OXFORD

Cerebral Cortex, 2022, **32**, 3081–3097
https://doi.org/10.1093/cercor/bhab402
Advance access publication date: 28 November 2021
Original Article

Prediction errors for aversive events shape long-term memory formation through a distinct neural mechanism

Felix Kalbe, Lars Schwabe*

Department of Cognitive Psychology, Institute of Psychology, Universität Hamburg, Hamburg 20146, Germany
*Corresponding author: Lars Schwabe. Email: lars.schwabe@uni-hamburg.de

Prediction errors (PEs) have been known for decades to guide associative learning, but their role in episodic memory formation has been discovered only recently. To identify the neural mechanisms underlying the impact of aversive PEs on long-term memory formation, we used functional magnetic resonance imaging, while participants saw a series of unique stimuli and estimated the probability that an aversive shock would follow. Our behavioral data showed that negative PEs (i.e., omission of an expected outcome) were associated with superior recognition of the predictive stimuli, whereas positive PEs (i.e., presentation of an unexpected outcome) impaired subsequent memory. While medial temporal lobe (MTL) activity during stimulus encoding was overall associated with enhanced memory, memory-enhancing effects of negative PEs were linked to even decreased MTL activation. Additional large-scale network analyses showed PE-related increases in crosstalk between the "salience network" and a frontoparietal network commonly implicated in memory formation for expectancy-congruent events. These effects could not be explained by mere changes in physiological arousal or the prediction itself. Our results suggest that the superior memory for events associated with negative aversive PEs is driven by a potentially distinct neural mechanism that might serve to set these memories apart from those with expected outcomes.

Key words: arousal; associative learning; medial temporal lobe; salience network; schema network.

Introduction

Imagine meeting Barack Obama in the supermarket. Most likely, this event would deviate strongly from what you expected during your grocery shopping, resulting in a prediction error (PE). PEs are considered a to be driving force in reinforcement learning, during which an organism learns incrementally to achieve pleasant and avoid unpleasant states (Niv 2009; Glimcher 2011). Moreover, it may be expected that single episodes encoded in the context of a high PE should be preferentially stored in episodic memory. Although this would aid behavioral adaptation (Shohamy and Adcock 2010; Gershman and Daw 2017), PEs received little attention in episodic memory research (for early exceptions, see Henson and Gagnepain 2010; Mizumori 2013). Only recently, behavioral evidence started to accumulate showing that PEs associated with appetitive or aversive events may promote episodic memory formation of nearby events (Greve et al. 2017; Rouhani et al. 2018; Jang et al. 2019; Ergo et al. 2020; Kalbe and Schwabe 2020). A fundamental question concerns how PEs boost long-term memory formation. While PEs in both rewarding and aversive contexts are strongly linked with the neurotransmitter dopamine (Schultz et al. 1997; Matsumoto and Hikosaka 2009; Papalini et al. 2020), the neural mechanisms underlying their modulatory effects on memory formation are still largely unknown.

One way through which PEs may promote memory for surrounding events is by enhancing well-known mechanisms of long-term memory formation strongly linked to the medial temporal regions, including the hippocampus and parahippocampal gyrus (Alvarez and Squire 1994; Eichenbaum 2001). It is further well established that hippocampal memory formation is enhanced by emotional arousal through a process thought to be mediated by the amygdala, which strengthens memory formation processes in the hippocampus, parahippocampal gyrus, and related areas that together form a "medial temporal encoding network" (MTEN, McGaugh and Roozendaal 2002; Richardson et al. 2004; Strange and Dolan 2004; Hermans et al. 2014). Thus, one hypothesis would be that PE-driven episodic memory enhancements are due to increases in medial temporal lobe activation.

Alternatively, PEs might drive long-term memory formation through mechanisms that are distinct from those known to underlie common memory formation. Initial behavioral evidence suggests that PE effects on episodic memory formation go beyond the effects of physiological

Received: June 25, 2021. **Revised:** September 9, 2021. **Accepted:** October 12, 2021

arousal (Kalbe and Schwabe 2020). Furthermore, events associated with high PEs have been suggested to create event boundaries and establish a new latent context resulting in a separate memory trace (Rouhani et al. 2020). These behavioral findings point to the alternative that PEs might induce a qualitative shift in mnemonic processing. Specifically, an alertness response in reaction to unexpected outcomes (Summerfield and Egner 2009; Metereau and Dreher 2013) may be mediated by the salience network (Ham et al. 2013; Fouragnan et al. 2018), mainly comprised of the bilateral anterior insula and the dorsal anterior cingulate cortex (dACC; Garrison et al. 2013; Ham et al. 2013). At the same time, if high PE events are processed separately from expected events that match existing knowledge structures represented in what is referred to as a schema (Ghosh and Gilboa 2014), it can be further predicted that PEs result in a decreased recruitment of the neural "schema-network," comprised mainly of the angular gyrus, the precuneus and the medial prefrontal cortex (mPFC; van Kesteren et al. 2012; Vogel et al. 2018a). Accordingly, this alternative view predicts that the enhanced memory for events encoded in the context of high PEs is due to an activation of the salience network, accompanied by an even reduced activation of areas implicated in memory formation for events that are in line with prior experience (i.e., the MTEN and "schema-network").

To test these alternative hypotheses, participants performed an incidental encoding task in which they saw a series of stimuli from different categories that were associated with different probabilities to receive a mild electric shock. Comparing shock expectancy ratings given by participants to the actual trial outcome allowed us to determine the direction and extent to which participants experienced a PE in each trial. In line with the terminology of the influential Rescorla-Wagner model (Rescorla and Wagner 1972), we labeled an unexpected omission of the reinforcer (i.e., an unexpected shock omission) a negative PE, while the unexpected delivery of the reinforcer (i.e., an unexpected shock) was labeled a positive PE (Schultz 1998; Delgado et al. 2008; McHugh et al. 2014). Memory was probed in a recognition test 24 h after encoding. To unravel the neural mechanisms underlying PE-related enhancements of episodic memory, we used behavioral modeling, arousal measurement and fMRI in combination with large-scale network analysis.

Materials and methods

Participants

Sixty-one healthy volunteers (35 women, 26 men; mean age ± SD = 24.97 ± 4.65 years) participated in this experiment. Eleven participants had to be excluded from the analysis due to excessive head motion in the scanner (>5 mm within a single experimental block; $N = 2$), incidental finding of a frontal lesion ($N = 1$), missing >25% of responses on the task ($N = 6$), selecting only extreme ratings (i.e., 0% and 100%; $N = 1$), or not returning for the second experimental day ($N = 1$). To determine the target sample size, we performed an a priori power analysis based on previous findings of binary aversive PE effects on episodic memory formation (Kalbe and Schwabe 2020). As this study used a conceptually similar generalized linear mixed-effect model, we applied a simulation-based approach using the *SIMR R* package (Green and MacLeod 2016). We assumed the same effect size but increased the number of trials from 60 to 120 to account for the modified design in the present study. This indicated that a sample size of $N = 50$ participants would result in a statistical power of above 0.95. All participants met safety criteria for MRI and electrodermal stimulation, had normal or corrected-to-normal vision, were right-handed, had never studied psychology nor neuroscience, did not suffer from any psychiatric or neurological conditions, and reported no alcohol abuse, nor use of any illicit drug. They were paid 45€ upon completion of the second experimental day. The study protocol was approved by the ethics board of the University of Hamburg and all participants provided written informed consent prior to their participation.

Experimental procedure

The experiment took place on two consecutive days. On the first experimental day, participants completed a combined incidental encoding and fear learning task in the MRI scanner (Fig. 1A). About 24 h later, they completed a surprise recognition test for stimuli presented during the encoding session.

At the beginning of the first experimental day, participants provided informed consent and were prepared for the MRI scanner by placing a pair of MRI-safe gelled disposable electrodes (BIOPAC systems, Goleta, CA) over the thenar eminence of the left hand to measure skin conductance as an indicator of physiological arousal during the encoding task using the BIOPAC MP-160 system (BIOPAC systems, Goleta, CA). Another pair of electrodes was placed on the right side of the right lower leg, approximately 20 cm above the ankle, and used to administer aversive electric shocks during the fear learning task. Shocks were applied using the BIOPAC STMISOC (BIOPAC systems, Goleta, CA) connected to a BIOPAC STM100C stimulator (BIOPAC systems, Goleta, CA). After participants were placed in the scanner, they first completed an unrelated task that included stimuli that were distinct from the stimuli used in this experiment.

Prior to the start of the fear learning task, shock intensity was adjusted to be unpleasant but not painful by administering a series of test shocks that increased in intensity until participants rated the shocks as not yet painful but highly unpleasant. Participants then received detailed written instructions about the following fear learning task. On each trial, participants saw an image that was presented centrally on a screen for 4.5 s (Fig. 1A). Beneath each image, participants saw a slider that always started at 50% and could be adjusted to any integer value between 0% and 100% by using the left and right buttons

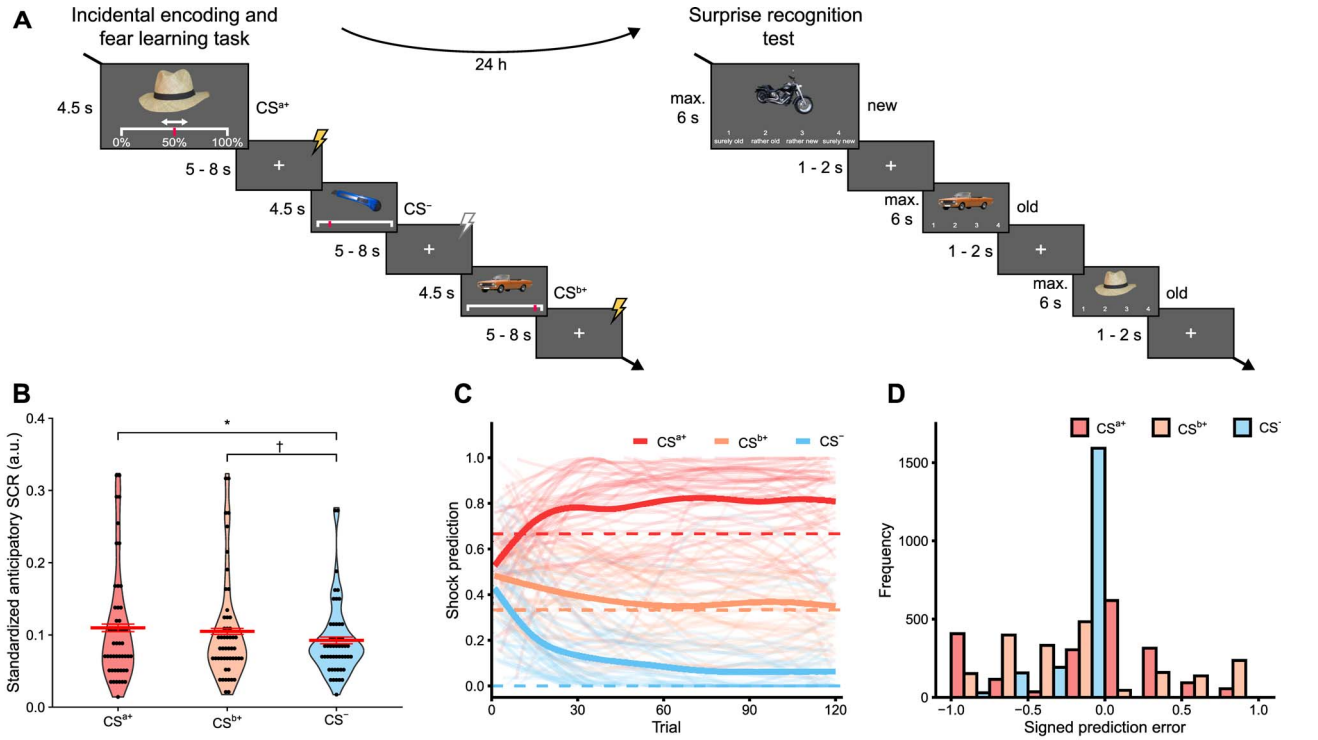

Fig. 1. Experimental task and performance parameters. (*A*) Participants completed a combined incidental encoding and fear learning task and a surprise recognition test for its contents about 24 h later. In the encoding task, participants saw a series of unique pictures from three different categories (clothing, vehicles, and tools) linked to fixed probabilities to receive an electric shock (CS^{a+}—67%, CS^{b+}—33%, and CS^{-}—0%). On each trial, participants indicated their shock expectation. Approximately 24 h later, they saw all pictures from the previous day intermixed with the same number of new pictures and categorized each picture as either "old" or "new." (*B*) Mean standardized anticipatory skin conductance responses (SCR) confirmed successful fear conditioning, as reflected in significantly elevated SCR to both CS^{a+} and CS^{b+} items compared with CS^{-} items. Black dots show data from individual participants. Thick red bar represents group mean, while thin red bars show ±1 standard error of the mean. (*C*) Participants' mean shock expectancy ratings (thick lines) approached the true shock probabilities (dotted lines) relatively fast, although there was a tendency to overestimate shock probabilities. Thin lines represent data from individual participants. (*D*) Signed PEs were prevalent in both the positive and negative domain for CS^{a+} and CS^{b+} pictures. PEs for CS^{-} pictures were mostly zero, reflecting that participants learned that items from this category were never paired with a shock. †$P < 0.05$, *$P_{corr} < 0.05$.

