Supplementary Material

Blazej M. Baczkowski[1,2,3,4] & co-authors

[1] IMPRS NeuroCom, Leipzig, Germany

[2] Department of Neurology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[3] Institute of Psychology, University of Leipzig, Leipzig, Germany

[4] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany

Supplementary Material

Supplementary Methods

Participants

The study was approved by the local ethical review board (CMO for Arnhem-Nijmegen region) and carried out in accordance with the Declaration of Helsinki. We recruited healthy right-handed volunteers between 18 and 35 years old with normal or corrected-to-normal vision from a student population in Nijmegen, the Netherlands. Eligibility criteria excluded pregnancy, current or a history of a neurological or psychiatric disorder, a disorder of the autonomic nervous system, heart conditions and weekly recreational drug-use. Further eligibility criteria excluded use of medication and excessive alcohol consumption 72 hours and 24 hours before the experiment, respectively. To increase the sample size, five volunteers who reported their regular, non-psychotropic medication (e.g., for allergy), were invited. Forty-four volunteers provided their written informed consent prior to the start of the study. Forty-one participants (29 females) between 19 to 34 (mode=20) years old completed the study. To ensure sufficient data quality, we specified two pre-registered exclusion criteria (see Table S1 for data overview). First, to ensure that we could assess the effects of relational memory structure, participants who exhibited low memory performance prior to the conditioning phase (less than 80% accuracy on the directly learned associations[1] on the previous day) were excluded from the analyses (n=5). Second, to ensure that we could assess the inferential expression of Pavlovian threat memory, which is dependent on successful acquisition of conditioned learning, participants with insufficient data quality (n=3 for skin conductance, and n=2 for pupil size, respectively), who did not comply to the study instructions (n=2), or who did not exhibit differential conditioned threat responses (numerical difference between CS+ and CS- greater than 0) in the second half of conditioning phase were excluded (n=6 for shock expectancy ratings, n=11 for skin conductance, and n=12 for pupil size response). Therefore, the final sample of participants who completed the experiment with sufficient data quality and met the inclusion criteria for memory performance and differential conditioning consisted of n=30 (24 females), n=26 (22

females), and n=24 (20 females) participants, respectively for each response modality, which met our pre-registered minimum sample size.

Procedure

Two experimental sessions took place on two consecutive days in the same test room. Stimulus presentations were generated with the Psychophysics Toolbox[2] for MATLAB 2016a (the MathWorks). Participants viewed the stimuli from a distance of approximately 60 cm on a 24" flat panel display (BenQ XL2420T, resolution 1920 x 1080, aspect ratio 16:9, refresh rate 60Hz) in a dimly lit room. In the first session (~90 min), participants were first asked to fill out four questionnaires: (a) the trait inventory of the State-Trait Anxiety Inventory-STAI[3]; (b) Childhood Trauma Questionnaire-CTQ[4,5]; (c) Intolerance of Uncertainty Scale-IUS[6]; and (d) Berkman-Syme Social Network Index-SNI[7]. Subsequently, they performed a paired associate learning (PAL) task during which they were explicitly informed that they would learn to associate images with one another in a trial and error fashion. Their task was to improve their performance level expressed in a percent of correct responses and their average speed throughout the whole phase.

The second session (~90 min) was scheduled to take place on the next day. It consisted of four experimental tasks.

To assess participants long-term memory for the relations among the images that they had learned during day 1, the experimental session started with a 2-Alternative Forced Choice (2-AFC) task. During the task, stimulus presentations within a trial followed the scheme from the paired associate learning (PAL) task but participants did not receive any feedback on their performance.

The session proceeded with the attachment of equipment to collect pupil size, skin conductance, and pulse oximeter as well as electrodes to administer mild electric shocks. To stabilise participants' heads, they were positioned on a chin rest in front of a computer display. An eye tracking camera was placed approx. 50 cm in front of the participants' eyes and adjusted to properly detect the size of their left pupil, followed by a 5-point calibration

procedure. Participants were asked to perform the tasks without vision correction if they were able to properly see the pictures on the screen. Skin conductance electrodes were attached to the non-dominant hand on the intermediate phalanges of the index and middle fingers. An optical heart-rate sensor (a photoplethysmogram, PPG) was attached to a ring finger on a non-dominant hand. Shock electrodes were attached to a dominant hand on the intermediate phalanges of the ring and pinky fingers. The shock intensity level was calibrated using an ascending staircase procedure starting with a low voltage (near a perceptible threshold) to reach a level deemed "maximally uncomfortable without being painful" by the participant. The intensity level was subsequently scored on a pain assessment scale from 0 (no sensation) to 9 (very high intensity) and ranged from 6 to 9 (mode=7) in the current study (n=41).

To impose Pavlovian threat memory within the pre-existing relational memory structure, participants were subsequently exposed to differential delay threat conditioning including two familiar images from the previous session that were allocated to the opposite ends of the relational memory graph. During conditioning, skin conductance, pupil size, and shock expectancy ratings were collected. Participants were told that some of the familiar images from the previous day may co-occur with a shock and instructed to predict receiving a shock based on an image they saw, but no explicit information was given regarding the shock-image contingencies so that they could learn it from reinforcement experience. To mitigate the potential that participants perceived threat conditioning as an unrelated task from the previous memory training, they were told to keep in mind the previously learned associations among images.

Next, to give participants a brief break, they watched a 7-minute abstract animation movie Inscapes[8]. They were informed that this was a break, no data was being collected, and they would not receive any shocks (the shocker was switched off). After the break, the shocker was switched on (the intensity level remained the same as indicated by the calibration procedure) and participants were exposed to the test for the inferential expression of Pavlovian threat memory during which all images from the previous day were presented.

Participants were told that the study continued as before including the same instructions as during the conditioning phase that were repeated.

After the inference test, all physiological measurements were stopped and the shock electrodes were removed. To test if the relational memory structure remained stable, in the last experimental phase of the study, participants performed again the 2-AFC task with the same instructions as at the beginning of the session. At the conclusion of the study participants were asked to rate the intensity of the shock felt during the session on a scale from 1 (not at all unpleasant) to 9 (extremely unpleasant), and how much fear they felt during the task with shocks from 1 (not at all afraid) to 9 (extremely afraid). They were also asked to estimate how many shocks they had received throughout the whole study (including during conditioning and the inference test, but not counting during the calibration phase) and identify the image that was exclusively paired with the shock among all images presented during the experiment. Finally, they were asked a binary question (yes or no) whether during the task with shocks they were thinking about relations among images that they had learned on the previous day. The study ended with a debriefing and participants were financially compensated for their participation.

Statistical analyses

To assess the structure of relational memory which requires integration of overlapping elements, we restricted the analyses of the 2-AFC task to only those participants who successfully retrieved relations of directly learned pairs presented on day 1, i.e., the premise associations[1]. To this end, we specified a pre-registered threshold of min. 80% correct responses achieved solely in trials probing an association between two elements that were allocated as neighbours on the memory graph. Given the 80% threshold for premise pairs, the chance level equaled to 72.8% (0.8 * 76 trials + 0.5 * 24 trials) in selected participants. To assure that the relational memory was integrated before Pavlovian conditioning, we applied the inclusion criterion only to the data from the 2-AFC task that was performed at the beginning of the session on day 2, i.e., prior to conditioning and test for the inferential expression of threat memory.

