1016/S1053-8119(03)00169-1
Marek, R., Strobel, C., Bredy, T.W., Sah, P., 2013. The amygdala and medial prefrontal cortex: Partners in the fear circuit. J. Physiol. 591, 2381–2391. https://doi.org/10.1113/jphysiol.2012.248575
Maris, E., Oostenveld, R., 2007. Nonparametric statistical testing of EEG- and MEG-data. J. Neurosci. Methods 164, 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024
Melinscak, F., Bach, D.R., 2020. Computational optimization of associative learning experiments. PLOS Comput. Biol. 16, e1007593. https://doi.org/10.1371/journal.pcbi.1007593
Navarro, D.J., 2015. Learning statistics with R: A tutorial for psychology students and other beginners.
Ojala, K.E., Bach, D.R., 2020. Measuring learning in human classical threat conditioning: Translational, cognitive and methodological considerations. Neurosci. Biobehav. Rev. 114, 96–112. https://doi.org/10.1016/j.neubiorev.2020.04.019
Ozawa, T., Ycu, E.A., Kumar, A., Yeh, L.-F., Ahmed, T., Koivumaa, J., Johansen, J.P., 2017. A feedback neural circuit for calibrating aversive memory strength. Nat. Neurosci. 20, 90–97. https://doi.org/10.1038/nn.4439
Palminteri, S., Khamassi, M., Joffily, M., Coricelli, G., 2015. Contextual modulation of value signals in reward and punishment learning. Nat. Commun. 6. https://doi.org/10.1038/ncomms9096
Pauli, W.M., Larsen, T., Collette, S., Tyszka, J.M., Seymour, B., O'Doherty, J.P., 2015. Distinct Contributions of Ventromedial and Dorsolateral Subregions of the Human Substantia Nigra to Appetitive and Aversive Learning. J. Neurosci. 35, 14220–14233. https://doi.org/10.1523/JNEUROSCI.2277-15.2015
Pauli, W.M., Nili, A.N., Michael Tyszka, J., 2018. Data Descriptor: A high-resolution probabilistic in vivo atlas of human

