Pre-existing knowledge of environmental structure guides inferential expression of Pavlovian threat memory

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

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

[2] Wilhelm Wundt Institute for Psychology, Leipzig University, Leipzig, Germany

[3] IMPRS NeuroCom, Leipzig, Germany

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

Author Note

This manuscript is a preprint and has not been peer reviewed. Blazej M. Baczkowski is now at the Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany. Correspondence concerning this article should be addressed to Blazej M. Baczkowski, blazej.baczkowski@uni-tuebingen.de, University of Tübingen.

Abstract

To protect from danger, individuals learn to recognise threats from exposure to aversive outcomes, as described by Pavlovian threat conditioning. However, Pavlovian threat memory alone is insufficient to gauge danger when previously encountered threats are absent. Here, we show that, to infer risk, individuals draw on their pre-existing knowledge of environmental relations among neutral cues built across multiple episodic experiences in safe contexts. We trained participants to associate pairs of novel images that together formed an abstract linear knowledge structure. The next day, the image anchored at one end of the structure acquired Pavlovian threat memory. When later exposed to the remaining images under the risk of a mild electric shock, participants showed defensive responses, both in self-report and physiological measures, that scaled with the associative distance between the neutral and threatening cues. These findings suggest that map-like relational memories can flexibly guide our defensive system, leveraging minimial aversive experiences.

Keywords: Pavlovian conditioning | fear | associative learning | relational memory | inference

Word count: intro: 720, methods: 1820, results: 1560, discussion: 1270

Pre-existing knowledge of environmental structure guides inferential expression of Pavlovian threat memory

Recognising dangerous situations is crucial for regulating defensive behaviours that are tailored to the imminence of potential harm[1,2]. Past life-threatening situations that have been overcome teach us about what might pose a risk of harm in the future. However, when learned or similar threats are not immediately perceptible in the current environment, the risk of harm, even if remote, may still exist. In such cases, it must be inferred rather than directly derived from a past aversive experience. How do individuals assess risk in such circumstances?

Learning about threats from a first-hand exposure to danger is well described by Pavlovian threat conditioning[3]. When a biologically neutral cue becomes associated with an unexpected aversive outcome through the engagment of the amygdala, Pavlovian threat memory is formed[4]. Later exposure to this cue activates the memory and elicits a set of (species-specific) behavioural and physiological responses aimed to cope with the anticipated harm[5]. This adaptive mechanism protects us from future threats, guiding defense in response to familiar cues, as well as those that resemble them based on perceptual or conceptual features[6,7]. However, it is ineffective in assessing risk on its own when such cues are not immediately perceptible in the environment.

To expand our defensive capacity and prepare for potential harm beyond what was learned from aversive experiences, Pavlovian threat memory could interact with other types of memories[8]. While aversive experiences are relatively rare, and thus their memories are sparse, we constantly form memories of relationships between neutral cues as we navigate the world around us. These episodic relational memories, likely supported by the hippocampal-entorhinal system[9], encode both immediate (adjacent) and multi-step (long-range) contingencies between environmental states, helping to predict what is likely to co-occur in the near and distant future[10,11]. The structure of our environment is reflected in how these memories are organized: events that frequently co-occur become strongly related and are represented ‘close to each other,’ so that recognizing one can accurately inform us

about the likelihood of the other[12,13]. Moreover, since all environmental contingencies are rarely present in a single learning episode, memories of related experiences – acquired over time and across different contexts – must be linked to form a complex network of stable relationships among objects, people, and places encountered in various circumstances[14,15]. Equipped with such knowledge, individuals can flexibly adjust their goal-oriented behavior, as it allows them to infer the outcomes of actions based on relationships between events that were never directly observed together[16–18]. Therefore, episodic-like relational knowledge, whose structure reflects stable environmental patterns, is a strong candidate for guiding our defensive system by building upon – and significantly expanding – the Pavlovian cue-outcome associations formed through first-hand aversive experiences.

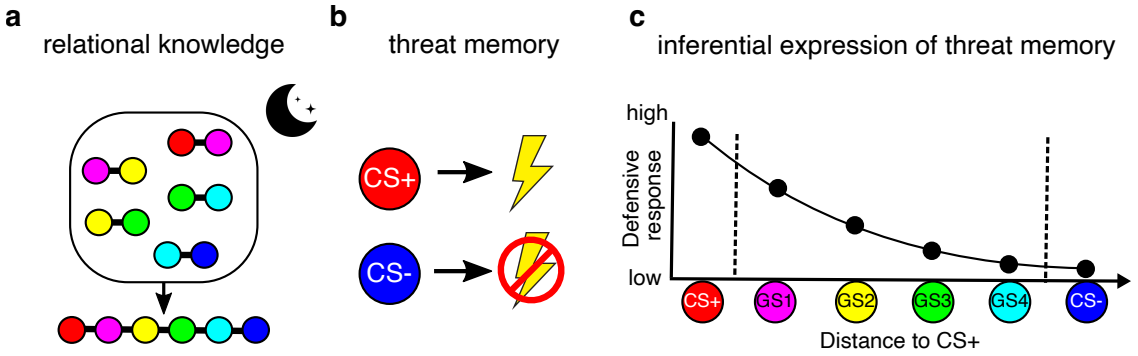

Figure 1. Study rationale. **a)** Distinct learning episodes of encoding overlapping pairs of neutral sensory cues can be integrated over night to form a stable relational knowledge characterised by an abstract linear graph of six nodes (sensory cues) and connecting edges (associations). **b)** Through a subsequent learning episode of Pavlovian threat conditioning, two sensory cues allocated to the opposite ends of the graph (CS+ & CS-) are differentially associated with an aversive outcome (illustrated as a lightning bolt). **c)** If individuals use the structure of pre-existing relational knowledge built across multiple episodes together with Pavlovian threat memory to infer the possibility of an aversive outcome, the magnitude of defensive behaviour in response to the cues that were never a part of conditioning (GS1-GS4) is expected to be a function of their relational distance to the cue previously associated with the aversive outcome.

In this pre-registered study, we investigated whether individuals combine Pavlovian threat memory with the structure of pre-existing relational knowledge, built across multiple episodes, to infer potential harm beyond the learned threat. We hypothesized that risk can be inferred from cues never directly experienced in an aversive context, with its level scaled according to the relational distance between these cues and the one directly associated with an aversive outcome (Figure 1). To this end, we used an experimental procedure across two consecutive days that combined elements of associative learning and inferential reasoning with Pavlovian cue conditioning and generalisation (Figure 2). Specifically, to form relational memory, participants learned multiple pairs of neutral images through trial and error. These

pairs shared images, allowing participants to later recognize that the overlapping associations followed an abstract linear graph of six nodes, which was not apparent from the feature similarity of the images or their temporal order during presentation. The memory training was followed by a one-day break to help integrate the overlapping associations into a complete six-element linear knowledge structure[19]. Next day, two images from opposite ends of the linear graph became differentially predictive of a mild electric shock in a Pavlovian conditioning procedure, such that only one of them was associated with the aversive outcome. To test whether participants infer the possibility of the aversive outcome based on the combined linear structure of pre-existing relational memory and Pavlovian threat memory, they were then exposed to the remaining images under the risk of receiving a shock, while their defensive responses were measured in three modalities: shock expectancy ratings, skin conductance, and pupil size changes.