To test for the linear organization of relational memory and its stability over time on day 2, we specified statistical models that quantify the linear (and polynomial) effect of the node in the memory graph at the beginning and at the end of the session on day 2 on representational scale values recovered through the maximum likelihood difference scaling (MLDS) based on binary responses in the 2-AFC task. To model the representational scale values that fall in the interval [0,1], we fitted a generalized linear mixed model (GLMM) with a *logit* link function and beta error structure. Because the beta distribution is a continuous probability distribution defined on the interval (0,1), we scaled the representational values to avoid 0s and 1s, using the following formula:

$$x' = (x * (N-1) + s)/N$$

where $s = 0.5$ and N is the sample size[9]. The full model included the linear, quadratic, and cubic trends of the node (scaled so that mean equals to zero) nested within a two-level factor of time as test predictors. To select the best model, the test predictors were successively dropped and compared to a null model without any of the test predictors. To account for the data non-independence, we nested a random intercept and a nested linear effect of node within a subject (36 levels) in every model.

To ensure that we could assess the inferential expression of Pavlovian threat memory, we included in the analyses only those participants who complied to the study instructions, exhibited sufficient data quality, and successfully acquired Pavlovian threat memory, i.e., showed numerical difference between CS+ and CS- greater than 0 in the 2nd half of conditioning phase for respective response modality, a procedure previously implemented in other reports[e.g., 10]. To describe our sample and quantify the magnitude of the conditioning effect in the selected participants, we calculated an effect size of the difference between the CS+ and the CS- in the 2nd half of conditioning for each modality. To measure the difference between the proportions of shock expectancy for the CS+ and CS-, we used odds ratio calculated as $exp(\beta)$, where β is extracted from the GLMM revealing the *logit* difference between the two conditions. To measure the difference between the CS+ and CS-

based on SCR and PSR, we calculated Hedges'g[11].

To test for the gradient of threat responses indicating the inferrential expression of Pavlovian threat memory based on the pre-existing relational knowledge, we quantified a linear trend across GS1-GS4 stimuli. For each response modality, we specified a G/LMM that included a test predictor of a linear trend across GS1-GS4 (coded as a numerical predictor [0, -3, -1, 1, 3, 0]) and control predictors of the CS+ and the CS-. To account for the potential perceptual similarity among images, we additionally included a six-level factor of stimulus identity as a control predictor. To model the proportion of responses indicating shock expectancy, we specified a GLMM with a *logit* link function and a binomial error structure while to model the magnitude of SCR/PSR, we specified LMMs with an identity link function and a Gaussian error structure. To account for the data non-independence, we aimed to include a full random structure of the predictors nested within a subject in every model[12]. When the full random structure led to convergence problems, it was subsequently reduced to achieve model convergance. To explore non-linearity in the gradients of the SCR/PSR, we specified two additional models: one including an additional test predictor of a quadratic trend and one with a *log* link function, instead of an identity link, that exponentiates the linear predictor and models an exponential decrease of expected values across GS1-GS4. The model fitted to the shock expectancy already included a non-linear predictor due to the *logit* link function.

To explore whether the gradient of threat responses evoked by GS1-GS4 stimuli is associated with the retrieval of relational memory prior to the inference test, we specified a linear model that quantified the effect of performance in the 2-AFC task on the individual slope of the gradient estimated from the G/LMMs for each modality. To make the analyses across the three modalities comparable, the slope estimates were z transformed.

Finally, to explore whether the gradient was associated with individual differences in trait anxiety, tolerance for uncertainty, and childhood traumatic experiences, we specified a multiple linear model that quantifies the effect of individual scores in three self-report questionnaires, i.e., STAI, IUS, and CTQ, on the individual slope of the gradient estimated

from the G/LMMs for each modality.

Skin conductance response estimation

Skin conductance was collected using BrainVision BrainAmp ExG (Brain Products GmbH, Germany) with Ag/AgCl electrodes for galvanic skin response at a sampling rate of 500 Hz. The quality of the skin conductance time series were visually inspected to check for artifacts suggesting malfunction of the recording system and/or lack of repeated increases to the presentation of the electric shocks. Continuous raw skin conductance time series were analysed in the window starting from 10 s prior to the onset of the first stimulus and terminating 16 s after the offset of the last stimulus in a task, i.e., either conditioning or inference test. The time series were filtered with a band-pass Butterworth filter (0.05-5.0 Hz). Skin conductance response (SCR) was estimated with a Trough-to-Peak-scoring method[10,13] such that responses were determined for each trial as the through-to-peak amplitude difference in skin conductance of the largest deflection in the latency window from 0-8 seconds after stimulus onset, i.e. maximum SCR value minus the minimum value that precedes the maximum value in time[14]. If a response did not meet these criteria, then the trial was scored as a zero. To eliminate SCR scaling differences caused by peripheral factors such as skin properties, within-subjects trial-wise responses for each task were range-corrected by dividing each response by the highest response (typically elicited by the shock)[15]. The magnitude of SCR was then computed as the mean value across condition-specific stimulus presentations for each task. To keep the number of trials comparable across conditions and avoid a response induced by the shock, the analysis was restricted only to non-reinforced trials, i.e., CS+ trials that co-terminated with the presentation of the US were excluded[16–20]. To normalise the distribution of the SCR values for the group level analysis, they were square-root transformed[10,13].

To verify the results of the peak-scoring method of SCR, we additionally estimated the magnitude of the SCR with a dynamic causal modeling (DCM) of anticipatory SCR[15,21] implemented in the software package PsPM v4.2.1[22,23]. The raw time series were imported to the software and trimmed to the window starting from 10 s prior to the onset of the first

stimulus and terminating 16 s after the offset of the last stimulus in a task, i.e., either conditioning or inference test. The DCM analysis was run using a canonical skin conductance response function and inversion of 2 trials at the same time. To this end, the trimmed data were first filtered with a unidirectional 1st order Butterworth high pass filter with cut off frequency at 0.0159 Hz[15] and resampled to 10 Hz sampling rate which requires a low-pass filter cut off frequency of 5 Hz[24]. The resulting data time series was z-transformed for each participant to account for inter-individual differences in responsiveness. A forward model was specified to include, for each trial: (1) an anticipatory response within a 3.8 s time window between CS/ GS onset and potential US occurrence; and (2) an evoked response at 3.8 s after CS/ GS onset, i.e., at the time point of a potential US for which the response was estimated. Hence, the model was not informed about the condition type or whether a US was presented or not[15,21]. The resulting trial-by-trial estimates of the amplitude were sorted by condition and averaged, excluding the reinforced CS+ trials.

Pupil size response estimation using GLM

To verify whether the PSR estimation is not biased by the choice of the temporal window in the main method, we additionally estimated PSR with an independent method using GLM[25]. Preprocessed pupil size time series were z-scored by subtracting the mean and dividing by the standard deviation within each participant and phase (i.e., conditioning or inference test). The GLM consisted of two components (a) stimulus onset (impulse function) and (b) a sustained component during the stimulus duration (a boxcar function) that reflects the anticipatory sympathetic arousal during stimulus presentation. The onset of the US was indicated only by the impulse function. The boxcar regressor was normalised by dividing its height by the number of samples in that particular interval, such that this regressor had the same norm as the transient, impulse regressor. Each regressor was then convolved with a canonical pupil impulse response function:

$$size(t) = t^w e^{-t \cdot w / t_{max}}$$

where w is the width and t is the time-to-peak (ms) of the impulse response function. Values of the w and t_{max} parameters come from the previous reports introducing the method in the attentional blink [$w = 10.1$ and $t_{max} = 930$ ms;[25]]. The measured pupil time series and convolved regressors were baseline-corrected by subtracting, from each value in each time series, the average value from all pretrial baseline intervals [-0.5 to 0 s from stimulus onset;[25]]. The convolved and baseline-corrected regressors were horizontally concatenated into the complete design matrix. Multiple linear regression yielded the best-fitting beta weights for each regressor type (i.e., temporal component of the pupil response). The beta parameter for the boxcar regressor was used as a condition-specific PSR.