of an MRI-compatible response box (Current Designs Inc., Philadelphia). Participants were instructed that while each image was present, they should adjust the slider to a value that corresponded with their prediction of the probability that a shock would follow. Participants were requested to confirm their rating by pressing the central button on the response box. In 40 out of the total of 120 trials, a 200 ms shock to the right lower leg followed immediately after image offset. Between trials, there was a jittered white fixation cross presented for 5–8 s. This relatively long inter-trial interval allowed us to observe the slowly emerging skin conductance response (SCR) to each outcome as well as to separate trials at the neural level. Critically, the probabilities of a shock were linked to image categories. While participants were explicitly instructed that they would see images of vehicles, clothing, and tools, they were not told that these categories would be linked to pre-defined shock contingencies. Participants were informed that their predictions would have no effect on the probability that a shock would occur, but that their aim should still be to improve their predictions over the course of the task. Out of 40 occurrences of the CS^{a+} category, 27 were followed by a shock, corresponding to a shock probability of approximately 2/3. Likewise, 40 occurrences of the CS^{b+} category were followed by a shock in 13 trials, leading to a shock probability of approximately one-third for the CS^{b+}. Finally, the 40 occurrences of the CS^{-} category were never followed by a shock. The six possible combinations of image categories (i.e., vehicles, clothing, tools) with conditioning categories (i.e., CS^{a+}, CS^{b+}, CS^{-}) were counterbalanced across participants. Participants completed four blocks with 30 trials each, resulting in a total of 120 trials. All images were selected to be unique exemplars of their respective categories. For example, there were not two different pictures of wristbands within the clothing category. Between each block, the experimenter asked participants whether they still perceived the shock as unpleasant but not painful and readjusted the intensity when needed. Upon completion of all four experimental blocks, participants rated the average unpleasantness of the shock over the task on a scale ranging from 1 ("not unpleasant at all") to 10 ("extremely unpleasant"). Their mean rating was 6.31 (SD = 1.49).

After an interval of 22–26 h, participants returned for a surprise recognition test outside of the MRI scanner. In this recognition test, they saw all 120 images that had been presented on the previous day randomly

intermixed with the same number of previously unseen ("new") images from the same three categories (40 new images per category). The allocation of images as learning items or lures was randomized and therefore unique for each participant. For each image, participants had a maximum of 6 s to indicate whether the current image had been presented on the previous day ("old") or not ("new") and how confident they were, using buttons corresponding to "definitely old," "maybe old," "maybe new," and "definitely new." Between each of the 240 trials of the recognition test (120 old, 120 new), a white fixation cross appeared centrally for 1–2 s.

MRI data acquisition

Functional MRI data were acquired during the incidental encoding session on a Siemens Magnetom Prisma 3T scanner equipped with a 64-channel head coil. For each of the four functional runs, approximately 185 volumes were recorded using a multi-band echo-planar imaging (EPI) sequence with the following parameters: 60 axial slices of 2 mm depth, slice orientation parallel to the AC-PC line, phase-encoding in AP direction, repetition time (TR) of 2000 ms, echo time (TE) of 30 ms, 60-degree flip angle, 224 × 224 mm field of view (FOV), 2 mm isotropic resolution, EPI factor of 112, echo distance of 0.58 ms. For each block, four images were recorded before the start of the behavioral task to ensure equilibrium magnetization. These initial images were discarded as dummy scans during further analyses. Following the last functional run, a T1-weigthed scan was acquired with 256 slices, coronal orientation, repetition time (TR) of 2300 ms, echo time (TE) of 2.12 ms, a 240 × 240 mm field of view (FOV), and a 0.8 × 0.8 × 0.9 mm voxel size.

Behavioral analysis

For each individual trial, the prediction uncertainty (*PU*) was derived from participants' shock predictions, while signed PEs (sPE) were calculated by contrasting predictions with actual outcomes. Specifically, the *PU* is a continuous variable that can take any value between 0 (least possible uncertainty) and 1 (maximum uncertainty) and was calculated as:

$$\mathrm{PU}(t) = 1 - |\mathrm{P}(t) - 0.5| \times 2$$

where P(t) is the continuous explicit shock prediction made by the participant in trial *t* (ranging from 0 to 1).

The sPE in trial *t* is a continuous variable that can take any value between −1 and +1 and was calculated as:

$$\mathrm{sPE}(t) = \mathrm{O}(t) - \mathrm{P}(t)$$

where O(t) is the binary outcome in trial *t* (coded 0 when no shock occurred and 1 when a shock occurred). Note that the sign of the sPE contains information about the outcome of the trial. sPEs < 0 could only occur in unshocked trials, while sPEs > 0 could occur when a shock occurred. Only for sPE = 0, the binary outcome of the trial is ambiguous.

The prediction uncertainty PU(t) for any trial *t* can also be calculated directly from the sPE (but not vice versa) using:

$$\mathrm{PU}(t) = 1 - ||\mathrm{sPE}(t)| - 0.5| \times 2$$

To test influences of uncertainty, PEs, and arousal (measured through SCRs) on episodic memory formation, we performed mixed-effects logistic regression at the level of individual trials, as implemented in the *lme4 R* package (Bates et al. 2015). The binary recognition of a previously presented item (collapsed over confidence ratings) was treated as the dependent variable, coded 0 for misses and coded 1 for hits. Following recommendations to maximize the generalizability of these models (Barr et al. 2013), we included the maximum random effects structure, estimating random intercepts and random slopes per predictor per subject. We did not include random intercepts per item to account for different baseline memorability as their inclusion led to singular fit in some models. For the central effect of PEs on memory formation, we considered three possible relationships: a linear effect of PEs, a quadratic effect of PEs (i.e., U-shaped), and a quadratic effect with negative versus positive PEs affecting memory in opposite directions (i.e., S-shaped). Notably, the later relationship could be modeled similarly well by a cubic regressor. However, as the previous literature from the reward domain reported quadratic effects (Rouhani et al. 2018) and both regressor differed only negligibly in terms of model fit (ΔAIC = 2.2), we only report results from the quadratic S-shaped regressor.

Skin conductance analysis

During the incidental encoding session of the first experimental day, we recorded electrodermal activity as a measure of physiological arousal. These data were analyzed in Ledalab Version 3.4.9 (Benedek and Kaernbach 2010) using a Continuous Decomposition Analysis (CDA) to derive the average phasic driver within given response windows. In short, the CDA aims to separate the continuous skin conductance data into a tonic, stimulus-independent component, and a phasic, stimulus-driven component. To obtain more precise estimates of the underlying sudomotor nerve activity compared with more traditional methods such as a through-to-peak analysis, the CDA only considers changes in the phasic component in response to an event. As a first measure, we defined anticipatory SCRs as reactions occurring from the onset of the decision in each trial (i.e., the confirmation of the shock rating) until the end of the stimulus presentation (i.e., exactly 4.5 s after stimulus onset). Additionally, we defined outcome-related SCRs to occur 0.5 s after the outcome of the current trial was revealed (i.e., whether a shock would occur or not) until

2.9 s after the outcome onset to ensure that this measure would capture activity evoked by the current trial, but not the following. Skin conductance data were downsampled from 1000 to 50 Hz. Model optimization was repeated for four sets of initial values to reduce the influence of local optima and identify the set of parameters leading to the overall lowest model error (Benedek and Kaernbach 2010). The minimum amplitude threshold was set to 0.01 μS for both anticipatory and the outcome-related SCRs. Individual physiological factors, such as the thickness of the corneum, can greatly affect the range of observed SCRs (Figner and Murphy 2011). To account for this interindividual variability, both the anticipatory and the outcome-related SCRs were standardized by dividing the average phasic driver estimate by the maximum average phasic driver value observed in any trial of this participant.

For post-hoc analyses of ANOVA results, we also report corrected P values under the label P_{corr}. In these cases, we corrected for multiple comparisons using the Bonferroni method.

fMRI preprocessing

Functional MRI data were preprocessed in MATLAB using SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12). First, functional volumes were spatially realigned to the first image in the time series. This step also yielded six motion parameters used in univariate analyses to control for motion-related activation artifacts. Realigned volumes were co-registered to each participant's structural image. Then, images were spatially normalized into standard stereotactic (MNI) space using unified segmentation. For univariate fMRI analyses, the normalized functional images were additionally smoothed with an 8 mm full-width half-maximum Gaussian kernel.

Univariate fMRI analyses

Based on theoretical accounts and results from our behavioral modeling, we identified 1) prediction uncertainty, 2) shock expectancy, 3) quadratic signed PEs, and 4) physiological arousal (measured through anticipatory and outcome-related SCR, respectively) as key variables to explain episodic memory formation in this fear learning task. To investigate the neural basis of these effects, we modeled the fMRI time series using generalized linear models (GLMs). These models included regressors of the onsets of the stimulus and outcome presentation as predictors of interest, and nuisance regressors to account for head movement (i.e., the six movement parameters derived from spatial realignment). As behavioral data suggested separable effects of positive versus negative quadratic prediction errors on memory formation, we fitted separate models for unshocked trials (corresponding to negative PEs) and shocked trials (corresponding to positive PEs). Both models featured onsets of stimuli with shock expectancy and prediction uncertainty as parametric modulators. To control for possible effects of arousal, standardized anticipatory SCRs were also placed as an additional parametric modulator on stimulus onsets. A second regressor featured onsets of outcomes with quadratic prediction errors as the critical parametric modulator. Again, we controlled for possible confounding effects of arousal by placing standardized outcome-related SCRs as an additional parametric modulator on outcome onsets.

For the estimation procedure, data from each of the four experimental blocks were concatenated using the spm_fmri_concatenate function in SPM12, a high-pass filter at 1/128 Hz was applied, and an AR(1) process was used to adjust for temporal autocorrelation. Second-level analysis were constructed from each subject's first level contrasts using a standard one-sample t-test approach in SPM. We thresholded all resulting t-maps using a whole-brain voxel-level family-wise error corrected P value of $P_{FWE} < 0.05$.