Methods

The hypotheses, methods, and analysis plan were preregistered [LINK] prior to data collection, with deviations reported in the Supplementary Material.

Participants

The study was approved by the local ethical review board (CMO for the Arnhem-Nijmegen region) and conducted in accordance with the Declaration of Helsinki. The eligibity and exclusion critera are detailed in the Supplementary Material. In brief, we recruited healthy volunteers aged 18–35 from a student population in Nijmegen, the Netherlands, using convenience sampling. Forty-four volunteers provided informed consent, and 41 participants (aged 19–34, mode=20 years, 29 females) completed the study. Applying our pre-registered exclusion criteria to ensure data quality, the final sample sizes were n=30 (24 females) for shock expectancy ratings, n=26 (22 females) for skin conductance, and n=24 (20 females) for pupil size, all meeting our pre-registered minimum requirements based on previous studies[20,21].

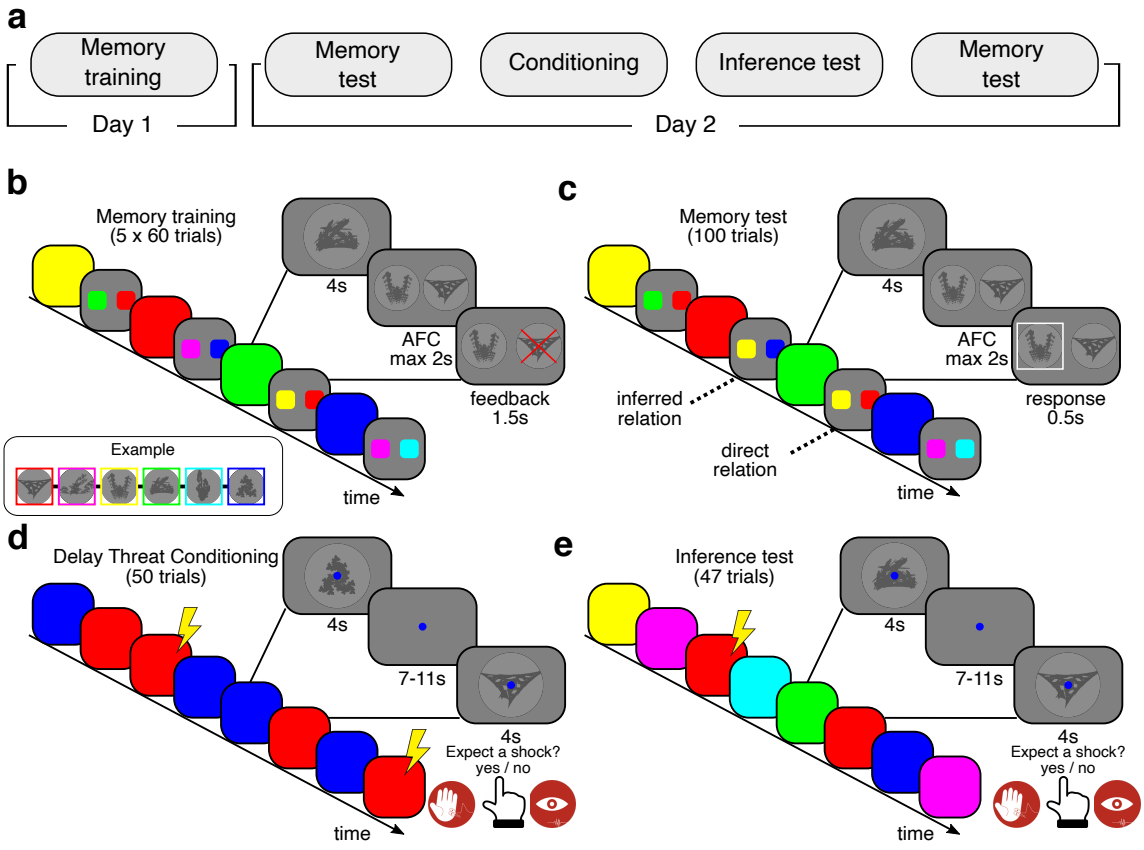

Figure 2. Procedure and behavioural tasks. **a)** Four experimental tasks were performed sequentially on two consecutive days. **b)** On a first day, participants learned to associate pairs of fractal-like images via trial and error. Each trial started with a cue image followed by a presentation of two other images, one of which was an associate of the cue. Participants chose one of two images and received feedback on accuracy. Unbeknown to participants, the learned pairs shared images between one another such that they followed an abstract linear graph of six nodes. **c)** The next day, participants completed a two-alternative forced-choice task to test their memory for image pairs and their reasoning of novel relations. Participants were shown a target image and asked to choose which of two other images they believed was most closely associated with it. The trial presentation followed the same scheme as the learning task, except that no feedback was provided, and some forced-choice trials included images not previously paired with the target. **d)** Next, two familiar images from opposite ends of the graph were used in a Pavlovian delay conditioning task, with one occasionally co-terminating with a mild shock. We measured defensive reactions in preparation for the anticipated shock in three modalities (shock expectancy ratings, skin conductance, and pupil size changes). **e)** Finally, to test whether participants infer the risk of an aversive outcome based on pre-existing relational knowledge, they were exposed to all six familiar images under the risk of receiving a shock, and their defensive responses were measured as during conditioning.

Stimuli

To minimize perceptual similarity and avoid prior associations, six custom-made fractal-like images were generated (http://www.apophysis.org/). These grayscale, isoluminant, circle-shaped images, were presented at a 20° diameter on a grey background (RGB: [128/255 128/255 128/255]).

Behavioural tasks

Paired associate learning (PAL) task. Participants acquired relational memory by learning image pairs through trial and error across 300 trials, split into five blocks. Image pairs were selected from a set of the six images, which were randomly assigned to the nodes of the linear memory graph. Each trial began with a 4-s cue image, followed by two other images (left/right counterbalanced), one of which was a direct neighbour of the cue. Participants had 2 s to select the correct associate, receiving 1.5 s of visual feedback on accuracy or a blank screen. The inter-stimulus interval (ISI) was 0.5 s, and the inter-trial interval (ITI) was 1 s. Trials were pseudorandomised so that no more than two consecutive trials could start with the same cue image.

Two-alternative forced choice (2-AFC) task. To assess the memory structure of previously learned associations, participants were presented with a cue imaged followed by two other options and asked to select the option most closely associated with the cue. Trials followed the same format and timing as the PAL task but without accuracy feedback. Unanswered trials were repeated at the end of the task, with repetitions ranging from 0 to 10 (mode = 0), and a maximum of 0 to 2 repetitions per trial (mode = 1). The task consisted of 100 trials, including 76 assessing direct neighbor relations (e.g., node1–node2). Accuracy on these trials (minimum 80%) served as a threshold for successful relational memory acquisition and retrieval[22]. Additionally, the task evaluated inferred associations relevant to nodes that would later serve as CS+/CS- (e.g., node1–node3), akin to paired-associate inference paradigms[23].