Supplementary Results

Skin conductance response (DCM)

To verify the results of the skin conductance responses (SCR) estimated with a peak-scoring method, we additionally estimated the magnitude of the SCR using a dynamic causal modeling (DCM) of anticipatory SCR[15,21]. The DCM analysis corroborated the results reported in the main text, albeit with weaker evidence obtained from a sample of 24 participants. The DCM method (Figure S2) revealed increased SCR to the CS+ than the CS- condition during the conditioning phase (CS+ = 1.55 [1.33, 1.78], CS- = 0.93 [0.77, 1.09]) with a large effect size (Hedges's g = 1.21). During the inference test, participants continued to exhibit differential SCR to the CS+ than the CS- condition as well as revealed a gradient of threat responses to GS1-GS4 as a function of their distance to the CS+, but with smaller magnitude to GS2 than GS3: CS+ = 1.68 [1.31, 2.06], GS1 = 1.24 [0.98, 1.52], GS2 = 0.97 [0.77, 1.17], GS3 = 1.03 [0.80, 1.27], GS4 = 0.89 [0.70, 1.08], CS- = 0.90 [0.70, 1.12] with the BF_{10} = 6.30 (estimate ± SE = -0.05 ± 0.02) and the effect size of marginal pseudo-R^2 = 0.17 and conditional pseudo-R^2 = 0.81.

Pupil size response (GLM)

To verify the results of the PSR obtained from baseline-corrected average, we additionally estimated PSR with an independent method using GLM[25]. The GLM method corroborated the results reported in the main text, showing similarly strong evidence for the gradient of threat response to GS1-GS4 stimuli, obtained from a sample of 27 participants (Figure S3). Conditioning phase: CS+ = 2.19 [1.95, 2.42], CS- = 1.67 [1.47, 1.87] (Hedges's g = 0.87). Phase including the inference test: CS+ = 2.45 [2.23, 2.68], GS1 = 1.97 [1.67, 2.28], GS2 = 1.65 [1.42, 1.88], GS3 = 1.58 [1.35, 1.80], GS4 = 1.49 [1.26, 1.74], CS- = 1.59 [1.38, 1.79] with the $\log BF_{10}$ = 6.51 (estimate ± SE = -0.07 ± 0.01) and the effect size of marginal pseudo-R^2 = 0.23 and conditional pseudo-R^2 = 0.81.

Post-conditioning relational memory

To explore whether the organization of the relational knowledge was affected by Pavlovian conditioning and the inference test, we inspected the representational scale values estimated from the MLDS in participants who exhibited differential conditioned threat responses in each response modality (shock expectancy, SCR, PSR) during the 2nd half of conditioning (Supplementary Figure S1). The visual inspection of the representational scale values in those participants revealed a highly similar profile of the results as in the main analysis, suggesting the linear and stable memory organization.

Non-linear gradients of threat responses in inference test

Second, we explored whether the gradients of the SCR and average event-related PSR (including subsequently PSR estimated with a GLM) to the GS1-GS4 stimuli could also be described with a non-linear (i.e., a quadratic or exponential) trend. The full-null model comparisons using approximated BF suggested that models with non-linear trends could also describe these gradients (Supplementary Figure S4). Yet, these models revealed smaller effect sizes than the models with the linear trend (e.g., in the case of PSR) and did not fully meet model assumptions (i.e., heteroscedasticity in residuals of the SCR model with an exponential trend), leaving the confirmation of (non-)linear trends for future research including a bigger sample.

Early vs. late trials during inference

Since the inference test based on a steady-state generalization test could be considered as discrimination learning, we also explored whether the gradients depended on early (first four) vs. late (last three) trials of the test. The data profiles of the three response modalities suggested that the gradient of shock expectancy was present mainly in the early trials while SCR and PSR data revealed mixed profiles (Supplementary Figure S5).

Individual differences

Finally, we explored whether individual differences in trait anxiety [STAI-T;[3]], intolerance for uncertainty [IUS;[6]], and childhood traumatic experience [CTQ;[4];[5]] were associated with the slope of the gradient. We did not find any evidence for such associations in any of the response modalities (BF_{10} = 0.008, 0.010, 0.012 for expectancy, SCR, and PSR, respectively).

291

Supplementary Figures

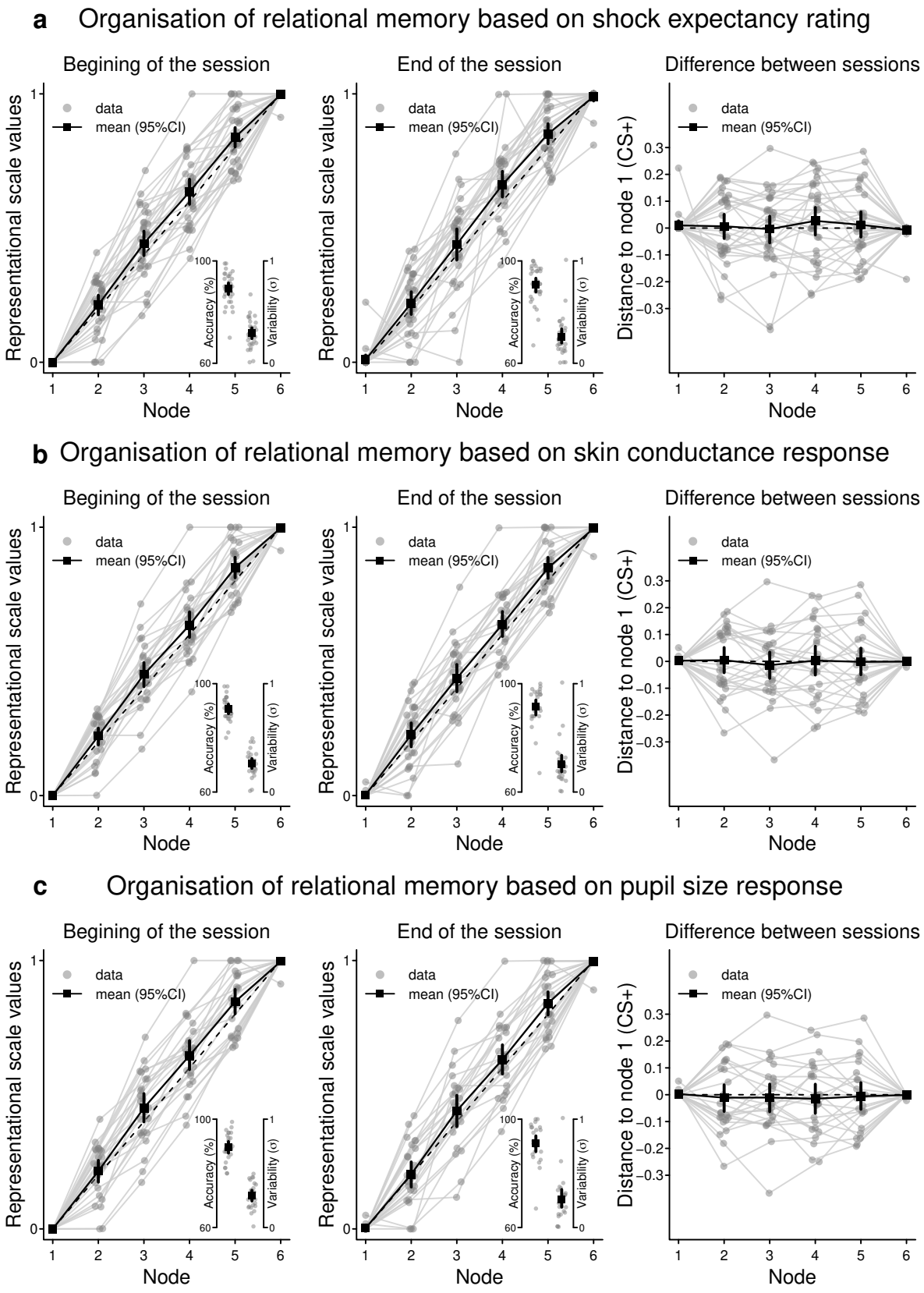

Figure S1. Organization of relational memory based on representational scale values estimated with the MLDS in participants who exhibited differential conditioned threat response during conditioning in each response modality.