To link differences in neural activity with subsequent recognition performance, we specified two additional univariate fMRI models: The first model aimed to identify clusters linked with subsequent recognition during the encoding of individual stimuli and used stimulus onsets a regressor with the binary subsequent recognition of an item as the sole parametric modulator. To elucidate the neural basis of the memory-enhancing effects of PEs, we specified an additional univariate fMRI model with onsets of outcomes (i.e., when a PE occurred) as a regressor and PEs (ranging between 0 and 1), the binary subsequent recognition of an item (coded 0 for misses and 1 for hits) and their interaction as parametric modulators. These models were estimated separately for shocked and unshocked trials to account for the opposite effects of negative versus positive PEs on memory using the same procedure as described above. Based on the vast literature linking structures of the medial temporal lobe with declarative memory formation (Alvarez and Squire 1994; Eichenbaum 2001), we defined the bilateral hippocampus, as well as the bilateral (posterior) parahippocampal gyrus as regions of interests and performed small volume corrections, which were additionally corrected for the number of search regions using a Bonferroni correction. We refer to these additionally Bonferroni-corrected P values with P_{corr}. Voxels belonging to each of these regions with a probability threshold of 50% were identified based on an existing anatomical atlas (Harvard-Oxford structural atlas; Desikan et al. 2006).

Large-scale network-connectivity analyses

We performed analyses of functional connectivity in the CONN toolbox (Whitfield-Gabrieli and Nieto-Castanon 2012) to assess how within- and between-network connectivities of memory-relevant brain networks differed depending on PE magnitudes. As this analysis did not allow for continuous parametric modulators, we instead split PEs into low ($|sPE| < 0.5$) versus high ($|sPE| \geq 0.5$). Our analyses focused on PE effects at outcome time for unshocked trials. However, in the specific GLM for this

analysis, we included onset regressors for each combination of the following factors: stimulus versus outcome onsets, shocked versus unshocked, and low versus high PEs. This resulted in a total of eight regressors in this model. In a first-level analysis, to denoise data, we applied a linear detrending and a standard band-pass filter of 0.008–0.09 Hz. Besides the just mentioned effects of PEs, we added white matter, cerebrospinal fluid, and movement regressors obtained from spatial realignment as additional confounds to the model. Further analysis focused on pre-defined regions of interest and networks implemented in the CONN toolbox: 1) dorsal anterior cingulate cortex, bilateral anterior insula, bilateral rostral prefrontal cortex and bilateral supramaginal gyrus forming the *salience network* (Menon 2011); 2) medial prefrontal cortex, bilateral angular gyrus and precuneus forming the schema network (van Kesteren et al. 2012; Vogel et al. 2018a); and 3) bilateral hippocampus, bilateral anterior parahippocampal gyrus, and bilateral posterior parahippocampal gyrus as the medial temporal encoding network (Fernández et al. 1999; Shrager et al. 2008).

Results

Successful fear learning

Physiological and explicit rating data indicated successful fear learning. Specifically, standardized anticipatory SCR differed significantly between CS categories, $F(2,98) = 3.62$, $P = 0.030$, $\eta_G = 0.011$ (Fig. 1B). Post hoc paired *t*-tests revealed that participants showed increased anticipatory SCRs to CS^{a+} pictures relative to CS^{-} pictures, $t(49) = 2.38$, $P_{corr} = 0.042$, $d_{av} = 0.27$. At trend level, they also showed greater anticipatory SCRs to CS^{b+} pictures compared with CS^{-} pictures, $t(49) = 2.10$, $P = 0.041$, $P_{corr} = 0.082$, $d_{av} = 0.20$ (Fig. 1B). Explicit shock ratings further showed that participants learned to associate picture categories with their respective shock probabilities over the course of the task (Fig. 1C). Participants had a significantly higher shock expectancy for CS^{a+} ($M = 0.78$, $SD = 0.14$) than for CS^{b+} ($M = 0.38$, $SD = 0.16$), $t(49) = 11.53$, $P_{corr} < 0.001$, $d_{av} = 2.67$, and for CS^{b+} than for CS^{-} ($M = 0.12$, $SD = 0.12$), $t(49) = 10.81$, $P_{corr} < 0.001$, $d_{av} = 1.87$.

From participants' explicit shock expectancy ratings, we derived signed PEs by contrasting each prediction with the binary outcome (i.e., unshocked or shocked) in the respective trial (see section Materials and Methods). Resulting PEs ranged from −1 to 1, with negative values in cases of unexpected shock omissions and positive values in cases of unexpected shocks, while greater distances from 0 in both directions indicated greater discrepancies between predictions and outcomes. Importantly, the distribution of signed PEs (Fig. 1D) showed a sufficient number of positive and negative PEs to allow reliable conclusions of their both their effects on memory formation. Moreover, the explicit shock ratings allowed us to directly assess participants' prediction uncertainty, which ranged from 0 (maximal certainty, corresponding to predictions of 0% or 100%) to 1 (maximal uncertainty, corresponding to a prediction of 50%).

Overall recognition memory performance

In the recognition test 24 h after encoding, participants performed overall very well, as indicated by markedly higher hit rates (i.e., the rate of correctly classifying previously seen pictures as "old") than false alarm rates (i.e., the rate of incorrectly classifying unseen pictures as "old"), $M_{hitrate} = 60.9\%$ ($SD = 0.15$), $M_{FArate} = 21.1\%$ ($SD = 0.098$). Participants were significantly more certain with their responses for hits ($M = 0.59$, $SD = 0.18$) than for false alarms ($M = 0.26$, $SD = 0.20$), $t(49) = 15.92$, $P < 0.001$, $d_{av} = 1.70$.

A repeated-measures ANOVA showed that hit rates differed significantly between CS categories, $F(2,98) = 7.29$, $P = 0.001$, $\eta_G^2 = 0.05$. For false alarm rates, on the other hand, there was no such difference between CS categories, $F(2,98) = 0.25$, $P = 0.77$, $\eta_G^2 = 0.003$, suggesting that the actual memory but not the response bias differed between CS categories. Post-hoc paired *t*-tests showed that hit rates were selectively enhanced for items from the CS^{a+} category, which was associated with a shock probability of 67%, compared with both items from the CS^{b+} category ($t(49) = 4.15$, $P_{corr} < 0.001$, $d_{av} = 0.54$), which was associated with a shock probability of 33%, and the CS^{-} category ($t(49) = 2.64$, $P_{corr} = 0.022$, $d_{av} = 0.40$; Fig. 2A), which was never followed by a shock. Enhanced recognition performance for CS^{a+} items was also obtained when hits and false alarms were integrated to the sensitivity d' based on signal detection theory: A repeated measures ANOVA confirmed that d' was generally different between CS categories ($F(2,98) = 3.70$, $P = 0.028$, $\eta_G^2 = 0.03$), with post-hoc *t*-tests showing a trend towards an increased memory sensitivity for CS^{a+} items ($M = 1.39$, $SD = 0.66$) compared with CS^{b+} items ($M = 1.17$, $SD = 0.51$), $t(49) = 2.28$, $P = 0.027$, $P_{corr} = 0.053$, $d_{av} = 0.38$. Further, memory sensitivity for CS^{a+} items was significantly greater than for CS^{-} items ($M = 1.17$, $SD = 0.60$), $t(49) = 2.42$, $P_{corr} = 0.038$, $d_{av} = 0.35$.

At first glance, one might assume that these differences are simply due to differences in (arousing) shock presentations between CS categories. However, our data did not support this interpretation. The greater proportion of shocked items could not explain the improved hit rate for the CS^{a+} category: A repeated-measures ANOVA to explain hit rates indicated no memory advantage for shocked over unshocked items per se ($F(1,49) = 1.12$, $P = 0.294$, $\eta_G^2 = 0.022$). Further, a 2×2 repeated-measures ANOVA confirmed increased hit rates for CS^{a+} over CS^{b+} items even after controlling for shocks ($F(1,49) = 19.47$, $P < 0.001$, $\eta_G^2 = 0.08$). Notably, this ANOVA even showed a tendency towards decreased hit rates for shocked items, $F(1,49) = 3.76$, $P = 0.058$, $\eta_G^2 = 0.006$), with no significant interaction ($F(1,49) < 0.001$, $P = 0.997$, $\eta_G^2 < 0.0001$ (Supplementary Fig. 1). These findings indicate that differences between CS categories in the number of presented shocks cannot

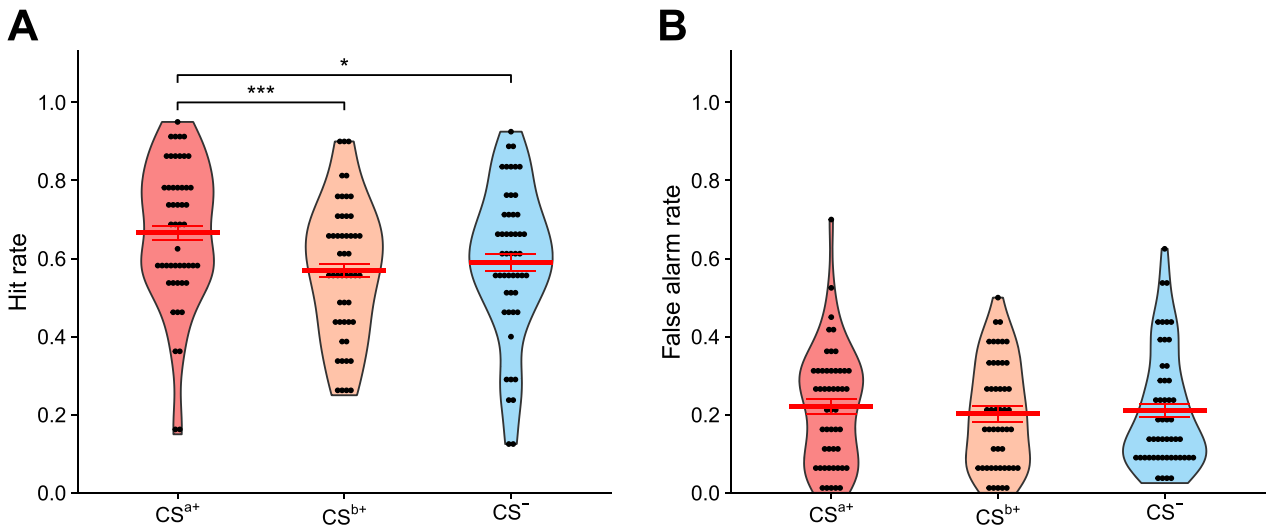

Fig. 2. General recognition performance by CS category. (A) Hit rates for items from the CS^{a+} category were significantly larger compared with both CS^{b+} and CS^{-} items. Although CS^{a+} items had the highest shock probability, this could not explain their increased hit probability as hit rates for shocked items even tended to be lower than for unshocked items (Supplementary Fig. 1). (B) False alarm rates were comparable for all three conditioning categories, showing that the CS-type did not affect the mnemonic response bias. Black dots show data from individual participants. Thick red bar represent group means, while thin red bars show ±1 standard error of the mean. $^{*}P_{corr} < 0.05$, $^{***}P_{corr} < 0.001$.

explain the differential memory performance and that other factors drive the boost in memory.

Aversive PEs and prediction uncertainty modulate episodic memory formation beyond arousal

To explain episodic memory formation in the incidental encoding task at trial level, we fitted generalized linear mixed-effects models (GLMMs) with a binary response variable (hit vs. miss) and a logit link function (i.e., mixed-effects logistic regression) using the *lme4 R* package (Bates et al. 2015). The dependent variable was the recognition of an item in the surprise recognition test, coded 0 for misses and 1 for hits. We applied the maximum random effects structure (Barr et al. 2013), estimating random intercepts and random slopes of all predictors per subject.