Pavlovian delay threat conditioning. Two images, previously assigned to opposite ends of the memory graph, served as CS+ and CS− in a Pavlovian delay threat conditioning task. The task included 20 presentations of each image, along with additional 10 CS+ trials that co-terminated with an unconditioned stimulus (US) – a mild electric shock – delivered 200 ms before stimulus offset and lasting until its end. Stimulus order was pseudorandomized to prevent more than two consecutive repetitions of the same condition. Each stimulus was displayed for 4 seconds, followed by a variable inter-trial interval (ITI) of 7, 9, or 11 seconds (mean = 9 s), during which a colored fixation dot (0.7° diameter, RGB:

[68/255 68/255 248/255]) appeared. A 12-second break was provided midway through the task. On each trial, participants indicated whether they expected a shock (yes/no) and were explicitly instructed that their responses had no effect on the outcome, reducing the chance of attributing the result to their choice or reaction time.

Test for the expression of Pavlovian threat memory (Inference test). To test for the expression of Pavlovian threat memory to non-conditioned stimuli, participants were presented with the CS+, CS- and four other images that belonged to the memory graph (GS1-GS4). Stimulus presentation followed the same procedure and timing as the conditioning task, with pseudorandomized trials to rule out temporal order effects. Each stimulus appeared seven times, with five additional CS+ trials co-terminating with the shock to prevent extinction and habituation (steady-state generalization testing)[20,24]. Participants rated shock expectancy (yes/no) for each image, but were again instructed that their button presses did not affect the outcome of a trial.

Measures

Representational scale values of relational memory. To recover the arrangement of memory items on the linear graph, we applied the maximum likelihood difference scaling (MLDS) approach, originally used to perceptual rather than representational space[25]. Binary responses from the 2-AFC task were treated as interval comparisons between memory items, subject to additive Gaussian noise (ϵ) with equal variance (σ). In each trial, the decision variable, D, was defined as the difference in absolute representational scale distances plus noise: $|\phi_a - \phi_b| - |\phi_a - \phi_c| + \epsilon$. If $D > 0$, the first pair was chosen; otherwise, the second. If two pairs were very similar, D approached 0, the decision depended on the additive random noise. The likelihood of selecting the first pair was given by $P(D > 0 \mid \epsilon, \sigma)$. The six-image arrangement was modeled by optimizing representational space values to maximize the likelihood. Values were normalized to [0,1], and optimization was performed using Matlab 2016a's *fminsearch* function to find representational scale values (ϕ) and variance (σ)[7].

Skin conductance response. Skin conductance time series was recorded using a BrainVision BrainAmp ExG system (Brain Products GmbH, Germany) with Ag/AgCl electrodes at a 500 Hz sampling rate. Data quality was visually inspected for artifacts indicating recording malfunctions or a lack of consistent responses to shocks. Preprocessing and response estimation was done using a custom script. Skin conductance response (SCR) was estimated with a Trough-to-Peak-scoring method[21,26] 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; otherwise the trial was scored as a zero[27]. Within-subjects trial-wise responses were range-corrected by dividing each response by the highest response (typically elicited by the shock). To ensure comparability across conditions and minimize shock-induced responses, only non-reinforced trials were included, excluding CS+ trials co-terminating with the US. SCR values were square-root transformed for group analysis. To verify robustness, we additionally estimated SCR using dynamic causal modeling (DCM) of anticipatory responses with its own preprocessing pipeline[28] (see the supplementary material).

Pupil size response. Pupil size and gaze direction (left eye) were recorded at 500 Hz using an EyeLink 1000 system (SR Research, Ottawa, Canada). Gaze direction was calibrated via a 5-point procedure in the EyeLink 1000 software. Blink-related missing data were removed and linearly interpolated 100 ms before and after each blink. Data points exceeding a predefined 7° visual angle threshold were also treated as missing and interpolated. Participants with more than 35% missing data in any acquisition phase were excluded[29]. Interpolated time series were band-pass filtered (0.05–4 Hz). Event-related pupil size responses (PSR) were computed by averaging pupil size from 1 to 4 s post-stimulus onset, normalized to a 1-s pre-stimulus baseline (-1 to 0 s). This window was chosen to minimize light reflex effects[30]. The average PSR was calculated for all condition-specific stimulus presentations across the full task or task halves. To ensure comparability, only non-reinforced trials were analyzed, excluding CS+ trials co-terminating with the US. To verify robustness, we also estimated PSR using a GLM-based method[31] (see the

supplementary material).

Procedure

Two experimental sessions took place on consecutive days in the same test room. Details of the procedure are described in the supplementary material. On Day 1 (~90 min), participants completed four self-report questionnaires on personality (trait anxiety, intolerance of uncertainty), childhood trauma, and social relations, followed by the PAL task. Day 2 (~90 min) included four experimental tasks. It began with the 2-AFC task, followed by the attachment of physiological recording equipment (pupil size and gaze, skin conductance, pulse oximeter) and electrodes for mild electric shocks. Shock intensity was calibrated using an ascending staircase procedure, adjusted to a level participants rated as "maximally uncomfortable without being painful" 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 sample. Participants then underwent differential delay threat conditioning. After a short break watching a 7-minute abstract animation (*Inscapes*[32]), they completed the test for the inferential expression of Pavlovian threat memory, maintaining the same shock intensity. To test if the relational memory structure remained stable, participant repeated the 2-AFC task after which the study concluded with debriefing questions.

Statistical analyses

Data pre-processing, plotting, and analysis were conducted in Matlab R2016a and R 3.3.3[33]. Sample means are reported with 95% confidence intervals (CI) obtained through bootstrapping (n=5000) and are indicated in square brackets.

To test each of our hypotheses, we compared the plausibility of two competing statistical models, i.e., a null and full model which corresponded to the null and alternative hypotheses, respectively[34]. The null model is the full model reduced by a test predictor. The comparison between models was quantified by the Bayes Factor (BF) that is interpreted as the weight of evidence coming from the data. We approximated the BF using Bayesian

Information Criterion (BIC) given by:

$$BIC(H_i) = -2 * LogL_i + k_i * log(n)$$

where n is the number of observations, k_i is the number of free parameters of model H_i, and L_i is the maximum likelihood for model H_i[34]. The BIC approximation of the BF is given by the ratio of prior predictive probabilities:

$$BF_{01} \approx \frac{P_{BIC}(D \mid H_0)}{P_{BIC}(D \mid H_1)} = \exp(\Delta BIC_{10}/2)$$

where $\Delta BIC_{10} = BIC(H_1) - BIC(H_0)$. We assumed *a priori* that the models under consideration are equally plausible. BF_{10} values were interpreted using the standard classification: 1-3 for weak evidence, 3-10 for moderate, and above 10 for strong evidence. Additionally, we performed frequentist statistical analyses that is a more common approach to statistical inference. To this end, the null and alternative models were compared using a likelihood-ratio test, with significance set at $p = .05$, and reported in the supplementary material.