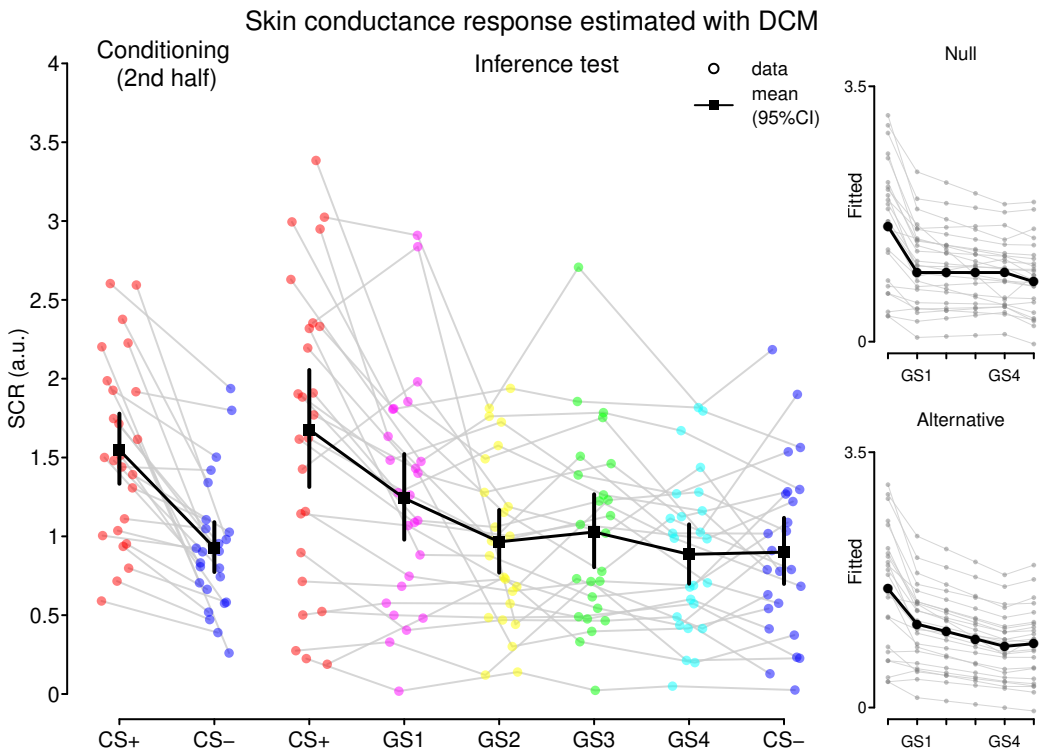

Figure S2. Results of conditioning and the test for the inferred risk of aversive outcome based on skin conductance response estimated with DCM. Corresponding subplots illustrate fitted values under the null and the alternative models (including stimulus identity as a control predictor) that were compared to test for the linear gradient of responses to GS1-GS4 stimuli.

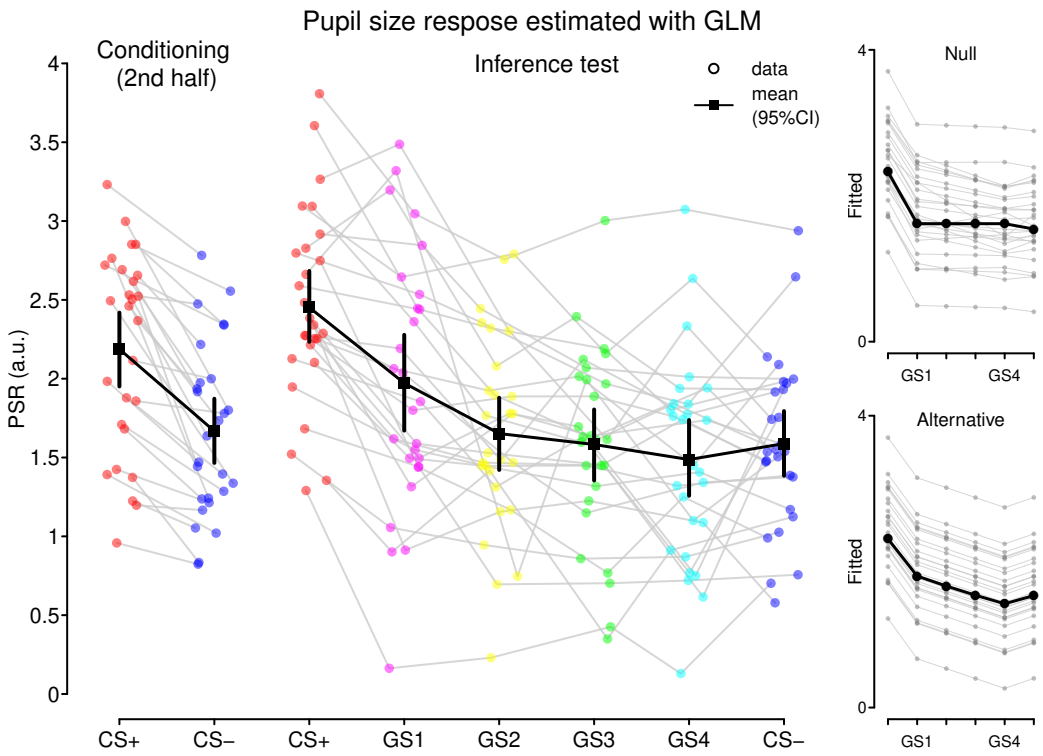

Figure S3. Results of conditioning and the test for the inferred risk of aversive outcome based on pupil size responses estimated with GLM. Corresponding subplots illustrate fitted values under the null and the alternative models (including stimulus identity as a control predictor) that were compared to test for the linear gradient of responses to GS1-GS4 stimuli.