To determine the shape of the putative relationship between PEs and memory formation (Fig. 3A), we fitted three initial models over all trials (including both negative and positive PEs) using 1) linear PEs, 2) quadratic PEs, and 3) a variant of quadratic PEs that assumes effect of negative versus positive PEs to be in opposite directions based on the following inverted S-shaped transformation:

$$f(x) = \begin{cases} x^2 & \text{if } x \leq 0 \\ -x^2 & \text{if } x > 0 \end{cases}$$

Model comparisons using the Akaike information criterion (AIC) to identify the optimal model while also considering increased model complexity favored the inverted S-shaped model (AIC = 7420.8) over both the linear (AIC = 7425.2) and the quadratic model (AIC = 7434.3). The absolute differences in AIC compared with the best-fitting inverted S-shaped regressor (ΔAIC) can be used to evaluate the evidence in favor of an alternative PE-memory relationship (Cavanaugh and Neath 2019). This comparison indicated that there was considerably less evidence for the linear model than for the inverted S-shaped model, ΔAIC = 4.4. Furthermore, there was essentially no evidence in favor of a quadratic (i.e., U-shaped) effect of PEs on memory, ΔAIC = 13.5. Therefore, we only considered the inverted S-shaped PE regressor in subsequent models of memory formation. Results from this inverted S-shaped model indicated that negative PEs enhanced memory formation, while positive PEs decreased memory formation, $\beta = 0.27$, 95%-CI [0.07, 0.47], $z = 2.68$, $P = 0.007$.

In a next step, we asked whether the PE-effect is mainly driven by the CS^{-} category, whose items were never followed by shock and could therefore only produce negative, but not positive PEs. Even after excluding all trials featuring CS^{-} items, the S-shaped PE-effect remained virtually unchanged, $\beta = 0.30$, 95%-CI [0.11, 0.49], $z = 3.01$, $P = 0.002$, suggesting that the observed PE effects was not primarily owing to CS^{-} items. Therefore, trials from all three conditioning categories (i.e., CS^{a+}, CS^{b+}, and CS^{-}) were included in the following analyses.

Results so far suggest that greater negative PEs and greater positive PEs had opposite effects on episodic memory formation, with the former increasing and the latter decreasing the probability of a subsequent hit. However, this model assumes both effects to be equally strong in each participant. To further investigate whether this assumption is justified, we next fitted models separately for negative and positive PE trials with quadratic PEs as the sole independent variable to explain the binary recognition of an item (Fig. 3B). For negative PEs, we again observed a memory enhancement with greater PE magnitude, $\beta = 0.49$, 95%-CI [0.15, 0.82],

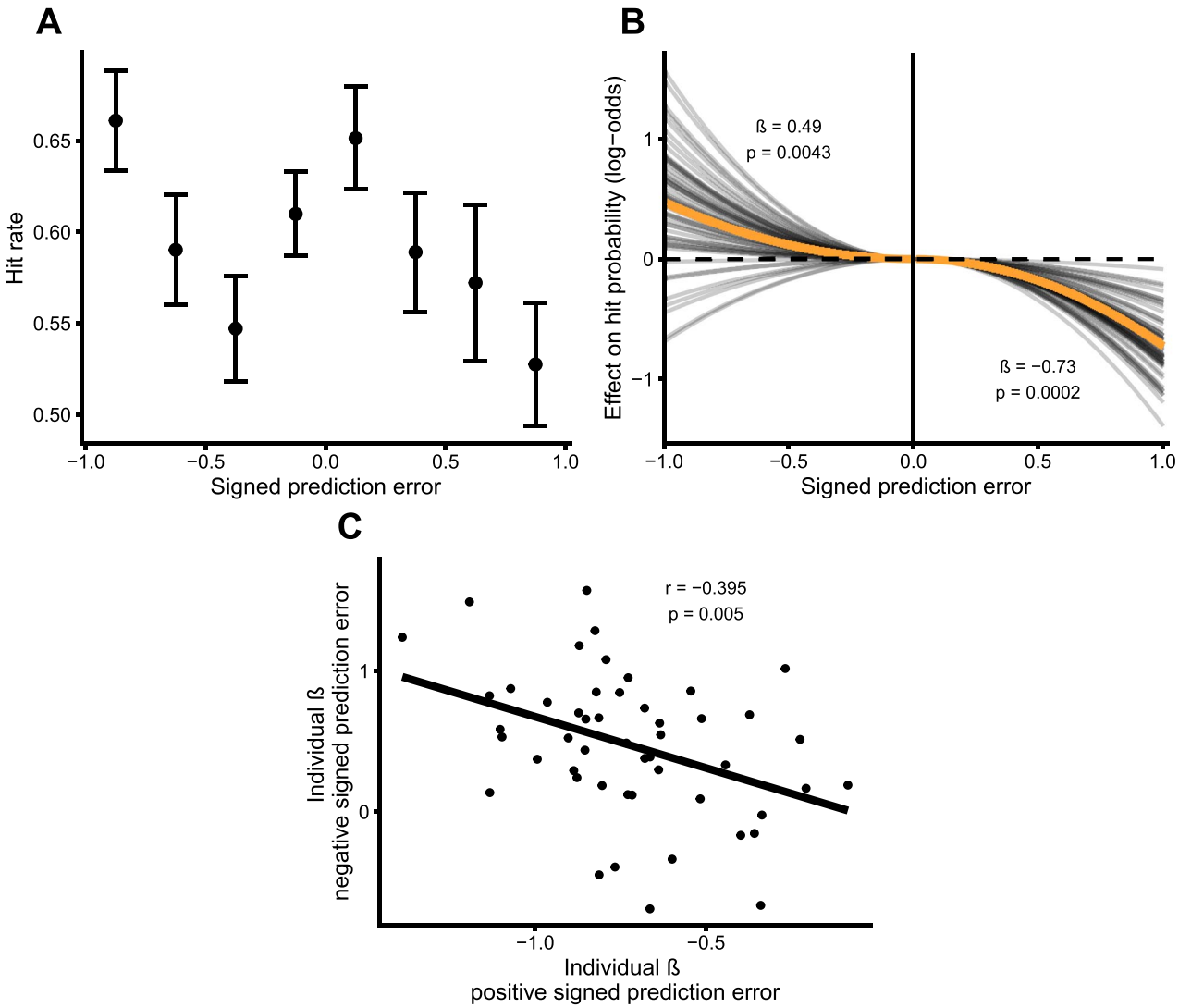

Fig. 3. Behavioral model of long-term memory formation reveals modulating influences of prediction errors. (A) Empirical relationship between signed prediction errors and recognition performance (hit rates). Points show the mean hit rate in each bin, with error bars indicating ±1 standard error of the mean. (B) Results from a trial-level mixed-effect logistic regression show opposite effects of positive and negative prediction errors on later memory. Quadratic negative prediction errors (associated with unexpected shock omissions; left half) were linked with improved memory formation for associated pictures. In contrast, quadratic positive prediction errors (associated with shocks; right half) were linked with decreased memory formation. Orange line indicates estimated fixed effects of PEs, while thin black lines show PE effects estimated separately per participant. (C) Effects of quadratic negative and positive PEs were negatively correlated at the level of participants.

$z = 2.85$, $P = 0.004$. The same model for positive PEs confirmed that greater PE magnitude was instead associated with decreased memory performance, $\beta = -0.73$, 95%-CI [−1.12, −0.34], $z = 3.67$, $P < 0.001$. Random βs per subject from both models were moderately negatively correlated, indicating that participants that showed a stronger memory benefit from negative PEs also showed a stronger memory decrease from positive PEs, $r = -.395$, $t(48) = 2.98$, $P = 0.005$ (Fig. 3C).

An alternative explanation for the observed effects of PEs on recognition memory could be that these PEs are partially confounded with several control variables that are known to enhance declarative memory formation. To investigate this possibility, we next build a joint model by including potential predictors in a stepwise fashion and evaluating whether the inclusion of each variable significantly improved the model fit, which we interpreted as evidence that this predictor additionally contributed to memory formation (Table 1). As a baseline model, we only estimated a random intercept per participant to explain the binary recognition of an item. Next, we added the binary trial outcomes (i.e., shocked or unshocked) to the model, which had no effect on the probability that an item would subsequently be recognized, $\beta = 0.09$, 95%-CI [−0.03, 0.21], $z = 1.52$, $P = 0.13$.

Classic models of episodic memory formation in aversive contexts emphasized the memory promoting role of physiological arousal (Cahill and McGaugh 1998; McGaugh 2018). Therefore, we tested in a next step whether differences in trial-level memory formation could be explained by physiological arousal. To this aim, we first added standardized anticipatory SCRs to the

Table 1. Comparison of behavioral models of memory formation.

Model	n_{par}	logLik	AIC	ΔAIC	χ^2	P	
Baseline	2	-3713.11	7430.22	31.73			
+ Outcomes	3	-3711.96	7429.91	31.41	2.31	0.128	
+ Anticipatory SCR	4	-3711.75	7431.51	33.01	0.4	0.526	
+ Outcome-related SCR	5	-3711.32	7432.64	34.14	0.87	0.351	
+ Uncertainty	6	-3705.24	7422.49	23.99	12.15	<0.001	***
+ Quadratic prediction	7	-3694.18	7402.35	3.86	22.13	<0.001	***
+ S-shaped PE	8	-3691.25	7398.50	0	5.86	0.015	*

Notes: n_{par}: Number of parameters in the model; ΔAIC: Difference in AIC compared with the model leading to the minimal AIC score; χ^2 and P refer to results from a likelihood ratio test whether the inclusion of the predictor significantly improves the model fit. $^{*}P < 0.05$. $^{***}P < 0.001$.

model, which had no significant effect on hit probabilities, $\beta = -0.11$, 95%-CI [−0.46, 0.23], $z = 0.64$, $P = 0.53$. Likewise, additionally added standardized outcome-related SCRs also had no effect on hit probabilities, $\beta = 0.16$, 95%-CI [−0.17, 0.49], $z = 0.93$, $P = 0.35$.

Finally, we added cognitive measures associated with possible shocks to the model. First, uncertainty about outcomes was linked with decreased subsequent recognition performance, $\beta = -0.35$, 95%-CI [−0.54, −0.15], $z = 3.49$, $P < 0.001$. Shock expectancy is a critical variable, because it forms, in association with shocks, the computational basis for PEs. As such, one might hypothesize that the expectancy component alone, rather than PEs, which contrasts it with actual outcomes, is the driving component for the observed memory modulation. Indeed, adding the quadratic shock expectancy (analogous to our quadratic conceptualization of PE-effects) to the model revealed that hit rates were improved when participants expected that a shock would follow, $\beta = 0.43$, 95%-CI [0.25, 0.60], $z = 4.68$, $P < 0.001$. Critically, adding our S-shaped PE-regressor to the model again confirmed that their previously described relationship with subsequent memory performance, even after controlling for several other candidate variables, $\beta = 0.85$, 95%-CI [0.16, 1.53], $z = 2.42$, $P = 0.015$. Interestingly, in this full model, quadratic shock expectancy was no longer significantly associated with subsequent memory, $\beta = -0.34$, 95%-CI [−0.99, 0.30], $z = 1.04$, $P = 0.30$. This full model including our S-shaped PE-regressor also led to the smallest (i.e., best-fitting) value in AIC compared with any of the more parsimonious models (Table 1).

Beyond these trial-unique measures, one might hypothesize that the perceived aversiveness of shocks could explain memory formation. To test this hypothesis, we added participants' subjective aversiveness rating as an additional predictor to our model. However, this indicated no significant effect of shock aversiveness ratings on recognition performance, $\beta = -0.03$, 95%-CI [−0.15, −0.10], $z = 0.40$, $P = 0.69$. Further, we found no improvement in model fit indices after adding individual shock ratings (AIC = 7400.3 for the model including shock ratings versus AIC = 7398.5 for the model without shock ratings) and therefore preferred the more parsimonious model.