When specifying a statistical model, we used (Generalised) Linear Mixed Models [G/LMM;[35]], implemented in the R packages glmmTMB[36] and glmmADMB[37,38]. We aimed to include a full random structure in each model but when a model exhibited convergence or singular fit problems, its random structure was successively reduced, detailed in the supplementary manuscriptaterial. To quantify the effect size, we calculated a conditional and marginal coefficient of determination (pseudo-R^2) for each (G)LMM using the R package MuMIn[39]. To assess the assumptions of LMMs, we inspected the normality of model residuals and their homoscedasticity against the fitted values. For GLMMs with a binomial error structure, we checked for overdispersion using the R package DHARMa that compares the dispersion of simulated residuals to the observed residuals using a non-parametric test[40].

Results

Learning overlapping pairs of neutral images led to a stable and systematically organised knowledge of their relations

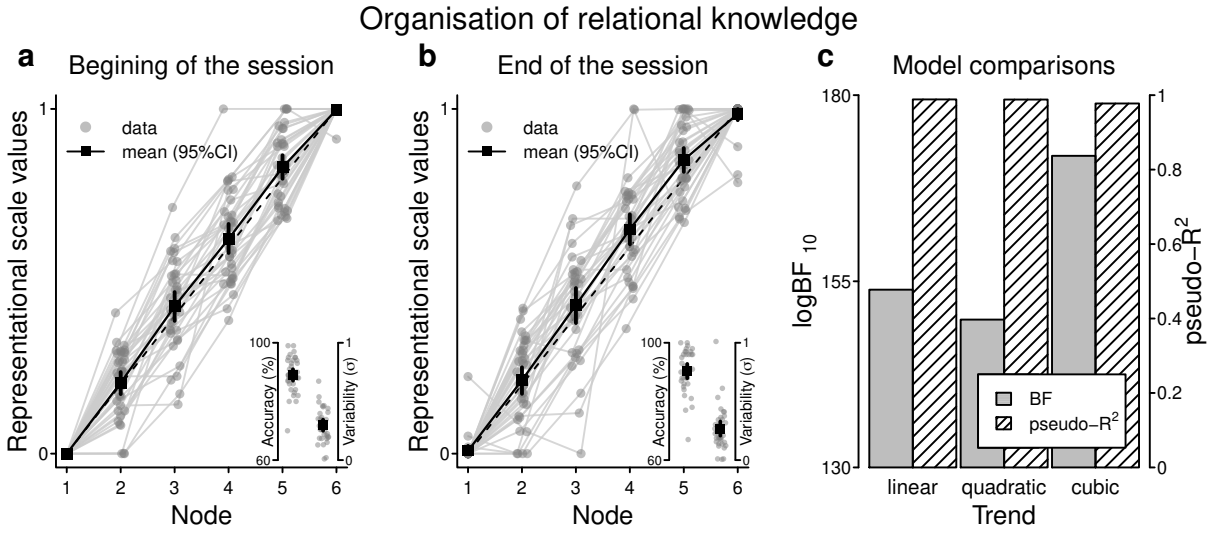

Figure 3. The acquired relational knowledge conforms to the imposed linear arrangement. Panel **a)** and **b)** illustrate the representational scale values estimated with maximum likelihood difference scaling (MLDS) using binary responses from the 2-AFC task performed on day 2 either at the beginning of the session (i.e., before threat conditioning) or at the end of the session (i.e., after inference test). The diagonal dashed line indicates the ideal linear arrangement of the representational values. Inset plots illustrate the variability of the binary responses together with their accuracy (the chance level is 72.8% due to the inclusion threshold). Panel **c)** illustrates comparisons of the null model against alternative models including polynomial trends to the 1st (linear), 2nd (quadratic), and 3rd (cubic) order using an approximated Bayes Factor (BF) together with pseudo-R^2 that is based on within-subject change of variance relative to an intercept-only model. Pseudo-R^2 reveals the linear trend as the major contributor in explaining the data variability.

Participants performed well in the memory training on day 1. Across five blocks of the paired-associate learning task, they showed successive improvement in their accuracy (block 1 = 67.78% [64.03, 71.06], block 2 = 81.44% [78.38, 84.26], block 3 = 88.75% [85.37, 91.71], block 4 = 91.76% [89.49, 93.89], block 5 = 93.15% [91.06, 95.05]) and reaction time (block 1 = 1.05s [0.98, 1.11], block 2 = 0.92s [0.86, 0.99], block 3 = 0.86s [0.80, 0.91], block 4 = 0.80s [0.75, 0.86], block 5 = 0.78s [0.73, 0.83]).

Next, we tested participants' knowledge of the full relational structure among the six images. We analyzed the 2-AFC task, conducted both at the beginning and at the end of the session (i.e., before threat conditioning and after the inference test), to assess the stability of the relational structure. The 2-AFC task indicated that indeed participants formed and maintained the memory for the relational structure by integrating overlapping pairs of neutral elements. The binary responses collected during the task showed that the

retrieval accuracy was above the 72.8% chance level both at the beginning (89.08% [87.06, 90.97]) and at the end of the session (90.42% [87.92, 92.78]).

While the accuracy scores indicated the correct inference of relationships between indirect (i.e., non-premise) pairs from the trained (i.e., premise) pairs, it is a coarse index of the underlying memory organisation. To recover a more fine grained arrangement of the memory elements, we used representational scale values estimated via maximum likelihood difference scaling (MLDS)[25]. This method estimated for each participant the distances between memory elements that best predict the binary responses in the 2-AFC task (Figure 3a-b). MLDS analysis revealed that participants exhibited low response variability, relatively to one unit of representational distance (0.2), both at the beginning (0.30 [0.25, 0.34]) and the end of the session (0.26 [0.21, 0.33]). Further, the averages of representational scale values closely resembled equispaced distances between adjacent nodes indicating their successive order, regardless of whether they were measured at the beginning (node 1 = 0.00 [0.00, 0.00], node 2 = 0.21 [0.17, 0.24], node 3 = 0.43 [0.39, 0.47], node 4 = 0.62 [0.58, 0.67], node 5 = 0.83 [0.80, 0.87], node 6 = 1.00 [0.99, 1.00]) or at the end of the session (node 1 = 0.01 [0.00, 0.02], node 2 = 0.21 [0.17, 0.25], node 3 = 0.43 [0.38, 0.48], node 4 = 0.65 [0.61, 0.69], node 5 = 0.85 [0.82, 0.89], node 6 = 0.99 [0.97, 1.00]).

To formally assess the profile of relational memory organisation recovered from the MLDS, we specified generalised linear mixed models (GLMM) with a *logit* link function and beta error structure that quantified a linear (and polynomial) effect of the graph node on the representational scale values nested within a time factor (the beginning and the end of the session). The comparison of statistical models (Figure 3c) revealed that the representational scale values were best fitted with the GLMM ($\log BF_{10} = 171.85$) including polynomial trends up to the 3rd order: linear (*logit* estimate $\pm$ SE = 1.46 $\pm$ 0.12 and 1.77 $\pm$ 0.12 for the beginning and end of the session, respectively), quadratic (*logit* estimate $\pm$ SE = -0.07 $\pm$ 0.07 and -0.10 $\pm$ 0.07 for the beginning and end of the session, respectively), and cubic (*logit* estimate $\pm$ SE = 0.62 $\pm$ 0.09 and 0.40 $\pm$ 0.09 for the beginning and end of the session, respectively). While the best model indicated the non-linearity in the

representational values, inspection of the effect sizes (pseudo-R^2 based on within-subject change of variance relative to an intercept-only model) suggested the linear trend as the major contributor in explaining the variation of the data (pseudo-R^2 = 0.99, 0.99, and 0.98 for the relative change in the effect size of the models including linear, quadratic, and cubic trend, respectively).