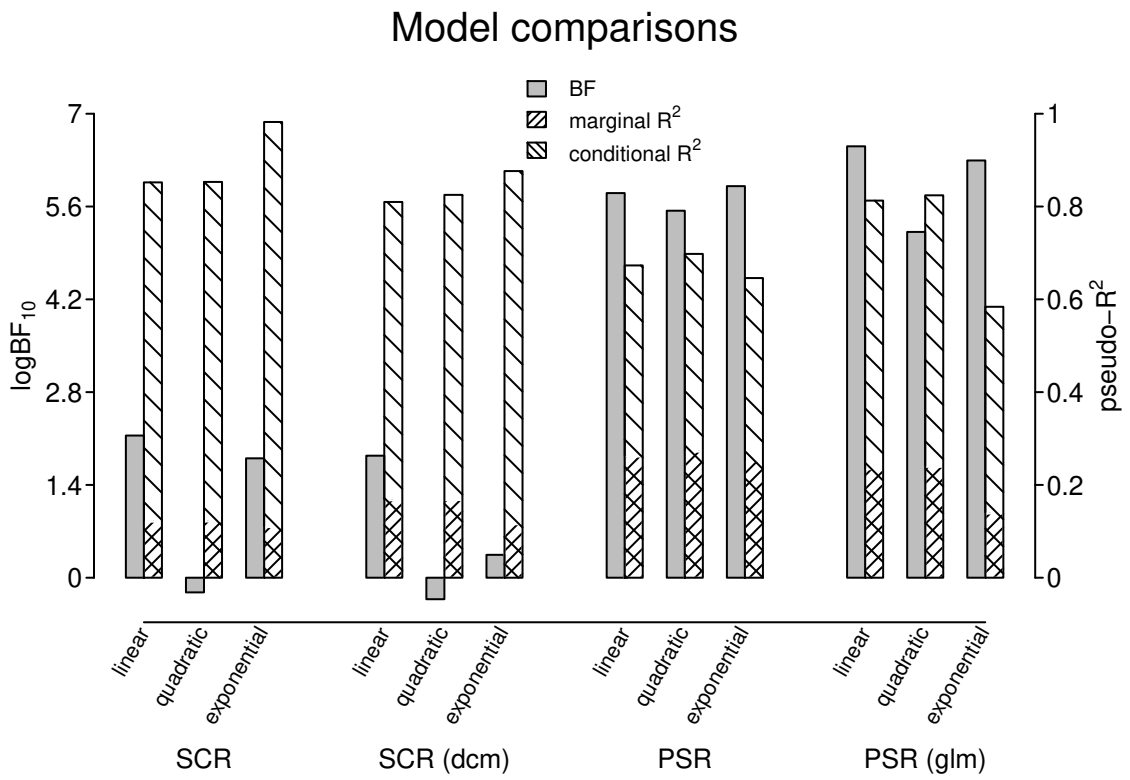

Figure S4. Full-null model comparisons with linear and non-linear (i.e., quadratic and exponential) trends for skin conductance response (SCR), skin conductance response (SCR) estimated with DCM, average event-related pupil size response (PSR), and event-related pupil size response estimated with GLM (PSR [glm]). Approximated Bayes Factor (BF) reflects the full-null model comparison for corresponding models while pseudo-R^2 (marginal and conditional pseudo-R^2) is a coefficient of determination calculated from the full model. Note: exponential model fitted to the SCR and SCR DCM revealed heteroscedasticity and long tails in a plot of residuals against fitted values, suggesting violations of model assumptions.

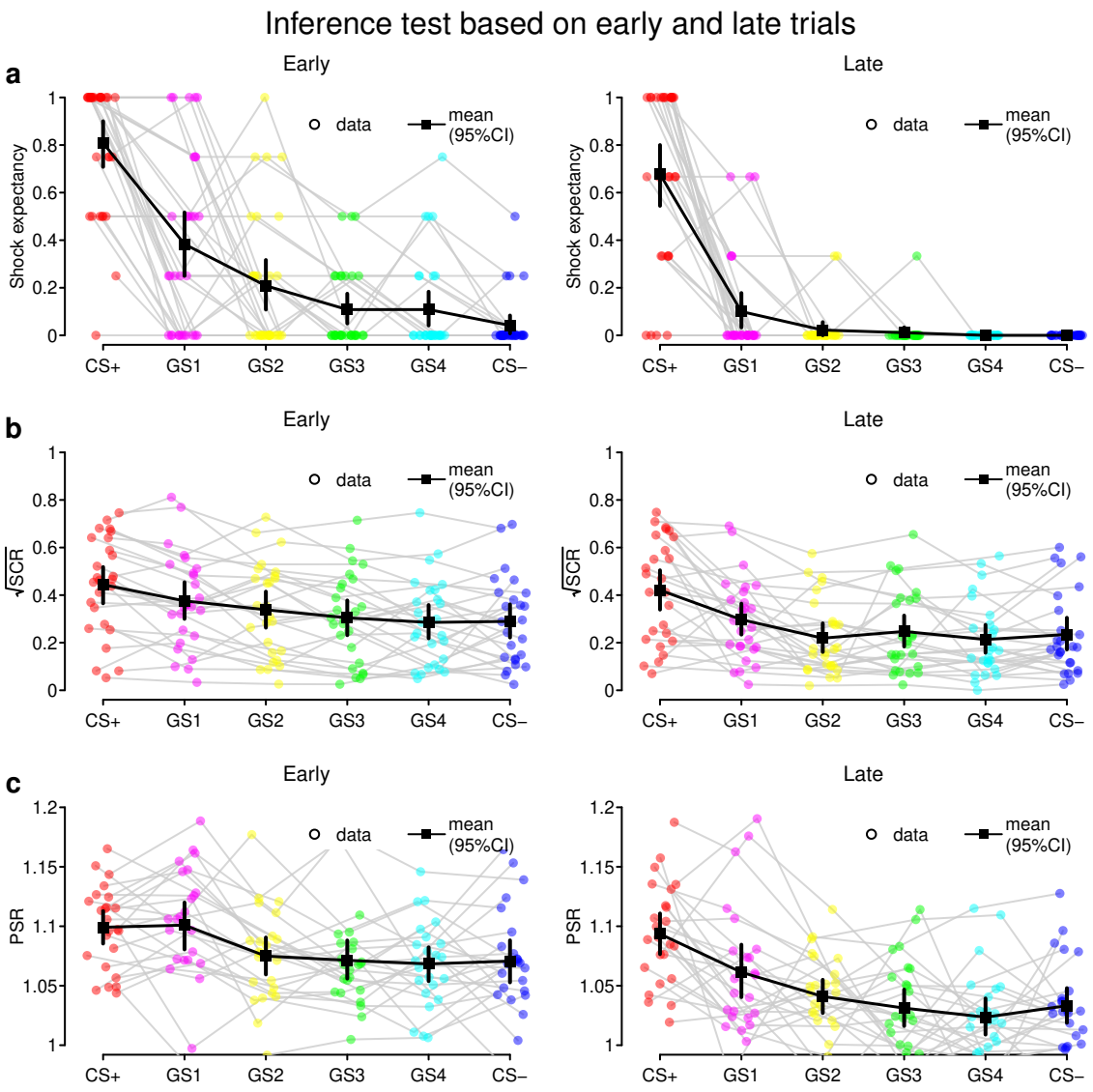

Figure S5. Results of the test for inferred risk of aversive outcome in each response modality split between early (first four) and late (last three) trials.

Supplementary Tables

Table S1
Data quality overview.

Participant id (age and sex)	Positive screening and informed consent obtained	Experiment completed (comments about the procedure)	Accuracy on memory task achieved[a]	Acceptable quality of skin conductance data[b]	Acceptable quality of pupil size and gaze data[c]	Differential conditioned response observed[d] – expectancy ratings	Differential conditioned response observed[d] – skin conductance	Differential conditioned response observed[d] – skin conductance (DCM)	Differential conditioned response observed[d] – pupil size (baseline corrected average[e])	Differential conditioned response observed[d] – pupil size (estimated with GLM[f])
1 (22f)	+	+	+	+	+	+	+	+	+	+
2 (23f)	+	+ (2)	+	+	- (2)	+	+	+	-	-
3 (21f)	+	+	+	+	+	+	-	-	+	+
4 (21m)	+	+ (3)	+	- (13)	-	+	- (20)	- (20)	- (20)	- (20)
5 (24f)	+	+	+	+ (14)	+ (vc)	+	+	+	+	+
6 (27m)	+	+ (4)	-	+	+ (vc)	+	+	+	+	+
7 (20f)	+	+	+	+	+	- (17)	+	-	+	+
8 (29f)	+	- (5,6)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
9 (20f)	+	+	+	+	+ (vc)	+	+	+	+	+
10 (20f)	+	+	+	+	+	+	-	-	+	-
11 (23f)	+	+ (7,8)	+	+	+ (vc)	+	+	+	-	-
12 (20f)	+	+	+	- (13)	+	-	-	+	-	-
13 (19f)	+	+ (9)	+	+	+	-	+	+	+	+
14 (23f)	+	+	+	+	+	+	-	-	-	+
15 (22m)	+	+ (8)	-	- (13)	+ (vc)	+	+	-	+	+
16 (23f)	+ (1a)	+	+	+	+	+	-	-	+	+
17 (21f)	+	+	+	+	+	+	+	+	+	+
18 (22m)	+ (1b)	+	+	+	+	+	+	+	+	+
19 (25m)	+	+	-	+	+	-	-	+	-	+
20 (22f)	+	+	+	+ (14)	+	+	+	- (21)	+	+