Medial temporal activity during stimulus presentation is associated with subsequent memory

To link neural data with memory formation, we first ran a subsequent memory analysis in which we asked which changes in brain activity during stimulus presentation would generally be predictive of the subsequent recognition of an item. Note that this analysis does not yet capture any effects of PEs, which only emerged at a later stage when the outcome of the respective trial was revealed. We modeled the pre-processed fMRI time series using a generalized-linear model (GLM) with stimulus onsets as a regressor and the binary subsequent recognition of an item as its sole parametric modulator (see section Materials and Methods). Based on the rich literature linking the medial temporal lobe with episodic memory formation (Alvarez and Squire 1994; Eichenbaum 2001), we specified the bilateral hippocampus and the bilateral posterior parahippocampal gyrus as two candidate regions predicting subsequent memory and performed a small volume correction. In line with the literature, results showed that improved memory formation during encoding was positively linked with clusters of activity in the left posterior parahippocampal gyrus, $t(49) = 4.90$, $P_{SVC} = 0.001$ (FWE-corrected), $P_{corr} = 0.002$ (see section Materials and Methods for details about the applied correction; Fig. 4), right posterior parahippocampal gyrus, $t(49) = 4.16$, $P_{SVC} = 0.006$ (FWE-corrected), $P_{corr} = 0.012$ and, at trend level, in the right hippocampus, $t(49) = 3.89$, $P_{SVC} = 0.031$ (FWE-corrected), $P_{corr} = 0.062$.

Negative PEs are associated with greater activation of the salience-network, paralleled by decreased activation of medial temporal lobe and schema-networks

To elucidate the neural basis of negative PE-related memory enhancements, we first asked which brain areas are modulated by negative PEs. Our results (all findings significant at the whole-brain level at $P < 0.05$, FWE corrected) show that negative PEs were associated with large clusters of increased activity in the bilateral anterior insula and the dACC, which are key regions of the salience network (Menon 2011; Fig. 5A,B). In addition, negative PEs were associated with significant decreases

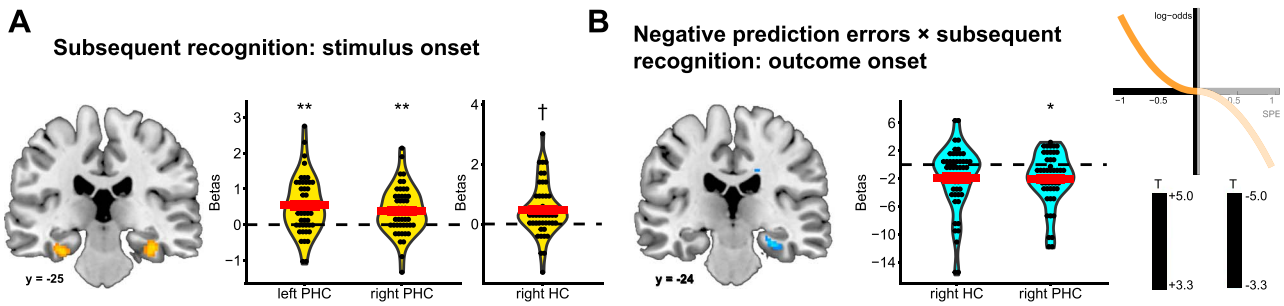

Fig. 4. Univariate fMRI analysis of subsequent memory. (*A*) Congruent with the existing literature on medial temporal lobe involvement in declarative memory formation, greater activation of the hippocampus (HC) as well as the posterior parahippocampal gyrus (PHC) during stimulus presentation were overall associated with improved subsequent memory performance. (*B*) Contrarily, for items associated with larger negative PEs that were later recognized, we found decreased BOLD responses in the right hippocampus and the right parahippocampal gyrus when the outcome of the trial was revealed. All displayed voxels were thresholded at $P < 0.001$ (uncorrected) for display purposes only. Black dots indicate beta estimates from individual participants, while the red line shows the mean beta estimate over all participants. †$P_{SVC} < 0.05$ (FWE-corrected) *$P_{corr} < 0.05$, **$P_{corr} < 0.01$.

in activation in large portions of the bilateral hippocampus and parahippocampal gyrus (Fig. 5*D*). Although it is important to note that this decrease in medial temporal lobe activity occurred only after outcomes were revealed and therefore after the offset of the to-be-remembered stimulus, this finding is in stark contrast to both our findings linking medial temporal activity during stimulus presentation with improved memory and earlier studies demonstrating this relationship (Fernández et al. 1999; Shrager et al. 2008). These findings therefore provide evidence that the PE-induced memory enhancement that we observed here might involve a neural mechanism that is different from standard modes of memory formation. In addition to decreased activation in the medial temporal lobe, we also observed decreased activity for negative PEs in the mPFC, precuneus, and left angular gyrus (Fig. 5*C*–*E*), all three of which have been described as part of the schema network that links current information to existing knowledge structures (van Kesteren et al. 2012; Vogel et al. 2018a). This finding might be taken as evidence that the superior memory for items associated with large negative PEs is associated with a distinct neural mechanism that sets these PE events apart from those with expected outcomes. Further, potential issues with collinearity in the GLM could not explain our findings, as indicated by small variance inflation factors for the critical PE regressor (Supplementary Table 1).

Same as negative PEs, prediction uncertainty in unshocked trials was associated with decreased activation in the prefrontal cortex, although this cluster was located significantly more dorsally for uncertainty (Supplementary Fig. 2*A*, Supplementary Table 4). Additionally, we observed decreased activation in the bilateral middle temporal gyrus (Supplementary Fig. 2*B*), likely reflecting decreased visual processing of stimuli associated with greater prediction uncertainty, which might explain the reduced memory for items associated with uncertainty.

For mere shock expectancy, we found no significant changes in activation in any areas that were previously linked with PEs (i.e., dACC, insula, hippocampus, mPFC, precuneus, angular gyrus). Instead, shock expectancy was only associated with changes in occipital areas, which might reflect visual processing of the slider that participants used to give their expectancy rating (Supplementary Table 5). This finding complements results from the behavioral models suggesting that the deviation of outcomes from predictions (i.e., PEs) is critical for memory modulation, rather than the mere expectation of an aversive stimulus.

Decreased medial temporal activation to larger negative PEs is linked to improved memory formation

In a next step, we assessed changes in brain activity that were directly associated with the enhanced memory for negative PEs. To this end, we fitted an additional univariate fMRI model with onsets of unshocked outcomes (rather than stimulus onsets) as a regressor and PEs, the binary subsequent recognition of an item and their interaction as parametric modulators (see section Materials and Methods). Our analysis focused on the interaction between PEs and subsequent recognition, as this specific interaction links the processing of PEs with their effects on memory formation. As in the previous analyses on subsequent memory, we focused our analysis on the hippocampus and the posterior parahippocampal gyrus using a small volume correction. In sharp contrast to our previous subsequent memory analysis at stimulus onset, we found for items associated with larger negative PEs that were subsequently recognized clusters of decreased BOLD activity in the right posterior parahippocampal gyrus, $t(49) = 3.87$, $P_{SVC} = 0.015$ (FWE-corrected), $P_{corr} = 0.030$ (Fig. 4*B*). Additionally, there was a similar non-significant trend in right hippocampus, $t(49) = 3.65$, $P_{SVC} = 0.062$ (FWE-corrected), $P_{corr} = 0.124$. These results suggest a distinct medial temporal lobe involvement in overall memory formation and PE-driven memory enhancements.

Negative PEs are associated with altered connectivity within and between memory-relevant neural networks

Based on the theoretical distinction between "standard" memory processing of events that are in line with

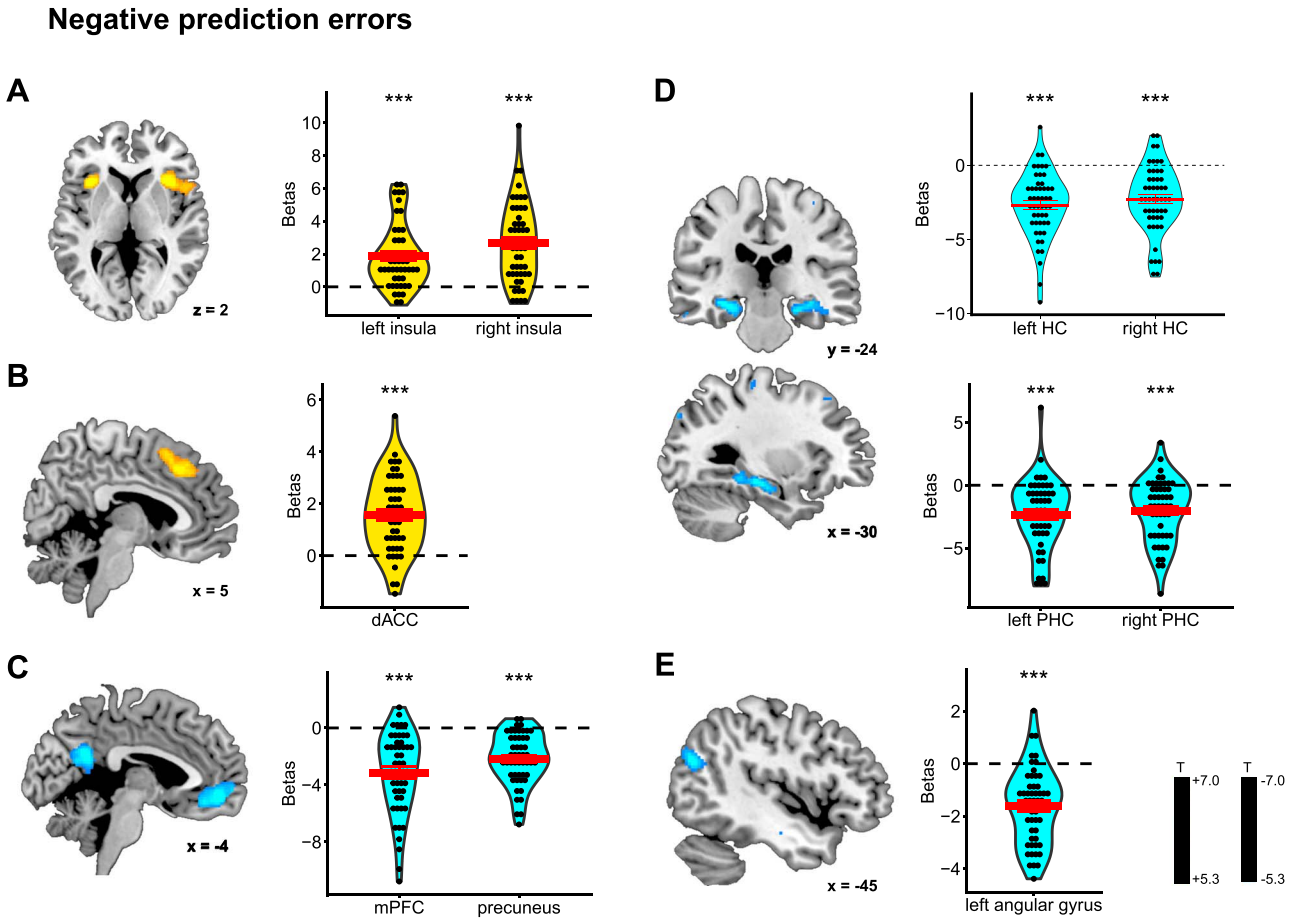

Fig. 5. Univariate fMRI analysis to identify regions associated with negative PEs. Negative PEs were linked with increased BOLD responses in the bilateral insula and the dorsal anterior cingulate cortex (dACC), (*A*, *B*) and decreased BOLD-responses in the medial prefrontal cortex (mPFC), precuneus, bilateral hippocampus (HC), bilateral parahippocampal gyrus (PHC), and left angular gyrus (*C*–*E*). Only voxels significant at $P < 0.05$ after whole-brain family-wise error (FWE) correction (peak level) are displayed. Black dots indicate beta estimates from individual participants, while the red line shows the mean beta estimate over all participants. ***$P_{FWE} < 0.001$.