To further corroborate the evidence that the memory organisation was independent from the time when it was measured (beginning vs. end of the session), we specified a GLMM with a *logit* link function and beta error structure that quantified the effect of a two-level factor of time on representational scale values, taking into account the effect of the graph node indicated in the previous GLMM. We found strong evidence ($logBF_{10}$ = -7.25) that indeed the memory organisation was stable over time (*logit* estimate ± SE = -0.31 ± 0.16, 0.04 ± 0.10, and 0.22 ± 0.12 for the interaction term with the linear, quadratic, and cubic trend, respectively).

Taken together, these results show that learning overlapping pairs of neutral images led to a stable and systematically organised knowledge of their relations that conformed to the imposed linear arrangement.

Threat responses to non-conditioned cues increased as their relational distance to the conditioned cue decreased

Participants revealed medium to large effect size of differences in conditioned responses between the CS+ and CS- condition for each response modality (Figure 4a-c): proportion of shock expectancy (CS+ = 0.68 [0.59, 0.78], CS- = 0.00 [0.00, 0.01], odds ratio = 126,354.60; magnitude of the SCR (CS+ = 0.43 [0.36, 0.50], CS- = 0.29 [0.24, 0.35]; Hedges's g =0.82), and magnitude of the PSR (CS+ = 1.08 [1.07, 1.09], CS- = 1.05 [1.04, 1.06]; Hedges's g =1.18).

Having established that participants successfully acquired the linear knowledge structure and exhibited differential conditioned responses to the CS+ and CS- images, we next examined the responses to the remaining four images of the graph (GS1-GS4). Here,

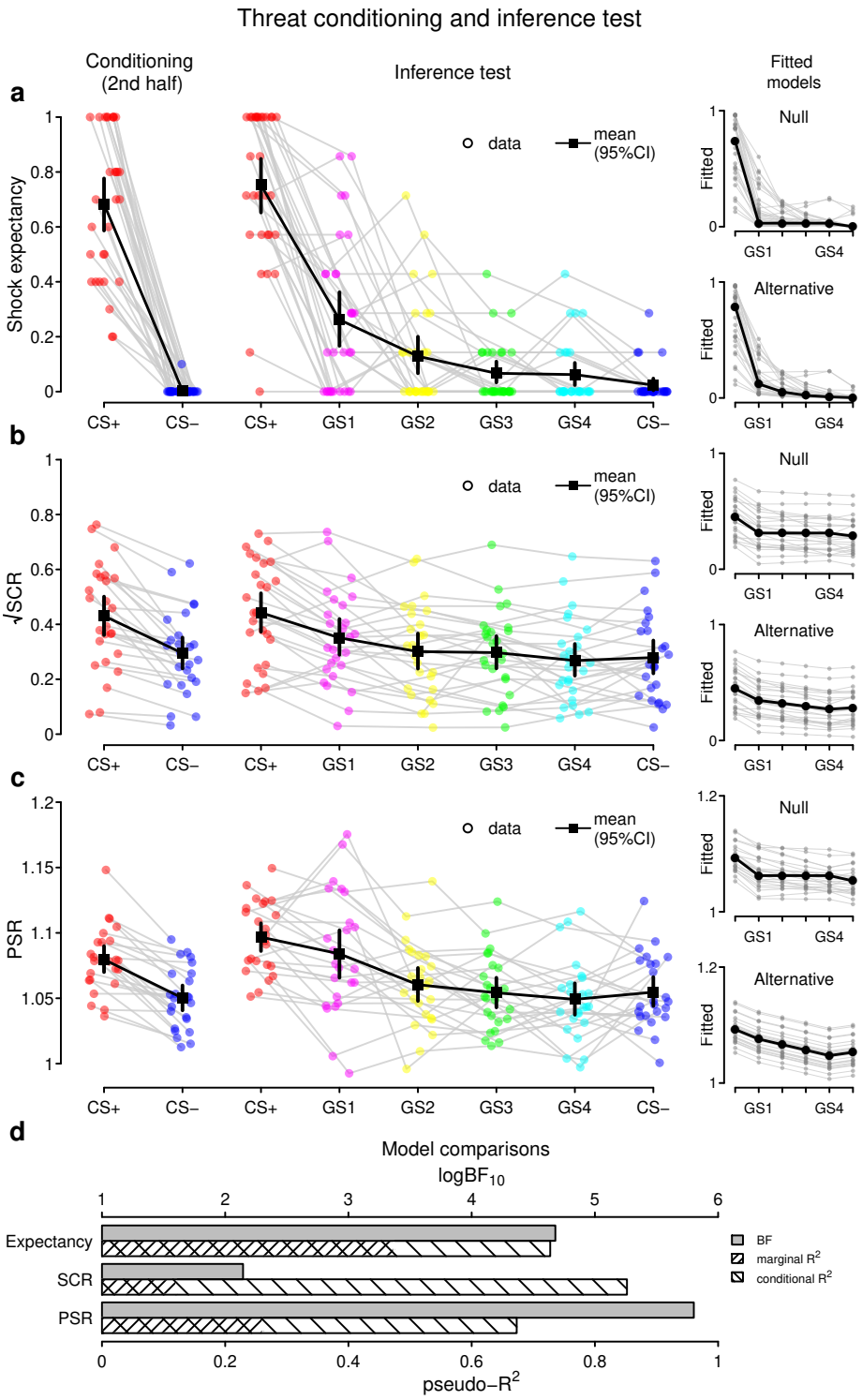

Figure 4. Results of Pavlovian cue conditioning and the inference test. Panel **a)**, **b)**, and **c)** illustrate threat responses for shock expectancy ratings, skin conductance response (SCR), and pupil size response (PSR), respectively. Corresponding subplots illustrate fitted values under the null and the alternative models (including stimulus identity as a nuisance term in the model) that were compared to test for the (non-)linear gradient of threat responses evoked by GS1-GS4 stimuli. Panel **d)** illustrates model comparisons using an approximated Bayes Factor (BF) together with the pseudo coefficient of determination (marginal and conditional pseudo-R^2), indicating strong evidence for a gradient of threat responses to GS1-GS4 that follows the linear structure of the pre-existing relational knowledge.

participants were exposed to all six familiar images presented one at a time while one of them (CS+) continued to occasionally co-terminate with the electric shock. During the

inference test (Figure 4a-c), participants continued to exhibit differential conditioned responses between the CS+ and CS- (expectancy: CS+ = 0.75 [0.65, 0.85], CS- = 0.02 [0.00, 0.05]; SCR: CS+ = 0.44 [0.37, 0.51], CS- = 0.28 [0.22, 0.34]; PSR: CS+ = 1.10 [1.09, 1.11], CS- = 1.05 [1.04, 1.07]) as well as gradients of responses to GS1-GS4 as a function of their distance to the CS+ based on the structure of the pre-existing relational knowledge (expectancy: GS1 = 0.26 [0.17, 0.36], GS2 = 0.13 [0.07, 0.20], GS3 = 0.07 [0.03, 0.11], GS4 = 0.06 [0.02, 0.10]; SCR: GS1 = 0.35 [0.29, 0.42], GS2 = 0.30 [0.24, 0.37], GS3 = 0.30 [0.24, 0.36], GS4 = 0.27 [0.21, 0.33]; PSR: GS1 = 1.08 [1.07, 1.10], GS2 = 1.06 [1.05, 1.07], GS3 = 1.05 [1.04, 1.07], GS4 = 1.05 [1.04, 1.06]).