subcortical brain nuclei. Sci. Data 5, 1–13. https://doi.org/10.1038/sdata.2018.63
Pearce, J.M., Hall, G., 1980. A model for stimulus generalization in Pavlovian conditioning: variation in the effectiveness of conditioned but not unconditioned stimuli. Psychol. Rev. 87, 532–552. https://doi.org/http://dx.doi.org/10.1037/0033-295X.87.6.532
Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., RCoreTeam, 2020. nlme: Linear and Nonlinear Mixed Effects Models.
Poser, B.A., Versluis, M.J., Hoogduin, J.M., Norris, D.G., 2006. BOLD contrast sensitivity enhancement and artifact reduction with multiecho EPI: Parallel-acquired inhomogeneity-desensitized fMRI. Magn. Reson. Med. 55, 1227–1235. https://doi.org/10.1002/mrm.20900
RCoreTeam, 2019. R: A language and environment for statistical computing.
Reijmers, L.G., Perkins, B.L., Matsuo, N., Mayford, M., 2007. Localization of a stable neural correlate of associative memory. Science (80-.). 317, 1230–1233. https://doi.org/10.1126/science.1143839
Rescorla, R.A., Wagner, A.R., 1972. A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement, in: Prokasy, A., Black, W.F. (Eds.), Classical Conditioning II: Current Research and Theory. Appleton-Century-Crofts, New York, NY, pp. 64–99. https://doi.org/10.1101/gr.110528.110
Roy, M., Shohamy, D., Daw, N., Jepma, M., Wimmer, G.E., Wager, T.D., 2014. Representation of aversive prediction errors in the human periaqueductal gray. Nat. Neurosci. 17, 1607–12. https://doi.org/10.1038/nn.3832
RStudioTeam, 2018. RStudio: Integrated Development for R.
Rutledge, R.B., Dean, M., Caplin, A., Glimcher, P.W., 2010. Testing the Reward Prediction Error Hypothesis with an Axiomatic Model. J. Neurosci. 30, 13525–13536. https://doi.org/10.1523/JNEUROSCI.1747-10.2010
Salinas-Hernández, X.I., Vogel, P., Betz, S., Kalisch, R., Sigurdsson, T., Duvarci, S., 2018. Dopamine neurons drive fear extinction learning by signaling the omission of expected aversive outcomes. Elife 7, 1–25. https://doi.org/10.7554/eLife.38818
Sassenhagen, J., Draschkow, D., 2019. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. Psychophysiology 56, 1–8. https://doi.org/10.1111/psyp.13335
Schultz, W., 2016. Dopamine reward prediction error coding. Nat. Rev. Neurosci. 17, 183–195. https://doi.org/10.1038/nrn.2015.26
Schultz, W., Dickinson, A., 2000. Neuronal Coding of Prediction Errors. Annu. Rev. Neurosci. 23, 473–500. https://doi.org/10.1146/annurev.neuro.23.1.473
Seymour, B., 2019. Pain: A Precision Signal for Reinforcement Learning and Control. Neuron 101, 1029–1041. https://doi.org/10.1016/j.neuron.2019.01.055
Seymour, B., O'Doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., Frackowiak, R., 2004. Temporal difference models describe higher order learning in humans. Nature 429, 664–667. https://doi.org/10.1038/nature02581
Seymour, B., O'Doherty, J.P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K.J., Dolan, R., 2005. Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. Nat. Neurosci. 8, 1234–1240. https://doi.org/10.1038/nn1527
Shih, Y.W., Tsai, H.Y., Lin, F.S., Lin, Y.H., Chiang, C.Y., Lu, Z.L., Tseng, M.T., 2019. Effects of positive and negative expectations on human pain perception engage separate but interrelated and dependently regulated cerebral mechanisms. J. Neurosci. 39, 1261–1274. https://doi.org/10.1523/JNEUROSCI.2154-18.2018
Skvortsova, V., Palminteri, S., Pessiglione, M., 2014. Learning to minimize efforts versus maximizing rewards: Computational principles and neural correlates. J. Neurosci. 34, 15621–15630. https://doi.org/10.1523/JNEUROSCI.1350-14.2014
Spoormaker, V.I., Andrade, K.C., Schröter, M.S., Sturm, A., Goya-Maldonado, R., Sämann, P.G., Czisch, M., 2011. The neural correlates of negative prediction error signaling in human fear conditioning. Neuroimage 54, 2250–2256. https://doi.org/10.1016/j.neuroimage.2010.09.042
Staib, M., Abivardi, A., Bach, D.R., 2020. Primary auditory cortex representation of fear-conditioned musical sounds. Hum. Brain Mapp. 41, 882–891. https://doi.org/10.1002/hbm.24846
Staib, M., Bach, D.R., 2018. Stimulus-invariant auditory cortex threat encoding during fear conditioning with simple and complex sounds. Neuroimage 166, 276–284. https://doi.org/10.1016/j.neuroimage.2017.11.009
Steinberg, E.E., Keiflin, R., Boivin, J.R., Witten, I.B., Deisseroth, K., Janak, P.H., 2013. A causal link between prediction errors, dopamine neurons and learning. Nat. Neurosci. 16, 966–973. https://doi.org/10.1038/nn.3413
Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15, 273–289. https://doi.org/10.1006/nimg.2001.0978
Tzovara, A., Korn, C.W., Bach, D.R., 2018. Human Pavlovian fear conditioning conforms to probabilistic learning. PLoS Comput. Biol. 14, e1006243. https://doi.org/10.1371/journal.pcbi.1006243
Van Essen, D.C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T.E.J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S.W., Della Penna, S., Feinberg, D., Glasser, M.F., Harel, N., Heath, A.C., Larson-Prior, L., Marcus, D.,

Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S.E., Prior, F., Schlaggar, B.L., Smith, S.M., Snyder, A.Z., Xu, J., Yacoub, E., 2012. The Human Connectome Project: A data acquisition perspective. Neuroimage 62, 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Wagenmakers, E.J., 2007. A practical solution to the pervasive problems of p values. Psychon. Bull. Rev. 14, 779–804. https://doi.org/10.3758/BF03194105

Walker, R.A., Wright, K.M., Jhou, T.C., McDannald, M.A., 2020. The ventrolateral periaqueductal grey updates fear via positive prediction error. Eur. J. Neurosci. 51, 866–880. https://doi.org/10.1111/ejn.14536

Wood, K.H., Kuykendall, D., Ver Hoef, L.W., Knight, D.C., 2013. Neural Substrates Underlying Learning-Related Changes of the Unconditioned Fear Response. Open Neuroimag. J. 7, 41–52. https://doi.org/10.2174/1874440001307010041

Yau, J.O.Y., McNally, G.P., 2018. Brain Mechanisms Controlling Pavlovian Fear Conditioning. J. Exp. Psychol. Anim. Learn. Cogn. 44, 341–357. https://doi.org/http://dx.doi.org/10.1037/xan0000181

Zhang, S., Mano, H., Ganesh, G., Robbins, T., Seymour, B., 2016. Dissociable Learning Processes Underlie Human Pain Conditioning. Curr. Biol. 26, 52–58. https://doi.org/10.1016/j.cub.2015.10.066