prior knowledge and an alternative mode of memory formation for events that are linked to unexpected outcomes, we further hypothesized that items associated with high negative PEs are particularly well remembered because they alter contributions of three main memory networks: 1) the salience network (represented by anterior insula and dACC; Seeley et al. 2007; Menon 2011; Ham et al. 2013; Metereau and Dreher 2013), 2) the medial-temporal encoding network (represented by bilateral hippocampus and bilateral parahippocampus), and 3) the schema network (represented by mPFC, precuneus, and angular gyrus; van Kesteren et al. 2012; Vogel et al. 2018a). To address this hypothesis, we analyzed functional connectivity within and between these networks depending on PE magnitudes. For this analysis, we defined a separate GLM with eight regressors based on combinations of the following factors: onset type (stimulus vs. outcome), outcome (shocked vs. unshocked), and PE magnitude (low if $|sPE| < 0.5$; high otherwise). After pre-processing the raw times series (see section Materials and Methods), we based our analysis on the implemented network atlas consisting of several ROIs each to compute within- and between-network correlations (Fig. 6A). Here, we focused on the contrast between high and low (negative) PEs at the time when the outcome of each trial was revealed. Results showed significant PE-related changes in the connectivity between large-scale networks. Specifically, for large versus small negative PEs we obtained significantly increased functional connectivity between the salience network and both the schema network ($t(49) = 2.68$, $P_{corr} = 0.030$, $d_{av} = 0.344$) and, at trend level, the medial-temporal encoding network ($t(49) = 2.18$, $P = 0.034$, $P_{corr} = 0.10$, $d_{av} = 0.355$; Fig. 6B); the connectivity between the schema network and the medial-temporal network did not depend on PEs in unshocked trials, $t(49) = 0.29$, $P = 0.773$, $P_{corr} = 1$, $d_{av} = 0.046$). When we correlated the two PE-related increases in between network connectivity with memory, we found that the increase in functional connectivity between the salience and schema networks was relevant for long-term memory formation, as indicated by its significant correlation with improved hit rates for high negative PE items, $r = 0.320$, $t(48) = 2.34$, $P_{corr} = 0.048$ (Fig. 6C); salience-MTEN correlation with hit rates for high negative PE items: $r = 0.147$, $t(48) = 1.03$, $P = 0.31$, $P_{corr} = 0.616$. Furthermore, within-network connectivity

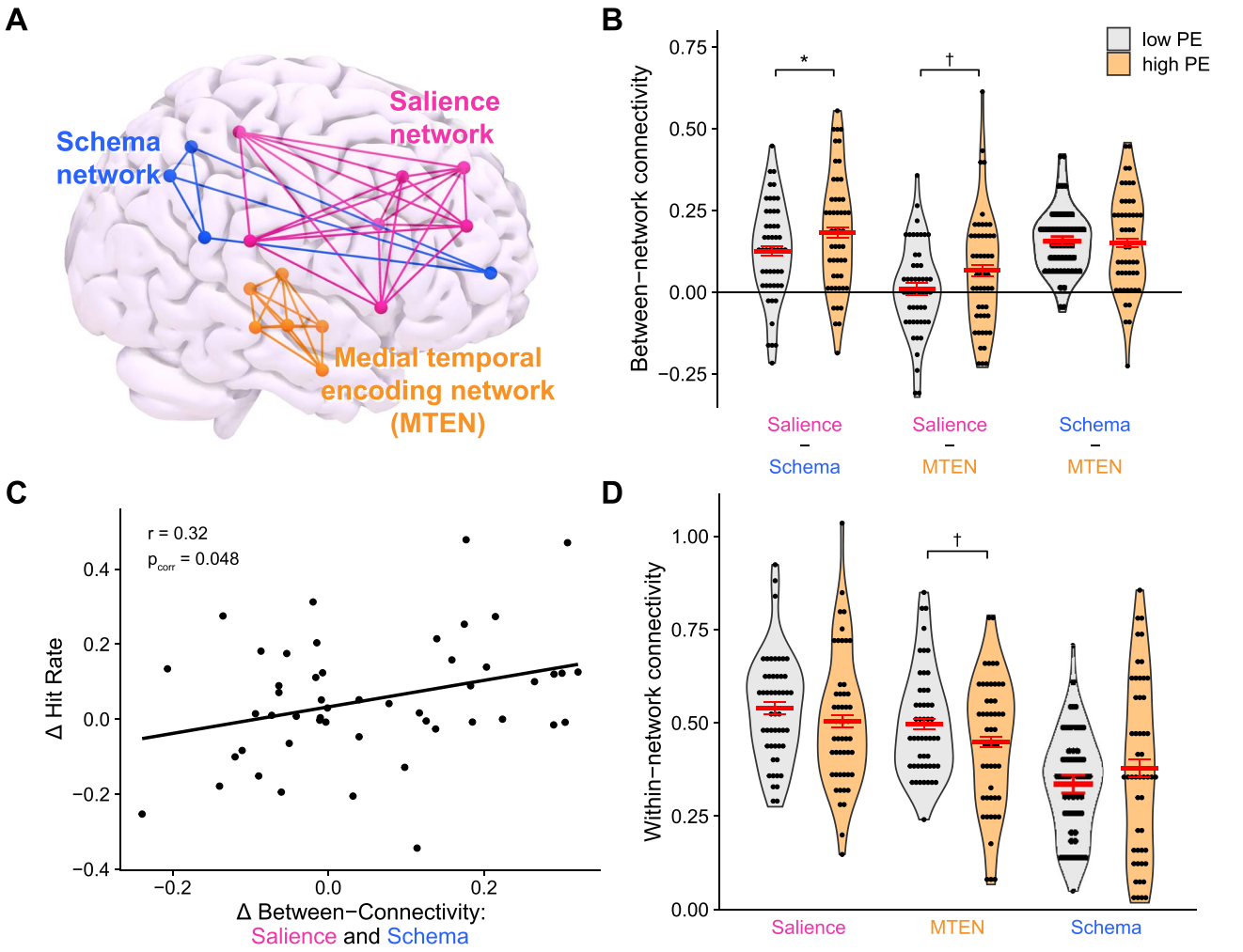

Fig. 6. Negative prediction error magnitude is associated with altered within- and between-network connectivity in memory-relevant networks. (*A*) We investigated PE-associated changes in the activity within and between the salience network (rostral prefrontal cortex, supramaginal gyrus, anterior insula, and dACC), schema network (mPFC, precuneus, and angular gyrus), and medial-temporal encoding network (hippocampus and anterior/posterior parahippocampal gyrus). (*B*) Large (vs. small) PEs were associated with significantly increased cross-network connectivity of the salience network with both the schema-network and the medial-temporal encoding network. Thick red bar represent group means, while thin red bars show ±1 standard error of the mean. (*C*) Increases in functional connectivity between salience network and schema network in response to large negative PEs correlated with greater memory enhancement for large negative PEs. (*D*) Large (vs. small) PEs were associated with significantly decreased within-network functional connectivity in the medial temporal encoding network. †$P < 0.05$, *$P_{corr} < 0.05$.

tended to be decreased for large compared with small negative PEs in the medial-temporal encoding network ($t(49) = 2.44$, $P = 0.018$, $P_{corr} = 0.055$, $d_{av} = 0.307$), but not in the salience network ($t(49) = 1.60$, $P = 0.115$, $P_{corr} = 0.346$, $d_{av} = 0.218$), nor in the schema network ($t(49) = 1.24$, $P = 0.221$, $P_{corr} = 0.664$, $d_{av} = 0.221$; Fig. 6*D*).

Positive PEs are associated with parietal and temporal lobe modulation

So far, our analysis focused on neural underpinnings of the memory-enhancing effects of negative PEs. However, our behavioral findings also pointed to a memory impairment related to positive PEs. To investigate the neural basis of this detrimental effect on memory, we specified parallel models for shocked trials. These revealed that larger positive PEs per se were associated with increased activity in two smaller clusters located in the left superior parietal lobule and the right middle temporal gyrus and decreased activity in the left supramarginal gyrus (see Supplementary Table 3). Again, an analysis of variance inflation factors showed no evidence that multicollinearity influenced results of the critical PE regressor (Supplementary Table 2).

To specifically investigate specific neural activity in response to positive PEs that might underlie their memory decreasing effects, we fitted a univariate fMRI model with onsets of shocked outcomes as a regressor and PEs, the binary subsequent recognition of an item and their interaction as parametric modulators (see section Materials and Methods). As in the parallel model for unshocked trials, our analysis focused on the interaction between PEs and subsequent recognition, as this specific interaction links the processing of PEs with their effects on memory formation. Again, we focused our analysis on the hippocampus and the posterior parahippocampal gyrus. Neither the hippocampus, nor the posterior

parahippocampal gyrus contained any voxels that specifically linked positive PEs with subsequent memory formation (all $P_{SVC} > 0.05$, FWE corrected). Even under a very liberal threshold of $P < 0.001$ (uncorrected), there were no significant voxels in any of the two regions. An additional explorative analysis at whole-brain level further showed no other clusters with increased or decreased levels of activation for the interaction of positive PEs with subsequent memory (all $P_{FWE} > 0.05$).

While prediction uncertainty was negatively associated with subsequent memory in behavioral results, we found no significant clusters that were specifically associated with uncertainty in shocked trials (all $P_{FWE} > 0.05$). For shock expectancy, which had behaviorally been positively linked with memory, we replicated the findings from unshocked trials. Specifically, shock expectancy was only associated with changes in occipital areas, possibly reflecting visual processing of the slider that participants used to give their expectancy rating (Supplementary Table 5). This lack of an overlap between the neural signatures of shock expectancy and positive PEs might be taken as evidence that these two reflect separate cognitive processes, in line with our behavioral findings that PE-effects on memory go beyond mere expectancy effects.

Discussion

For decades, PEs have been known to act as teaching signals in reinforcement learning (Sutton and Barto 1981; Schultz 1998; Cohen 2008). However, it was only rather recently discovered that PEs may shape memory formation for episodes preceding the PE event (Ergo et al. 2020). Here, we combined fMRI with behavioral modeling and large-scale network connectivity analyses to elucidate the mechanisms through which PEs associated with aversive events modulate the formation of long-term memories. Our results provide evidence that negative PEs for aversive events promote memory formation for preceding stimuli through a mechanism that might be distinct from common mechanisms of long-term memory formation. Importantly, the proposed PE-related memory storage mechanism could not be attributed to well-known effects of physiological arousal on memory formation or the effect of a specific prediction itself.

Traditionally, enhanced episodic memory formation has been linked to the medial temporal lobe, including the hippocampus and the parahippocampal gyrus (Reed and Squire 1997; Fernández et al. 1999; Davachi and Wagner 2002; Eichenbaum 2004; Mayes et al. 2007; Shrager et al. 2008). In line with this assumption, we found that activity in the hippocampus and posterior parahippocampal gyrus during stimulus presentation was linked to subsequent memory performance. The negative PE-related memory enhancement, however, was not linked to enhanced but even to decreased medial temporal lobe activity. Further, when participants experienced a negative PE, the connectivity within the medial-temporal encoding network tended to be reduced. While activity in the medial temporal lobe was reduced for negative PEs, we obtained significantly increased activity in the anterior insula and dACC for negative PE events. Both of these regions have previously been implicated in error monitoring, conscious perception of errors, and aversive PE signaling (Taylor et al. 2007; Preuschoff et al. 2008; Ullsperger et al. 2010; Garrison et al. 2013; Bastin et al. 2016; Fazeli and Büchel 2018). Moreover, both the anterior insula and the dACC are key regions of the salience network (Menon 2011; Ham et al. 2013), which signals biologically relevant events and the need for a behavioral or cognitive change (Kerns 2004; Dosenbach et al. 2006). Furthermore, the salience network has been proposed to dynamically change the control of other large-scale networks (Sridharan et al. 2008). In line with this idea, we obtained here a trend for increased functional connectivity between the salience network and the medial-temporal encoding network for negative PEs.