To formally test for the gradient of the inferred risk based on the structure of the pre-existing relational knowledge, we quantified a trend across responses to GS1-GS4 stimuli in a G/LMM that took into account responses to the CS+/CS- and the stimulus identity. GLMM fitted to the proportion of the shock expectancy revealed strong evidence for its non-linear gradient (*logit* estimate ± SE = -0.45 ± 0.11, $\log BF_{10} = 4.68$) and good fit to the data (marginal pseudo-$R^2 = 0.47$ and conditional pseudo-$R^2 = 0.73$) with the dispersion parameter ($\phi = 0.86$ without significant evidence for under-dispersion ($p = .149$[40]). LMM fitted to the magnitude of the SCR revealed moderate evidence for its linear gradient (estimate ± SE = -0.013 ± 0.004, $\log BF_{10} = 2.14$) and good fit to the data (marginal pseudo-$R^2 = 0.12$ and conditional pseudo-$R^2 = 0.85$). Finally, LMM fitted to the magnitude of the PSR revealed strong evidence for its linear gradient (estimate ± SE = -0.005 ± 0.001, $\log BF_{10} = 5.80$) and good fit to the data (marginal pseudo-$R^2 = 0.26$ and conditional pseudo-$R^2 = 0.67$). These results were further corroborated by supplementary analyses using different estimation methods of the SCR and PSR (see Supplementary Materials). The supplementary analyses also report estimates of all individual predictors included in each model, together with the likelihood ratio test for the frequentist inference, which was in line with the presented results.

To gain a deeper understanding of our results, we subsequently ran a number of exploratory analyses (for details, see Supplementary Materials). Among others, we explored

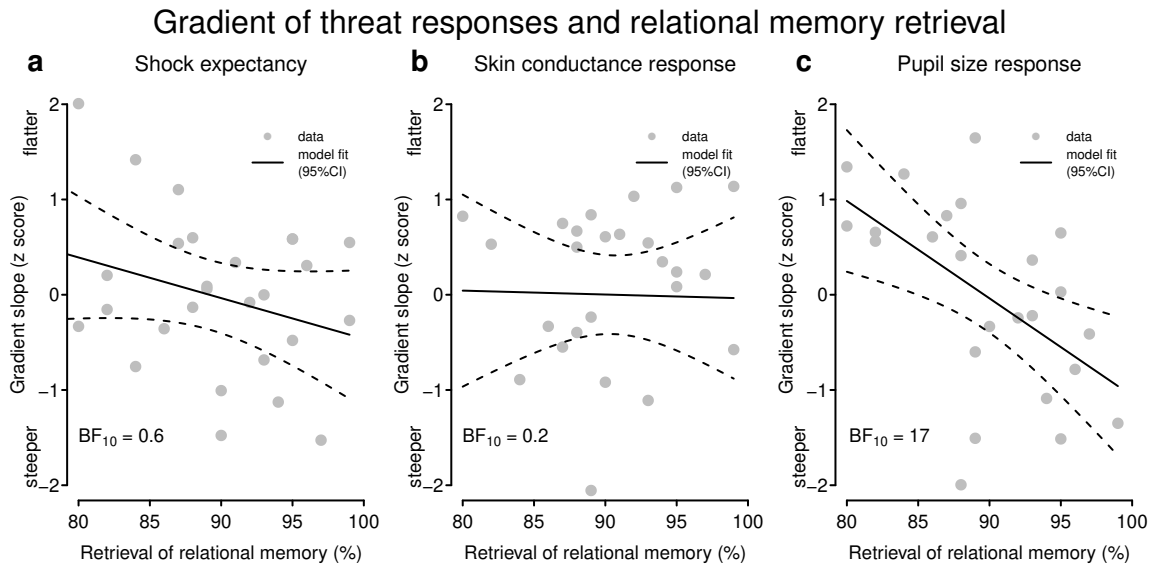

Figure 5. Accuracy of relational memory retrieval at the beginning of the session (i.e., before threat conditioning) predicts the gradient of threat responses evoked by non-conditioned cues. Panel **a)**, **b)**, and **c)** indicate the association for each response modality. Slopes of the gradients were retrieved from the full (G)LMMs and z transformed for standarisation. The approximated Bayes Factor (BF) indicates how much more likely the data are under the assumption that memory accuracy is related to the slope of the gradient.

whether the slope of the response gradient evoked by GS1-GS4 stimuli could be predicted from the retrieval accuracy of the relational knowledge during the 2-AFC task performed at the beginning of the session (Figure 5). We found evidence that better memory performance predicted a steeper slope for PSR (estimate $\pm$ SE = -0.10 $\pm$ 0.03, BF_{10} = 17.29, adjusted R^2 = 0.28), but not for the shock expectancy (estimate $\pm$ SE = -0.04 $\pm$ 0.03, BF_{10} = 0.60, adjusted R^2 = 0.04) nor SCR (estimate $\pm$ SE = 0.00 $\pm$ 0.04, BF_{10} = 0.20, adjusted R^2 = -0.04). In contrast, the slope was independent from individual differences in personality traits relevant for the expression of Pavlovian threat memory such as self-report anxiety, intolerance for uncertainty, and childhood traumatic experiences (for details, see Supplementary Material). These results suggest that the inference about the aversive outcome, based on pre-existing relational knowledge, may partially depend on the ease with which this knowledge is retrieved.

Taken together, these results demonstrate that the gradient of responses to the non-conditioned cues depended on the structure of pre-existing relational knowledge, which encoded information about their associative distance to the cue that directly predicted the aversive outcome.

Discussion

Here, we investigated whether individuals draw on the structure of pre-existing relational knowledge – built across multiple episodic experiences – together with Pavlovian threat memory to infer aversive outcomes from cues never experienced in an aversive context. We found that when participants encoded the relationships among neutral images into an abstract linear structure and were later exposed to these images under the risk of a mild electric shock, their threat responses to the images that were never a part of conditioning scaled with the relational distance to the image previously associated with the shock. The generalisation gradient was observed across both subjective shock expectancy ratings and physiological indices of arousal, including skin conductance and pupil size. These findings suggest that individuals assess the level of danger by combining associative learning across aversive and neutral contexts.