21 (22f)	+	+ (8)	+	+	+	+	+	-	+	+
22 (20f)	+	+	+	+	+	+	+	+	+	+
23 (22f)	+ (1c)	+	+	+	+	+	+	+	+	+
24 (21m)	+	+ (8)	+	+	+ (vc)	+	+	+	+	+
25 (20f)	+	+	+	+	+	+	+	+	+	+
26 (30f)	+	+ (10)	+	+	+	+	+	+	+	+
27 (25m)	+	+	+	+	+	+	+	+	-	+
28 (24f)	+	+	+	+	+	+	+	+	-	+
29 (21f)	+	+	-	+	+	+	+	+	+	+
30 (24m)	+	+	+	+	+ (vc)	- (18)	- (18)	- (18)	- (18)	- (18)
31 (20f)	+	+	+	+	+ (vc)	-	+	+	-	+
32 (20f)	+	+	+	+	+	+	+	+	+	+
33 (21m)	+	+	+	+	+ (15)	+	-	+	+	+
34 (34f)	+	+	+	+	+	+	+	+	+	+
35 (26f)	+	+	+	+	+ (vc)	+	-	+	+	+
36 (27m)	+	- (11)	+ (12)	N/A	N/A	N/A	N/A	N/A	N/A	N/A
37 (24f)	+	+	+	+	+	+	+	+	-	-
38 (24m)	+	- (5)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
39 (34m)	+	+	+	+	+ (16)	- (19)	- (19)	- (19)	- (19)	- (19)
40 (20f)	+	+	+	+	+	+	+	+	-	-
41 (23f)	+ (1d)	+	+	+	+ (vc)	+	+	+	+	+
42 (26f)	+	+	+	+	+	+	+	+	+	+
43 (28m)	+ (1e)	+	+	+	+	+	+	+	+	+
44 (25m)	+	+	-	+	+	+	+	-	+	+
Sample size (n)	44	41	36	38	39	34	30	29	28	32
					Final sample size (n)[g]	30	26	25	24	27

Note. **Abbreviations:** Male (m); Female (f); Uses vision correction but performed experiment without (vs); Inclusion (+); Exclusion (-); Not available (N/A). **Criteria and Definitions:** (a) Memory accuracy threshold: $\geq 80\%$ correct responses on directly learned associations before conditioning; (b) Data were visually inspected for recording system artifacts and electric shock responses; (c) Usable data: $< 35\%$ missing values and repeated increases to electric shocks (see Methods); (d) Differential conditioned threat response: Modality-specific difference between unreinforced CS+ and CS- > 0 (during second half of conditioning); (e) Pupil size response: Baseline-corrected event-related average (see Methods); (f) Pupil size estimated via gamma impulse response function using GLM (see Methods); (g) Final sample includes only participants meeting all inclusion criteria. **Participant-Specific Notes:** 1. Used regular medication for (a) hypothyroidism, (b) allergy, (c) skin acne, (d) inflammatory bowel disease, or (e) recent pain (diclofenac) within 72h prior to the experiment; 2. Kept one eye closed during generalisation (pupil data unavailable); 3. Had difficulty sitting still (hand/head movement); 4. Used a strategy for association learning task (day 1) to remember which of the two images *did not* go with the first image; 5. Did not attend the second session; 6. The paired association learning task restarted after few trials due to misunderstanding of instructions; 7. Generalisation task restarted due to eye-tracker failure but before the first trial; 8. Appeared drowsy during generalisation; 9. Used 3-point instead of 5-point calibration (5-point unsuccessful); 10. No movie shown during break (toilet break); 11. Refused further shocks during calibration, only completed memory test; 12. Achieved memory accuracy on the task prior to conditioning but did not complete the memory task at the end of the experiment; 13. Noisy signal made skin conductance estimation impossible or unreliable; 14. Showed lack of fluctuations but consistent responses to shocks and occasionally to CS+ trials (2nd half of conditioning and generalisation); 15. 35.8% missing data during conditioning but only 1.5% during generalisation (instructed to blink less); 16. Underwent eye surgery for vision correction; 17. Did not provide behavioral responses during conditioning; 18. in addition to the lack of differential conditioned responses, the participant misidentified the CS+ picture during debriefing (potential non-compliance to the task instructions); 19. Reported believing shocks were manually triggered by experimenter (potential non-compliance); 20. Showed differential conditioned threat responses but insufficient data quality; 21. DCM software produced NaN values during generalisation.

Table S2
Shock expectancy rating (full model)

Family:	binomial	logit			
Fixed effects					
Term			estimate	se	z
Intercept (GS mean)			-3.35e+00	4.52e-01	-7.40e+00
GS (gradient)*			-4.53e-01	1.14e-01	-3.99e+00
CS+			4.66e+00	5.40e-01	8.64e+00
CS-			-2.74e+00	1.93e+00	-1.42e+00
image2			1.01e+00	4.71e-01	2.13e+00
image3			1.17e+00	4.54e-01	2.57e+00
image4			-3.87e-01	5.35e-01	-7.23e-01
image5			4.32e-01	4.38e-01	9.87e-01
image6			7.68e-01	4.81e-01	1.59e+00
Random effects					
Subject	(30 levels)				
	intercept	GS			
intercept	1.52e+00	-2.23e-01			
GS	-2.23e-01	2.45e-01			
Subject	(30 levels)				
	CS+				
CS+	2.74e+00				
Subject	(30 levels)				
	CS-				
CS-	2.12e+01				
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	13	4.14e+02			
Full	14	3.99e+02	1.46e+01	1	1.36e-04

Note. * test predictor; random effects present var-cov matrix.

Table S3
Skin conductance response (full model)

Family:	gaussian	identity			
Fixed effects					
Term			estimate	se	z
Intercept (GS mean)			3.08e-01	3.16e-02	9.75e+00
GS (gradient)*			-1.26e-02	3.75e-03	-3.36e+00
CS+			1.39e-01	1.97e-02	7.03e+00
CS-			-2.50e-02	1.66e-02	-1.50e+00
image2			1.13e-02	2.12e-02	5.31e-01
image3			6.22e-03	2.07e-02	3.00e-01
image4			-1.99e-02	2.10e-02	-9.50e-01
image5			-7.15e-03	2.13e-02	-3.36e-01
image6			-1.29e-02	2.14e-02	-6.01e-01
Random effects					
Subject	(26 levels)				
	intercept				
intercept	2.05e-02				
Subject	(26 levels)				
	GS				
GS	1.34e-04				
Subject	(26 levels)				
	CS+				
CS+	4.2e-03				
Subject	(26 levels)				
	CS-				
CS-	1.25e-03				
Residual var					4.42e-03
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	13	-2.73e+02			
Full	14	-2.83e+02	9.34e+00	1	2.24e-03

Note. * test predictor; random effects present var-cov matrix.

Table S4
Skin conductance response estimated via DCM (full model)

Family:	gaussian	identity			
Fixed effects					
Term			estimate	se	z
Intercept (GS mean)			9.91e-01	1.20e-01	8.28e+00
GS (gradient)*			-5.07e-02	1.56e-02	-3.26e+00
CS+			6.38e-01	1.18e-01	5.38e+00
CS-			-1.16e-01	7.09e-02	-1.63e+00
image2			1.48e-01	9.41e-02	1.58e+00
image3			1.02e-02	9.18e-02	1.11e-01
image4			4.70e-02	9.53e-02	4.93e-01
image5			2.81e-02	9.94e-02	2.82e-01
image6			-4.93e-03	9.45e-02	-5.21e-02
Random effects					
Subject	(25 levels)				
	intercept	GS			
intercept	2.52e-01	-7.7e-03			
GS	-7.7e-03	1.47e-03			
Subject	(25 levels)				
	CS+				
CS+	2.35e-01				
Subject	(25 levels)				
	CS-				
CS-	3.62e-03				
Residual var					8.77e-02
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	14	1.78e+02			
Full	15	1.7e+02	8.69e+00	1	3.2e-03

Note. * test predictor; random effects present var-cov matrix.