In addition to the negative PE-related decrease in medial temporal activity, there was also a marked decrease in the activity of angular gyrus, precuneus, and mPFC for events associated with negative PEs. Together, these areas form a "schema-network," in which the mPFC is thought to detect a congruency of events with prior knowledge and to then integrate these events into existing knowledge representations (van Kesteren et al. 2012; Vogel et al. 2018b). When the organism experiences large PEs, this indicates that new information conflicts with prior knowledge and should therefore be stored separately from existing schema-congruent memories (van Kesteren et al. 2012). This idea is supported by the obtained negative PE-associated decrease in areas constituting the schema network. Moreover, there was also increased connectivity between the salience network and the schema-network when individuals experienced a negative PE and this PE-related change in large-scale network connectivity was directly correlated with the negative PE-driven memory enhancement. Together these findings suggest that the negative PE-induced enhancement of episodic memory is not driven by an enhancement of common medial temporal mechanisms of memory formation but by a potentially distinct mechanism that is linked to the salience network and separates PE events from experiences that are in line with prior knowledge.

The salience network has often been related to physiological arousal (Xia et al. 2017; Young et al. 2017), which is well known to mediate the superior memory for emotionally arousing events (Cahill and McGaugh 1998; McGaugh 2018). Although one might assume that high negative PEs may have elicited arousal which then enhanced memory storage, our data speak against this alternative and suggest that negative PE-related memory enhancement was not due to increased physiological arousal. First, aversive shocks per se had no influence on memory formation. Even in a combined model featuring

both anticipatory and outcome-related SCRs in addition to PEs, we still found clear evidence for complementary effects of PEs beyond these arousal measures. Importantly, specific neural clusters associated with negative PEs were identified in a model that controlled for physiological arousal. These results indicate that the effects of PEs on episodic memory formation cannot be explained by traditional arousal-based models. One could speculate whether this dissociation is also reflected in a different neuroendocrinological basis of arousal- versus PE-based effects. While the effects of arousal on memory formation have been strongly linked to the action of noradrenaline (Cahill and McGaugh 1998; Strange and Dolan 2004), PEs in both the reward and aversive domain are typically associated with dopamine (Schultz et al. 1997; Papalini et al. 2020). Future studies could test this account through pharmacological manipulations.

Another challenge is separating specific effects of PEs from those of mere shock expectancy. This is complicated by the fact that PEs incorporate information about both shock predictions and actual outcomes, making them conceptually close to an interaction between these two. The presence of such interaction effects in regression models often comes with high collinearity, which makes estimated regression coefficients less stable by inflating standard errors (Echambadi and Hess 2007). However, we found consistent evidence for memory-modulating effects of PEs under various circumstances. First, the memory-modulating effects our PEs could be detected in a simple model featuring them as the sole predictor, which could therefore not have been affected by collinearity. Further, adding our PE-regressor to a model comprised of several control variables significantly improved the model fit and confirmed the assumed S-shaped relation to memory formation. This speaks against an alternative account of our findings in which the mere prediction of an aversive event, possibly through increased attention to the predictive stimulus, is sufficient to explain our observed effects on memory formation.

It is also important to note that our findings go above and beyond previous results showing an enhanced memory for novel or surprising stimuli (Strange and Dolan 2004; Cycowicz and Friedman 2007). We show here that, rather than the novelty of a stimulus, the discrepancy between expected and experienced consequences of a stimulus affected its memorability. This is particularly remarkable as these consequences were only revealed after a stimulus had already disappeared, thus ruling out a simple increase of attentional processing.

Previous behavioral findings could not differentiate effects of negative and positive PEs in an aversive context (Kalbe and Schwabe 2020) and studies on the role of reward-related PEs yielded inconsistent findings as to whether the direction of the PE matters for episodic memory formation (Rouhani et al. 2018; Jang et al. 2019; Ergo et al. 2020). Interestingly, we found that memory effects depended on the sign of PEs, with negative PEs being associated with better recognition performance and larger positive PEs showing opposite, negative effects on recognition performance.

The neural signature of positive PEs was clearly distinct from the neural underpinnings of negative PEs. Positive PEs were associated with clusters of increased activation in the left superior parietal lobule and the right middle temporal gyrus and decreased activation of the left supramarginal gyrus. The superior parietal lobe has been linked to internal representations of sensory inputs before (Wolpert et al. 1998) as well as to contralateral sensorimotor coding of body parts (Wolbers 2003). As the electric shock was applied to the right leg and increased superior marginal activation was observed in the left hemisphere, the observed activity pattern might point to increased processing of the electric shock. Furthermore, the supramarginal gyrus has been previously associated with motor planning (Potok et al. 2019) and unexpected somatosensory feedback perturbation (Golfinopoulos et al. 2011). Thus, it is tempting to speculate that positive PEs resulted in more pronounced processing of the (unexpected) electric shock, which distracted from the mnemonic processing of the encoded stimulus and hence led to decreased subsequent recognition memory.

Closely related but conceptually distinct from PEs is prediction uncertainty. While PEs only become apparent after an outcome has been revealed, uncertainty emerges as soon as a potentially threatening stimulus is presented. We found that uncertainty about the possible occurrence of a shock was associated with decreased recognition performance. At the neural level, uncertainty was paralleled by decreased activation in bilateral medial occipital areas, possibly reflecting diminished visual processing of stimuli associated with uncertain outcomes, which might explain the uncertainty-related impairment in recognition. In addition, uncertainty was associated with reduced mPFC activation, a region implicated in beliefs and the inference of hidden states (Yoshida and Ishii 2006; Starkweather et al. 2018).

In summary, we provide behavioral and neural evidence for a critical impact of aversive PEs on long-term memory formation for events preceding the PE, thereby bridging the traditionally separated fields of associative learning and declarative long-term memory. In addition to the magnitude of the PE, our results show that the direction of the PE affects memory formation. Whereas positive PEs reduced subsequent memory, negative PEs promoted memory formation. In particular for negative PEs, our results suggest a qualitative shift in the contributions of large-scale neural networks to memory formation. Negative PEs reduced the processing of events in the schema network and the medial-temporal encoding network both of which are involved in "standard" long-term memory formation. Instead, such schema-incongruent experiences might be particularly well remembered because they are encoded

distinctly from more mundane experiences, perhaps at an exemplar-level, in a process that is likely mediated through the salience network. Importantly, these memory enhancements and related neural changes could not be explained by the prediction itself or mere changes in physiological arousal, thus pointing to a rather "cognitive" mechanism of memory enhancement. Although the salience and front-parietal network changes related to negative PEs might be considered as an extension of established ideas about episodic memory formation, in particular the opposite changes in MTL activity related to memory formation at stimulus onset versus PE occurrence suggests that potentially distinct mechanisms might be involved in memory for high PE events. These findings may have relevant implications for the treatment of fear-related mental disorders, suggesting that it might be beneficial to explicitly activate patients' negative outcome expectations prior to the exposure to the feared stimulus, as the absence of the feared consequence in the therapeutic context should produce strong fear-incongruent memories. More generally, our results provide novel insights into the mechanisms underlying the exceptional memory for episodes in the context of unexpected events, such as meeting Barack Obama in the supermarket.

Supplementary material

Supplementary material can be found at *Cerebral Cortex* online.

Notes

We gratefully acknowledge the support of Friederike Baier, Jan-Ole Großmann, Vincent Kühn, and Ricarda Vielhauer during data collection. *Conflict of Interest:* The authors declare no competing financial interests.

Funding

Universität Hamburg.

Author contributions

Conceptualization, F.K. and L.S.; Methodology, F.K. and L.S.; Formal analysis, F.K.; Investigation, F.K.; Writing – Original Draft, F.K. and L.S.; Writing –Review and Editing, F.K. and L.S.; Funding Acquisition, L.S.; Resources, L.S.; Supervision, L.S.

Data availability

Behavioral, SCR, and fMRI data that support the findings of this study are available at OSF: https://osf.io/3atyr/.

Code availability

Custom code used to analyze and model the data is available at OSF: https://osf.io/3atyr/.

References

Alvarez P, Squire LR. 1994. Memory consolidation and the medial temporal lobe: a simple network model. *Proc Natl Acad Sci*. 91(15): 7041–7045. 10.1073/pnas.91.15.7041.

Barr DJ, Levy R, Scheepers C, Tily HJ. 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. *J Mem Lang*. 68(3):255–278. 10.1016/j.jml.2012.11.001.

Bastin J, Deman P, David O, Gueguen M, Benis D, Minotti L, Hoffman D, Combrisson E, Kujala J, Perrone-Bertolotti M, et al. 2016. Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cereb Cortex*. 27(2): 1545–1557. 10.1093/cercor/bhv352.

Bates D, Mächler M, Bolker B, Walker S. 2015. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 67(1):1–48. 10.18637/jss.v067.i01.

Benedek M, Kaernbach C. 2010. A continuous measure of phasic electrodermal activity. *J Neurosci Methods*. 190(1):80–91. 10.1016/j.jneumeth.2010.04.028.

Cahill L, McGaugh JL. 1998. Mechanisms of emotional arousal and lasting declarative memory. *Trends Neurosci*. 21(7):294–299. 10.1016/S0166-2236(97)01214-9.

Cavanaugh JE, Neath AA. 2019. The Akaike information criterion: background, derivation, properties, application, interpretation, and refinements. *Wiley Interdiscip Rev Comput Stat*. 11(3):e1460. 10.1002/wics.1460.

Cohen MX. 2008. Neurocomputational mechanisms of reinforcement-guided learning in humans: a review. *Cogn Affect Behav Neurosci*. 8(2):113–125. 10.3758/CABN.8.2.113.

Cycowicz YM, Friedman D. 2007. Visual novel stimuli in an ERP novelty oddball paradigm: effects of familiarity on repetition and recognition memory. *Psychophysiology*. 44(1):11–29. https://doi.org/10.1111/j.1469-8986.2006.00481.x.

Davachi L, Wagner AD. 2002. Hippocampal contributions to episodic encoding: insights from relational and item-based learning. *J Neurophysiol*. 88(2):982–990. 10.1152/jn.2002.88.2.982.

Delgado MR, Li J, Schiller D, Phelps EA. 2008. The role of the striatum in aversive learning and aversive prediction errors. *Philos Trans R Soc B Biol Sci*. 363(1511):3787–3800. 10.1098/rstb.2008.0161.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, et al. 2006. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*. 31(3):968–980. 10.1016/j.neuroimage.2006.01.021.

Dosenbach NUF, Visscher KM, Palmer ED, Miezin FM, Wenger KK, Kang HC, Burgund ED, Grimes AL, Schlaggar BL, Petersen SE. 2006. A core system for the implementation of task sets. *Neuron*. 50(5): 799–812. 10.1016/j.neuron.2006.04.031.

Echambadi R, Hess JD. 2007. Mean-centering does not alleviate collinearity problems in moderated multiple regression models. *Marketing Science*. 26(3):438–445. 10.1287/mksc.1060.0263.

Eichenbaum H. 2001. The hippocampus and declarative memory: cognitive mechanisms and neural codes. *Behav Brain Res*. 127 (1–2):199–207. 10.1016/S0166-4328(01)00365-5.

Eichenbaum H. 2004. Hippocampus: cognitive processes and neural representations that underlie declarative memory. *Neuron*. 44(1): 109–120. 10.1016/j.neuron.2004.08.028.