How do individuals infer relationships between cues and outcomes to guide their defensive behavior, even when these relationships are learned across separate episodes? Previous studies using a sensory pre-conditioning paradigm[41] have shown that separate episodes of associative learning and Pavlovian conditioning can be combined to anticipate the outcome value from a pre-conditioned cue[42,43]. In this paradigm, individuals first learn that cue A predicts cue B. Later, cue B becomes associated with a US through Pavlovian conditioning. As a result, individuals begin to respond to cue A as though they expect the delivery of the US. This behavior is typically interpreted in two ways: either as a direct association between cue A and the US, established through inference during conditioning when A is reactivated by B, i.e., mediated learning[43], or as an indirect association formed via inference during the test phase when A is explicitly present[18,42]. The mediated learning view does not fully explain the effects of extinction learning. When the first-order cue B undergoes extinction, responding to the pre-conditioned cue A is also eliminated[42,44]. If cue A had directly acquired an association with the US during conditioning, its response should remain unaffected. In contrast, the inference at the test suggests that individuals reason through the associations rather than forming direct links – if A leads to B and B leads to the US,

then A is inferred to eventually lead to the US, which could explain the effect of extinction. Yet, the pre-conditioned cue A does not seem to support conditioned reinforcement[45,46], i.e., a procedure used to directly assess cue value by measuring an individual's willingness to work for a cue in the absence of a reward[47]. The fact that cue A fails to support instrumental responses (unlike cue B) suggests that it neither directly accrues value during conditioning nor gains value by default during subsequent test. Instead, cue A may serve a purely informational role, signaling the occurrence of the conditioned cue B, which in turn enables individuals to infer the probability of a valued outcome[16]. If pre-conditioned cues primarily signal the conditioned cue, the magnitude of responses to these cues should depend on how their associations with the conditioned cue are structured[11]. In line with this view, we showed that by organizing such associations in an abstract linear pattern, pre-conditioned cues elicited threat responses that scaled with their associative distance to the conditioned cue. Therefore, combining the literature on Pavlovian sensory pre-conditioning and episodic (relational) memory provides a putative explanation on how individuals may flexibly gauge danger using cues encountered in safe contexts across multiple episodic experiences.

Using pre-existing relational knowledge of the environmental structure to infer risk builds upon the concept of a 'cognitive map', supported by the hippocampal-entorhinal system[48,49]. This framework, derived from research on spatial navigation in animals, suggests that experiences can be mentally simulated by traversing a map that encodes the relationships between entities in the world, guiding goal-oriented behavior. For instance, Gauthier et al.[50] identified a population of neurons in the hippocampus that encode the location of rewards in physical space. These cells may help animals plan their preferred routes by simulating future trajectories before movement such as when foraging for food[51] or avoiding danger[52]. Similarly, in our study, participants anticipated the possibility of receiving a mild electric shock as if navigating through an abstract linear structure in their mental models considering the relational distance. This mirrors the behaviour of animals running on a linear track in physical space, who slow down proportionally when approaching a localized threat and speed up once past it[53]. Together, these observations suggest that the same mechanism used to regulate defensive responses to threat proximity in physical space may

also be employed to estimate the likelihood of an aversive event from discrete cues, based on their proximity to a known threat in an abstract cognitive (memory) space.

Using knowledge of environmental structure together with Pavlovian memories that are stored in separate brain regions[54] might be adaptive due to their complementary properties. Pavlovian threat memories are characterized by permanence and a strong resistance to being unlearned by new experiences[55,56]. In contrast, relational memories of neutral elements are flexible, allowing them to incorporate new information and adjust their structure to track changes in environmental contingencies[57]. This flexibility, paired with the durability of Pavlovian threat memory, may help individuals infer risk in a dynamic environment, especially when new elements are introduced or old ones are removed.

Our results open new avenues for future studies. First, while our results suggest that individuals combine the relationships between neutral cues and biologically relevant cues acquired separately through conditioning, it would be interesting to understand better whether the function of this interaction – and, by extension, the inference process – depends on the underlying structure of these relationships. For instance, cues may be predictive of one another, where one cue appears as the other disappears (A->B), or they may partially overlap in time (A->AB). To better understand the function of pre-conditioned cues, future research could perhaps directly compare various structures of the relationship between pre-conditioned and conditioned cues. In some cases, pre-conditioned cues may serve as warning signals that predict threats (threat prediction), while in other circumstances, they might be inferred as threats themselves (threat detection). Our current design cannot conclusively determine one possibility over the other.

Second, to indicate the inference outcome, we measured defensive responses using both subjective and physiological indices. It would be interesting to know whether the interaction of two memories translates into an instrumental behaviour. Goal-oriented actions dependent on planning are mainly studied in the appetitive rather than aversive domain[5]. Few studies using computer games focused on an approach-avoidance conflict, where foraging for food coincides with the risk of predation[58–60]. Future research could use virtual

lab games to explore how cues, never experienced in an aversive context, influence strategic defense decisions to reduce the likelihood of threat encounters.

Third, future research could explore whether neutral cues signaling proximity to potential harm trigger avoidance behavior typically associated with anxiety-related disorders. For example, when asked to walk through a city, patients with agoraphobia avoided passing through an open market on their way to a meeting point[61]. Some patients chose a route that avoided the market square in advance, suggesting they prospectively considered the market as if simulating the future trajectory of their actions. This behavior highlights how individuals may mentally map out potential threats ahead of time, influencing their actions, which could shed new light on psychopathology.

In conclusion, our results demonstrate that individuals can assess the level of danger even in the absence of a previously encountered threat, by drawing on structured memories formed from distinct episodic experiences acquired in safe contexts. This ability to infer risk from cues never encountered in an aversive context highlights how people can protect themselves from danger despite minimal first-hand exposure to aversive outcomes.

Research transparency

All study materials are publicly available at [LINK]. Data: All primary data are publicly available at [LINK]. Analysis scripts: All analysis scripts are publicly available at [LINK].

Funding

BMB was supported by the Max Planck Society, the International Max Planck Research School "NeuroCom", Leipzig University, and Leibniz Programm of the Research Academy Leipzig.

References

1. Mobbs, D., Headley, D. B., Ding, W. & Dayan, P. Space, Time, and Fear: Survival Computations along Defensive Circuits. *Trends in Cognitive Sciences* **24**, 228–241 (2020).

2. Fanselow, M. S. The Role of Learning in Threat Imminence and Defensive Behaviors. *Current opinion in behavioral sciences* **24**, 44–49 (2018).

3. Rescorla, R. A. Pavlovian conditioning: It's not what you think it is. *American Psychologist* **43**, 151–160 (1988).

4. Johansen, J. P. *et al.* Hebbian and neuromodulatory mechanisms interact to trigger associative memory formation. *Proceedings of the National Academy of Sciences* **111**, E5584–E5592 (2014).

5. LeDoux, J. E. & Daw, N. D. Surviving threats: Neural circuit and computational implications of a new taxonomy of defensive behaviour. *Nature Reviews Neuroscience* **19**, 269–282 (2018).