Table S5
Pupil size response (full model)

Family:	gaussian	identity			
Fixed effects					
Term			estimate	se	z
Intercept (GS mean)			1.06e+00	6.59e-03	1.61e+02
GS (gradient)*			-4.74e-03	9.83e-04	-4.82e+00
CS+			3.09e-02	5.00e-03	6.18e+00
CS-			-8.52e-03	4.93e-03	-1.73e+00
image2			1.56e-02	6.15e-03	2.54e+00
image3			2.17e-04	6.10e-03	3.55e-02
image4			2.45e-03	6.13e-03	4.00e-01
image5			-1.22e-02	6.20e-03	-1.97e+00
image6			1.03e-03	6.21e-03	1.66e-01
Random effects					
Subject	(24 levels)				
	intercept				
intercept	5.61e-04				
Subject	(24 levels)				
	GS				
GS	9.66e-12				
Residual var					4.45e-04
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	11	-6.35e+02			
Full	12	-6.51e+02	1.66e+01	1	4.67e-05

Note. * test predictor; random effects present var-cov matrix.

Table S6
Pupil size response estimated via GLM (full model)

Family:	gaussian	identity			
Fixed effects					
Term			estimate	se	z
Intercept (GS mean)			1.70e+00	1.23e-01	1.38e+01
GS (gradient)*			-6.77e-02	1.37e-02	-4.96e+00
CS+			7.54e-01	6.93e-02	1.09e+01
CS-			-6.56e-02	6.83e-02	-9.61e-01
image2			1.84e-01	8.61e-02	2.13e+00
image3			-5.53e-02	8.54e-02	-6.48e-01
image4			2.36e-02	8.58e-02	2.75e-01
image5			-1.84e-01	8.59e-02	-2.15e+00
image6			-1.10e-01	8.65e-02	-1.27e+00
Random effects					
Subject	(27 levels)				
	intercept				
intercept	3.02e-01				
Subject	(27 levels)				
	GS				
GS	7.21e-11				
Subject	(27 levels)				
	CS+				
CS+	2.98e-11				
Subject	(27 levels)				
	CS-				
CS-	2.23e-12				
Residual var					9.75e-02
Model comparison					
Model	npar	Deviance	χ^2	Df	p
Reduced	13	1.81e+02			
Full	14	1.63e+02	1.81e+01	1	2.09e-05

Note. * test predictor; random effects present var-cov matrix.

Deviations from the pre-registration

Sample size and stopping rule

In the pre-registration, we aimed to employ a modified Sequential Bayes Factor design with maximal sample size[26]. The initial sample size based on previous reports[10,13,16,27,28] was set to n=24 and the maximum sample size to n=36 participants meeting all inclusion criteria and exhibiting usable data for each response modality. After having collected data from approximately 28 participants, the inclusion rate was lower than expected – about 50-65%. To obtain the minimum sample size of n=24, we would have to collect data from about 40 participants. Similarly, the maximum sample size of n=36 was foreseen to require about 70 participants. Due to the limited time frame for data collection, which was constrained by the duration of the first author's guest research stay, this became unfeasible. As a consequence, we decided to stop the data collection after obtaining the pre-registered initial sample size of n=24 participants with usable data in at least one measure of peripheral physiology, i.e. pupil size or skin conductance response. As a result, our sampling plan became fixed and suitable for frequentist inference, which we present in the supplementary material, alongside the inference based on the approximated Bayes Factor reported in the main text. It is worth to highlight, however, that the original plan for the Bayes Factor design was implemented because the data from the (baseline-corrected) pupil size response showed the evidence for the gradient of threat responses to GS1-GS4 stimuli exceeding the BF of 10 after having obtained the data from the initial sample size of n=24.

Participation eligibility

Five volunteers who used medication (e.g., anti-inflammatory or anti-allergic drugs) and/or reported moderate alcohol consumption (e.g., two or three glasses) 72h before the experiment, were invited to participate in the study, given the lower inclusion rate. Given the pre-defined within-subjects inclusion criteria of relational and Pavlovian threat memory acquisition as well as individual calibration of the shock intensity, we do not expect this to have any impact on data quality or results. None of the participants used psychoactive drugs.

Analysis: Skin conductance response estimation with GLM and canonical SCR function

The initial analysis plan included additionally an alternative method for the estimation of SCR using a GLM with canonical SCR function. The GLM approach assumes that the sympathetic input is short and occurs at constant latency[29]. This assumption is not met in threat (fear) conditioning tasks[15,21,23]. Therefore,we performed an analysis using a dynamic causal modeling (DCM) of anticipatory skin conductance response, which is recommended over the GLM analysis[15,21].

For transparency, we report here the results obtained with the original plan for the SCR estimation using the GLM method. The raw time series were imported to the software and trimmed to the window starting from 10 s prior to the onset of the first stimulus and terminating 16 s after the offset of the last stimulus in a task, i.e., either conditioning or inference. For the GLM analysis, the trimmed data were filtered with a unidirectional 1st order Butterworth high pass filter with cut off frequency 0.05 Hz[29] and resampled to 10 Hz sampling rate which requires a low-pass filter cut off frequency of 5 Hz[24]. The resulting data time series was z-transformed for each participant to account for inter-individual differences in responsiveness. To construct a design matrix, each and every condition in the respective experimental phase (conditioning or inference) as well as the US was modeled with a Dirac delta function centered on the event onset, convolved with a canonical skin conductance response function and its first derivative[24]. Next, the magnitude of the condition-specific response was taken from the estimated parameter for the corresponding experimental condition.

This analysis corroborated the results reported in the main analysis, albeit with weaker evidence. Twenty-four participants revealed increased SCR to the CS+ than the CS- condition during the conditioning phase (CS+ = 1.19 [0.77, 1.63], CS- = 0.29 [0.11, 0.49]) with a large effect size (Hedges's g = 1.03). During the inference test, participants continued to exhibit differential SCR to the CS+ than the CS- condition (CS+ = 1.59 [0.99, 2.22], CS- = 0.43 [0.20, 0.70]). They also exhibited gradient of threat responses to

GS1-GS4 as a function of their distance to the CS+, but with smaller magnitude to GS2 than GS3 (GS1 = 0.93 [0.50, 1.46], GS2 = 0.46 [0.22, 0.75], GS3 = 0.58 [0.28, 0.93], GS4 = 0.38 [0.15, 0.65]). LMM fitted to the magnitude of the SCR revealed weak evidence for the linear gradient (estimate ± SE = -0.08 ± 0.03, BF_{10} = 2.98) and good fit to the data (marginal pseudo-R^2 = 0.18 and conditional pseudo-R^2 = 0.78).

Analysis: Pupil size response estimation with GLM

To verify the results of the pupil size response (PSR) estimated with event-related averaging, the initial plan included an additional estimation of the PSR with GLM based on a single impulse function at the stimulus onset that is convolved with a canonical pupil size response function[25]. While the method accounts for the initial response to a stimulus onset (i.e., a parasympathetically regulated pupil constriction), it does not account for a sustained component of the response that reflects the anticipatory sympathetic arousal, i.e., conditioned threat response. Therefore, we modified the GLM to include not one but two transient events: (1) stimulus onset (impulse function) and (2) a sustained component during the stimulus duration (a boxcar function), the estimation of which was used in the further analyses as a pupil size response of interest. The analysis was based on a previous report where both components were estimated[25]. In brief, the boxcar regressor was normalised by dividing the height of the boxcar by the number of samples in that particular interval, such that this regressor had the same norm as the transient, impulse regressor. The measured pupil time series and convolved regressors were baseline-corrected by subtracting, from each value in each time series, the average value from all pre-trial baseline intervals (-0.5 to 0 s from stimulus onset). The convolved and baseline-corrected regressors were horizontally concatenated into the complete design matrix. Multiple linear regression yielded the best-fitting beta weights for each regressor type (i.e., temporal component of the pupil response).