Ergo K, De Loof E, Verguts T. 2020. Reward prediction error and declarative memory. *Trends Cogn Sci*. 24(5):388–397. 10.1016/j.tics.2020.02.009.

Fazeli S, Büchel C. 2018. Pain-related expectation and prediction error signals in the anterior insula are not related to

aversiveness. *J Neurosci*. 38(29):6461–6474. https://doi.org/10.1523/JNEUROSCI.0671-18.2018.

Fernández G, Effern A, Grunwald T, Pezer N, Lehnertz K, Dümpelmann M, Van Roost D, Elger CE. 1999. Real-time tracking of memory formation in the human rhinal cortex and hippocampus. *Science*. 285(5433):1582–1585. 10.1126/science.285.5433.1582.

Figner B, Murphy RO. 2011. Using skin conductance in judgment and decision making research. In: Schulte-Mecklenbeck M, Kuehberger A, Ranyard R, editors. *A handbook of process tracing methods for decision research: A critical review and user's guide*. New York, NY: Psychology Press, pp. 163–184.

Fouragnan E, Retzler C, Philiastides MG. 2018. Separate neural representations of prediction error valence and surprise: evidence from an fMRI meta-analysis. *Hum Brain Mapp*. 39(7):2887–2906. 10.1002/hbm.24047.

Garrison J, Erdeniz B, Done J. 2013. Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies. *Neurosci Biobehav Rev*. 37(7):1297–1310. 10.1016/j.neubiorev.2013.03.023.

Gershman SJ, Daw ND. 2017. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu Rev Psychol*. 68(1):101–128. 10.1146/annurev-psych-122414-033625.

Ghosh VE, Gilboa A. 2014. What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*. 53:104–114. 10.1016/j.neuropsychologia.2013.11.010.

Glimcher PW. 2011. Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proc Natl Acad Sci*. 108(Supplement_3):15647–15654. 10.1073/pnas.1014269108.

Golfinopoulos E, Tourville JA, Bohland JW, Ghosh SS, Nieto-Castanon A, Guenther FH. 2011. fMRI investigation of unexpected somatosensory feedback perturbation during speech. *Neuroimage*. 55(3):1324–1338. 10.1016/j.neuroimage.2010.12.065.

Green P, MacLeod CJ. 2016. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol Evol*. 7(4):493–498. 10.1111/2041-210X.12504.

Greve A, Cooper E, Kaula A, Anderson MC, Henson R. 2017. Does prediction error drive one-shot declarative learning? *J Mem Lang*. 94:149–165. 10.1016/j.jml.2016.11.001.

Ham T, Leff A, de Boissezon X, Joffe A, Sharp DJ. 2013. Cognitive control and the salience network: an investigation of error processing and effective connectivity. *J Neurosci*. 33(16):7091–7098. 10.1523/JNEUROSCI.4692-12.2013.

Henson RN, Gagnepain P. 2010. Predictive, interactive multiple memory systems. *Hippocampus*. 20(11):1315–1326. 10.1002/hipo.20857.

Hermans EJ, Battaglia FP, Atsak P, de Voogd LD, Fernández G, Roozendaal B. 2014. How the amygdala affects emotional memory by altering brain network properties. *Neurobiol Learn Mem*. 112:2–16. 10.1016/j.nlm.2014.02.005.

Jang AI, Nassar MR, Dillon DG, Frank MJ. 2019. Positive reward prediction errors during decision-making strengthen memory encoding. *Nat Hum Behav*. 3(7):719–732. 10.1038/s41562-019-0597-3.

Kalbe F, Schwabe L. 2020. Beyond arousal: prediction error related to aversive events promotes episodic memory formation. *J Exp Psychol Learn Mem Cogn*. 46(2):234–246. 10.1037/xlm0000728.

Kerns JG. 2004. Anterior cingulate conflict monitoring and adjustments in control. *Science*. 303(5660):1023–1026. 10.1126/science.1089910.

van Kesteren MTR, Ruiter DJ, Fernández G, Henson RN. 2012. How schema and novelty augment memory formation. *Trends Neurosci*. 35(4):211–219. 10.1016/j.tins.2012.02.001.

Matsumoto M, Hikosaka O. 2009. Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*. 459(7248):837–841. 10.1038/nature08028.

Mayes A, Montaldi D, Migo E. 2007. Associative memory and the medial temporal lobes. *Trends Cogn Sci*. 11(3):126–135. 10.1016/j.tics.2006.12.003.

McGaugh JL. 2018. Emotional arousal regulation of memory consolidation. *Curr Opin Behav Sci*. 19:55–60. 10.1016/j.cobeha.2017.10.003.

McGaugh JL, Roozendaal B. 2002. Role of adrenal stress hormones in forming lasting memories in the brain. *Curr Opin Neurobiol*. 12(2):205–210. 10.1016/S0959-4388(02)00306-9.

McHugh SB, Barkus C, Huber A, Capitao L, Lima J, Lowry JP, Bannerman DM. 2014. Aversive prediction error signals in the amygdala. *J Neurosci*. 34(27):9024–9033. 10.1523/JNEUROSCI.4465-13.2014.

Menon V. 2011. Large-scale brain networks and psychopathology: a unifying triple network model. *Trends Cogn Sci*. 15(10):483–506. 10.1016/j.tics.2011.08.003.

Metereau E, Dreher J-C. 2013. Cerebral correlates of salient prediction error for different rewards and punishments. *Cereb Cortex*. 23(2):477–487. 10.1093/cercor/bhs037.

Mizumori SJY. 2013. Context prediction analysis and episodic memory. *Front Behav Neurosci*. 7:132. 10.3389/fnbeh.2013.00132.

Niv Y. 2009. Reinforcement learning in the brain. *J Math Psychol*. 53(3):139–154. 10.1016/j.jmp.2008.12.005.

Papalini S, Beckers T, Vervliet B. 2020. Dopamine: from prediction error to psychotherapy. *Transl Psychiatry*. 10(1):164. 10.1038/s41398-020-0814-x.

Potok W, Maskiewicz A, Króliczak G, Marangon M. 2019. The temporal involvement of the left supramarginal gyrus in planning functional grasps: a neuronavigated TMS study. *Cortex*. 111:16–34. 10.1016/j.cortex.2018.10.010.

Preuschoff K, Quartz SR, Bossaerts P. 2008. Human insula activation reflects risk prediction errors as well as risk. *J Neurosci*. 28(11):2745–2752. 10.1523/JNEUROSCI.4286-07.2008.

Reed JM, Squire LR. 1997. Impaired recognition memory in patients with lesions limited to the hippocampal formation. *Behav Neurosci*. 111(4):667–675. 10.1037//0735-7044.111.4.667.

Rescorla RA, Wagner AR. 1972. A theory of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In: *Classical conditioning ii: current research and theory*. New York: Appleton-Century-Crofts, pp. 64–99.

Richardson MP, Strange BA, Dolan RJ. 2004. Encoding of emotional memories depends on amygdala and hippocampus and their interactions. *Nat Neurosci*. 7(3):278–285. 10.1038/nn1190.

Rouhani N, Norman KA, Niv Y. 2018. Dissociable effects of surprising rewards on learning and memory. *J Exp Psychol Learn Mem Cogn*. 44(9):1430–1443. 10.1037/xlm0000518.

Rouhani N, Norman KA, Niv Y, Bornstein AM. 2020. Reward prediction errors create event boundaries in memory. *Cognition*. 203:104269. 10.1016/j.cognition.2020.104269.

Schultz W. 1998. Predictive reward signal of dopamine neurons. *J Neurophysiol*. 80(1):1–27. 10.1152/jn.1998.80.1.1.

Schultz W, Dayan P, Montague PR. 1997. A neural substrate of prediction and reward. *Science*. 275(5306):1593–1599. 10.1126/science.275.5306.1593.

Seeley WW, Menon V, Schatzberg AF, Keller J, Glover GH, Kenna H, Reiss AL, Greicius MD. 2007. Dissociable intrinsic connectivity networks for salience processing and executive control. *J Neurosci*. 27(9):2349–2356. 10.1523/JNEUROSCI.5587-06.2007.

Shohamy D, Adcock RA. 2010. Dopamine and adaptive memory. *Trends Cogn Sci*. 14(10):464–472. 10.1016/j.tics.2010.08.002.

Shrager Y, Kirwan CB, Squire LR. 2008. Activity in both hippocampus and Perirhinal cortex predicts the memory strength of subsequently remembered information. *Neuron*. 59(4):547–553. 10.1016/j.neuron.2008.07.022.

Sridharan D, Levitin DJ, Menon V. 2008. A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proc Natl Acad Sci*. 105(34):12569–12574. 10.1073/pnas.0800005105.

Starkweather CK, Gershman SJ, Uchida N. 2018. The medial prefrontal cortex shapes dopamine reward prediction errors under state uncertainty. *Neuron*. 98(3):616–629.e6. 10.1016/j.neuron.2018.03.036.

Strange BA, Dolan RJ. 2004. Adrenergic modulation of emotional memory-evoked human amygdala and hippocampal responses. *Proc Natl Acad Sci*. 101(31):11454–11458. 10.1073/pnas.0404282101.

Summerfield C, Egner T. 2009. Expectation (and attention) in visual cognition. *Trends Cogn Sci*. 13(9):403–409. 10.1016/j.tics.2009.06.003.

Sutton RS, Barto AG. 1981. Toward a modern theory of adaptive networks: expectation and prediction. *Psychol Rev*. 88(2):135–170. 10.1037/0033-295X.88.2.135.

Taylor SF, Stern ER, Gehring WJ. 2007. Neural systems for error monitoring: recent findings and theoretical perspectives. *Neuroscientist*. 13(2):160–172. 10.1177/1073858406298184.

Ullsperger M, Harsay HA, Wessel JR, Ridderinkhof KR. 2010. Conscious perception of errors and its relation to the anterior insula. *Brain Struct Funct*. 214(5–6):629–643. 10.1007/s00429-010-0261-1.

Vogel S, Kluen LM, Fernández G, Schwabe L. 2018a. Stress leads to aberrant hippocampal involvement when processing schema-related information. *Learn Mem*. 25(1):21–30. 10.1101/lm.046003.117.

Vogel S, Kluen LM, Fernández G, Schwabe L. 2018b. Stress affects the neural ensemble for integrating new information and prior knowledge. *Neuroimage*. 173:176–187. 10.1016/j.neuroimage.2018.02.038.

Whitfield-Gabrieli S, Nieto-Castanon A. 2012. Conn: a functional connectivity toolbox for correlated and anticorrelated brain networks. *Brain Connect*. 2(3):125–141. 10.1089/brain.2012.0073.

Wolbers T. 2003. Contralateral coding of imagined body parts in the superior parietal lobe. *Cereb Cortex*. 13(4):392–399. 10.1093/cercor/13.4.392.

Wolpert DM, Goodbody SJ, Husain M. 1998. Maintaining internal representations: the role of the human superior parietal lobe. *Nat Neurosci*. 1(6):529–533. 10.1038/2245.

Xia C, Touroutoglou A, Quigley KS, Feldman Barrett L, Dickerson BC. 2017. Salience network connectivity modulates skin conductance responses in predicting arousal experience. *J Cogn Neurosci*. 29(5):827–836. 10.1162/jocn_a_01087.

Yoshida W, Ishii S. 2006. Resolution of uncertainty in prefrontal cortex. *Neuron*. 50(5):781–789. 10.1016/j.neuron.2006.05.006.

Young CB, Raz G, Everaerd D, Beckmann CF, Tendolkar I, Hendler T, Fernández G, Hermans EJ. 2017. Dynamic shifts in large-scale brain network balance as a function of arousal. *J Neurosci*. 37(2):281–290. 10.1523/JNEUROSCI.1759-16.2016.