6. Dunsmoor, J. E. & Murphy, G. L. Stimulus typicality determines how broadly fear is generalized. *Psychological Science* **25**, 1816–1821 (2014).

7. Onat, S. & Büchel, C. The neuronal basis of fear generalization in humans. *Nature Neuroscience* **18**, 1811–1818 (2015).

8. Baczkowski, B. M., Haaker, J. & Schwabe, L. Inferring danger with minimal aversive experience. *Trends in Cognitive Sciences* **27**, 456–467 (2023).

9. Eichenbaum, H. Memory: Organization and Control. *Annual Review of Psychology* **68**, 19–45 (2017).

10. Momennejad, I. Learning Structures: Predictive Representations, Replay, and Generalization. *Current Opinion in Behavioral Sciences* **32**, 155–166 (2020).

11. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nature Neuroscience* **20**, 1643–1653 (2017).

12. Garvert, M. M., Dolan, R. J. & Behrens, T. E. A map of abstract relational knowledge in the human Hippocampal–Entorhinal cortex. *eLife* **6**, e17086 (2017).

13. Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B. & Botvinick, M. M. Neural representations of events arise from temporal community structure. *Nature Neuroscience* **16**, 486–492 (2013).

14. Zeithamova, D., Schlichting, M. L. & Preston, A. R. The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience* **6**, 70 (2012).

15. Schlichting, M. L. & Preston, A. R. Memory integration: Neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences* **1**, 1–8 (2015).

16. Barron, H. C. *et al.* Neuronal Computation Underlying Inferential Reasoning in Humans and Mice. *Cell* **183**, 228–243.e21 (2020).

17. Wang, F., Schoenbaum, G. & Kahnt, T. Interactions between human orbitofrontal cortex and hippocampus support model-based inference. *PLOS Biology* **18**, e3000578 (2020).

18. Sadacca, B. F. *et al.* Orbitofrontal neurons signal sensory associations underlying model-based inference in a sensory preconditioning task. *eLife* **7**, e30373 (2018).

19. Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D. & Walker, M. P. Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 7723–7728 (2007).

20. 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).

21. 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).

22. 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).

23. Zeithamova, D., Dominick, A. L. & Preston, A. R. Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron* **75**, 168–179 (2012).

24. Blough, D. S. Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes* **1**, 3–21 (1975).

25. Maloney, L. T. & Yang, J. N. Maximum likelihood difference scaling. *Journal of Vision* **3**, 5–5 (2003).

26. 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).

27. 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).

28. 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).

29. Korn, C. W. & Bach, D. R. A solid frame for the window on cognition: Modeling event-related pupil responses. *Journal of Vision* **16**, 28 (2016).

30. Hermans, E. J. *et al.* Persistence of Amygdala-Hippocampal Connectivity and Multi-Voxel Correlation Structures During Awake Rest After Fear Learning Predicts Long-Term Expression of Fear. *Cerebral Cortex* **27**, 3028–3041 (2017).

31. 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).

32. 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).

33. R Core Team. *R: A language and environment for statistical computing*. (R Foundation for Statistical Computing, 2017).

34. Wagenmakers, E.-J. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review* **14**, 779–804 (2007).

35. Baayen, R. H., Davidson, D. J. & Bates, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* **59**, 390–412 (2008).

36. Brooks, M. E. *et al.* glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal* **9**, 378–400 (2017).

37. Fournier, D. A. *et al.* AD Model Builder: Using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. *Optimization Methods and Software* **27**, 233–249 (2012).

38. Skaug, H. J., Fournier, D. A., Bolker, B. M., Magnusson, A. & Nielsen, A. glmmADMB: Generalized Linear Mixed Models using 'AD Model Builder'. (2016).

39. Bartoń, K. MuMIn: Multi-model inference. (2018).

40. Hartig, F. DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models. (2018).

41. Brogden, W. J. Sensory pre-conditioning. *Journal of Experimental Psychology* **25**, 323–332 (1939).

42. Jones, J. L. *et al.* Orbitofrontal cortex supports behavior and learning using inferred but not cached values. *Science (New York, N.Y.)* **338**, 953–956 (2012).

43. Wimmer, G. E. & Shohamy, D. Preference by Association: How Memory Mechanisms in the Hippocampus Bias Decisions. *Science (New York, N.Y.)* **338**, 270–273 (2012).

44. Rizley, R. C. & Rescorla, R. A. Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology* **81**, 1–11 (1972).

45. Sharpe, M. J., Batchelor, H. M. & Schoenbaum, G. Preconditioned cues have no value. *eLife* **6**, e28362 (2017).

46. Sharpe, M. J. *et al.* Dopamine transients do not act as model-free prediction errors during associative learning. *Nature Communications* **11**, 1–10 (2020).

47. Burke, K. A., Franz, T. M., Miller, D. N. & Schoenbaum, G. The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. *Nature* **454**, 340–344 (2008).

48. Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490–509 (2018).

49. O'Keefe, J. & Nadel, L. The hippocampus as a cognitive map. in *The hippocampus as a cognitive map* – (Clarendon Press, 1978).

50. Gauthier, J. L. & Tank, D. W. A Dedicated Population for Reward Coding in the Hippocampus. *Neuron* **99**, 179–193.e7 (2018).

51. Johnson, A. & Redish, A. D. Neural Ensembles in CA3 Transiently Encode Paths Forward of the Animal at a Decision Point. *Journal of Neuroscience* **27**, 12176–12189 (2007).

52. Wu, C.-T., Haggerty, D., Kemere, C. & Ji, D. Hippocampal awake replay in fear memory retrieval. *Nature Neuroscience* **20**, 571–580 (2017).

53. Girardeau, G., Inema, I. & Buzsáki, G. Reactivations of emotional memory in the hippocampus-amygdala system during sleep. *Nature Neuroscience* **20**, 1634–1642 (2017).

54. Wong, F. S., Westbrook, R. F. & Holmes, N. M. 'Online' integration of sensory and fear memories in the rat medial temporal lobe. *eLife* **8**, e47085 (2019).

55. Bouton, M. E. Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. *Biological Psychiatry* **52**, 976–986 (2002).

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

57. Milivojevic, B., Vicente-Grabovetsky, A. & Doeller, C. F. Insight reconfigures hippocampal-prefrontal memories. *Current biology: CB* **25**, 821–830 (2015).

58. Bach, D. R. Anxiety-Like Behavioural Inhibition Is Normative under Environmental Threat-Reward Correlations. *PLOS Computational Biology* **11**, e1004646 (2015).

59. Mobbs, D. *et al.* When Fear Is Near: Threat Imminence Elicits Prefrontal-Periaqueductal Gray Shifts in Humans. *Science (New York, N.Y.)* **317**, 1079–1083 (2007).

60. Qi, S. *et al.* How cognitive and reactive fear circuits optimize escape decisions in humans. *Proceedings of the National Academy of Sciences* **115**, 3186–3191 (2018).

61. Walz, N., Mühlberger, A. & Pauli, P. A Human Open Field Test Reveals Thigmotaxis Related to Agoraphobic Fear. *Biological Psychiatry* **80**, 390–397 (2016).