Analysis: Testing acquisition of relational memory

The initial analysis plan included a test for stability of the memory organization over the course of the session on day 2. To this end, we planned to use repeated measures

ANOVA on a difference in representational scale values between the beginning and the end of the session. This, however, tests primarly for the *change* in the memory rather than its *stability*. To mitigate this drawback, we performed an alternative analysis that quantified an effect of the graph node on the representational scale values *nested within a two-level factor of time*. Additionally, given that the data values were restricted to the interval [0,1], we used a Generalized Linear Mixed Models (GLMM) with *logit* link function and beta error structure as an appropirate analysis strategy.

Analysis: Testing acquisition of Pavlovian threat memory

The initial analysis plan included a test for the acquisition of Pavlovian threat memory, i.e., statistical inference on data from the conditioning phase. Since we included in the analysis only those participants who exhibited differential conditioned response (i.e., numerical difference between CS+ and CS- is higher than 0), which was included in the pre-registered inclusion criteria, any inference test here would be intrinsically biased. Therefore we did not perform any statistical inference on conditioning data. Instead, we reported respective effect sizes to quantify the magnitude of the conditioning effect in the selected participants.

Analysis: Testing Pavlovian threat generalization/ inferential expression of Pavlovian threat memory

The initial analysis plan considered simple linear regression to test for linear decrease in the means of the GS1-GS4. This strategy, however, is not appropriate for a repeated-measure design, i.e. when data is clustered within an individual, because it assumes independence between data points. Therefore, we decided to test the same hypothesis with an appropirate approach using (Generalized) Linear Mixed Models (G/LMM). G/LMMs treat dependent data as clustered within an individual that is considered as a random effect, and are recommended in the analysis of generalization gradients[30]. Moreover, their generalized form gave us flexibility to incorporate data that comes from other than normal distribution, i.e., shock expectancy ratings.

426 **Analysis: Follow-up analyses**

427 The initial analysis plan considered the possibility to re-run the analyses where GS
428 stimuli are ordered according to their arrangement recovered from the MLDS to adjust for
429 the participant-specific distortions of the memory graph. There were only three participants
430 in the sample who showed a substantially distorted memory graph. Given this and the fact
431 that the originally planned analyses worked well, we did not perform this follow-up analysis
432 due to its redundancy.

References

1. Schlichting, M. L., Mumford, J. & Preston, A. Learning-related representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications* **6**, 8151 (2015).

2. Brainard, D. H. The Psychophysics Toolbox. *Spatial Vision* **10**, 433–436 (1997).

3. Spielberger, C. D., Gorsuch, R. L., Lushene, R., Vagg, P. R. & Jacobs, G. A. *Manual for the State-Trait Anxiety Inventory*. (Consulting Psychologists Press, 1983).

4. Bernstein, D. P. *et al.* Initial reliability and validity of a new retrospective measure of child abuse and neglect. *The American Journal of Psychiatry* **151**, 1132–1136 (1994).

5. Bernstein, D. P. *et al.* Development and validation of a brief screening version of the Childhood Trauma Questionnaire. *Child Abuse & Neglect* **27**, 169–190 (2003).

6. Buhr, K. & Dugas, M. J. The Intolerance of Uncertainty Scale: Psychometric properties of the English version. *Behaviour Research and Therapy* **40**, 931–945 (2002).

7. Berkman, L. F. & Syme, S. L. Social networks, host resistance, and mortality: A nine-year follow-up study of Alameda County residents. *American Journal of Epidemiology* **109**, 186–204 (1979).

8. Vanderwal, T., Kelly, C., Eilbott, J., Mayes, L. C. & Castellanos, F. X. Inscapes: A movie paradigm to improve compliance in functional magnetic resonance imaging. *NeuroImage* **122**, 222–232 (2015).

9. Smithson, M. & Verkuilen, J. A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods* **11**, 54–71 (2006).

10. Dunsmoor, J. E., Otto, A. R. & Phelps, E. A. Stress promotes generalization of older but not recent threat memories. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 9218–9223 (2017).

11. Goulet-Pelletier, J.-C. & Cousineau, D. A review of effect sizes and their confidence intervals, Part I: The Cohen's d family. *The Quantitative Methods for Psychology* **14**, 242–265 (2018).

12. Barr, D. J., Levy, R., Scheepers, C. & Tily, H. J. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* **68**, 10.1016/j.jml.2012.11.001 (2013).

13. Dunsmoor, J. E., Kroes, M. C. W., Braren, S. H. & Phelps, E. A. Threat intensity widens fear generalization gradients. *Behavioral Neuroscience* **131**, 168–175 (2017).

14. Kroes, M. C. W., Dunsmoor, J. E., Lin, Q., Evans, M. & Phelps, E. A. A reminder before extinction strengthens episodic memory via reconsolidation but fails to disrupt generalized threat responses. *Scientific Reports* **7**, 1–14 (2017).

15. Staib, M., Castegnetti, G. & Bach, D. R. Optimising a model-based approach to inferring fear learning from skin conductance responses. *Journal of Neuroscience Methods* **255**, 131–138 (2015).

16. Dunsmoor, J. E., Martin, A. & LaBar, K. S. Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology* **89**, 300–305 (2012).

17. Klumpers, F., Morgan, B., Terburg, D., Stein, D. J. & van Honk, J. Impaired acquisition of classically conditioned fear-potentiated startle reflexes in humans with focal bilateral basolateral amygdala damage. *Social Cognitive and Affective Neuroscience* **10**, 1161–1168 (2015).

18. Kroes, M. C. W. *et al.* How Administration of the Beta-Blocker Propranolol Before Extinction can Prevent the Return of Fear. *Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology* **41**, 1569–1578 (2016).

19. Milad, M. R. *et al.* Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry* **62**, 446–454 (2007).

20. Schiller, D., Levy, I., Niv, Y., LeDoux, J. E. & Phelps, E. A. From fear to safety and back: Reversal of fear in the human brain. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* **28**, 11517–11525 (2008).

21. Bach, D. R., Daunizeau, J., Friston, K. J. & Dolan, R. J. Dynamic causal modelling of anticipatory skin conductance responses. *Biological Psychology* **85**, 163–170 (2010).

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

23. Bach, D. R. *et al.* Psychophysiological modeling: Current state and future directions. *Psychophysiology* **55**, e13214 (2018).

24. Bach, D. R. *et al.* Human hippocampus arbitrates approach-avoidance conflict. *Current biology: CB* **24**, 541–547 (2014).

25. de Gee, J. W., Knapen, T. & Donner, T. H. Decision-related pupil dilation reflects upcoming choice and individual bias. *Proceedings of the National Academy of Sciences of the United States of America* **111**, E618–625 (2014).

26. Schönbrodt, F. D. & Wagenmakers, E.-J. Bayes factor design analysis: Planning for compelling evidence. *Psychonomic Bulletin & Review* **25**, 128–142 (2018).

27. Dunsmoor, J. E., Mitroff, S. R. & LaBar, K. S. Generalization of conditioned fear along a dimension of increasing fear intensity. *Learning & Memory* **16**, 460–469 (2009).

28. Dunsmoor, J. E. *et al.* Event segmentation protects emotional memories from competing experiences encoded close in time. *Nature Human Behaviour* **2**, 291–299 (2018).

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

30. Vanbrabant, K. *et al.* A new approach for modeling generalization gradients: A case for hierarchical models. *Frontiers in Psychology* **6**, 652 (2015).