American Psychological Association

Journal of Experimental Psychology:
Learning, Memory, and Cognition

© 2019 American Psychological Association
0278-7393/20/$12.00

2020, Vol. 46, No. 2, 234–246
http://dx.doi.org/10.1037/xlm0000728

Beyond Arousal: Prediction Error Related to Aversive Events Promotes Episodic Memory Formation

Felix Kalbe and Lars Schwabe
University of Hamburg

Stimuli encoded shortly before an aversive event are typically well remembered. Traditionally, this emotional memory enhancement has been attributed to beneficial effects of physiological arousal on memory formation. Here, we proposed an additional mechanism and tested whether memory formation is driven by the unpredictable nature of aversive events (i.e., aversive prediction errors). In a combined Pavlovian fear conditioning and incidental memory paradigm, participants saw initially neutral pictures from 2 distinct categories, 1 of which was associated with a risk to receive an electric shock. During encoding, we measured both physiological arousal and explicit prediction errors to explain memory differences in a surprise recognition test that followed approximately 24 hr later. In a first experiment, we show that physiological arousal, expressed as outcome-related skin conductance responses, was associated with improved recognition memory, corroborating arousal-based models. Critically, unsigned binary prediction errors derived from explicit shock expectancy ratings in each trial were also linked to enhanced recognition and model fits showed that the impact of prediction errors on memory was dissociable from the impact of arousal. In a second experiment, we replicated and extended the findings of the first experiment by demonstrating that the memory-promoting effect of prediction errors remained even after controlling for arousal. The present data point to prediction error-related learning as a cognitive mechanism that contributes to the emotional enhancement of memory, above and beyond the well-established effects of arousal in emotional memory formation.

Keywords: episodic memory, prediction errors, arousal, associative learning, fear conditioning

Information that is encoded within close temporal proximity to an aversive event is typically well remembered (Cahill & McGaugh, 1998; Christianson & Loftus, 1987; Christianson, Loftus, Hoffman, & Loftus, 1991; LaBar & Cabeza, 2006; Schwabe, Joëls, Roozendaal, Wolf, & Oitzl, 2012). Although generally adaptive as it might help to avoid threatening situations in the future (Nairne, Thompson, & Pandeirada, 2007), the superior memory for stimuli encoded around the time of an aversive event may also contribute to fear-related psychopathologies such as phobia or posttraumatic stress disorder (de Quervain, Schwabe, & Roozendaal, 2017; Dunsmoor & Paz, 2015; Pitman, 1989).

The enhanced memory for information linked to an aversive event is exemplified by Pavlovian fear conditioning, in which an initially neutral conditioned stimulus (conditional stimulus [CS]$^+$) precedes an aversive unconditioned stimulus (UCS; LaBar & Cabeza, 2006; Maren, 2001). Several studies demonstrated that subsequent memory for the CS$^+$ is much better than for another stimulus (CS$^-$) that was also repeatedly presented but never paired with the UCS (Dunsmoor, Murty, Davachi, & Phelps, 2015; Schwarze, Bingel, & Sommer, 2012). Recent evidence shows that the memory boosting effect of aversive events is not limited to individual items but might also operate at the category level. For example, when several pictures of one category (e.g., animals, CS$^+$) were followed by an aversive shock, even nonshocked pictures from that category were better remembered in a subsequent surprise memory test compared with pictures from a nonshocked control category (e.g., tools, CS$^-$; Dunsmoor et al., 2015).

Classically, the emotional enhancement of memory in general and the superior memory for CS$^+$ versus CS$^-$ items, in particular, has been attributed to the physiological arousal that is elicited by aversive stimuli such as the CS$^+$ in fear learning (Cahill, Prins, Weber, & McGaugh, 1994; LaBar & Cabeza, 2006; McGaugh, 2018; Schwarze et al., 2012). More specifically, aversive experiences are well-known to prompt the secretion of catecholamines, including the release of adrenaline and noradrenaline (Joëls & Baram, 2009). In the periphery, adrenergic arousal is reflected, for instance, in increased skin conductance responses (SCRs). At the brain level, adrenergic arousal increases the activity of the basolateral amygdala, which then strengthens memory formation processes in other areas such as the hippocampus (LaBar & Cabeza, 2006; McGaugh & Roozendaal, 2002; Pape & Pare, 2010; Phelps, 2004).

This article was published Online First June 6, 2019.

Felix Kalbe and Lars Schwabe, Institute of Psychology, University of Hamburg.

Lars Schwabe received funding from the German Research Foundation (DFG) in the context of the collaborative research center “Fear, Anxiety, Anxiety Disorders” (TRR 58) and from the Landesforschungsfoerderung Hamburg (LFF FV-38). We gratefully acknowledge the assistance of Leandra Feldhusen, Vincent Kühn, and Moana Lamm during data collection.

Correspondence concerning this article should be addressed to Lars Schwabe, Department of Cognitive Psychology, University of Hamburg, Hamburg 20146, Germany. E-mail: lars.schwabe@uni-hamburg.de

While the role of physiological arousal in the enhanced memory for items encoded shortly before aversive events is well documented, there may still be other mechanisms contributing to this memory enhancement. In particular, aversive events are often unpredictable in nature and characterized by a discrepancy between expectations and outcomes, so-called *prediction errors*. The Rescorla-Wagner model (Rescorla & Wagner, 1972), a classic model in the domain of associative learning, describes how prediction error signals can prompt learning. At the core of this model, the strength of association between a CS and a UCS is updated iteratively after each trial through a prediction error that is weighted by both the salience of the CS and a learning rate parameter linked to the UCS (Walkenbach & Haddad, 1980). The prediction error is formalized as the difference between the actual US presented in a given trial and the summed predicted values of all the cues present on this trial (Miller, Barnet, & Grahame, 1995). Mathematically, this surprise signal is obtained by subtracting the expected signal from the observed outcome signal. The prediction error is therein conceptualized as a continuous variable, meaning that prediction errors can differ in magnitude depending on the extent to which observed and predicted outcomes differ. In the Rescorla-Wagner model, prediction errors are also treated as a signed variable, meaning that they will be negative when expectations exceed observed outcomes for the given trial and positive when outcomes exceed expectations.

Various basic cognitive domains, such as visual processing (Hosoya, Baccus, & Meister, 2005; Rao & Ballard, 1999), auditory processing (Baldeweg, 2006; Smith & Lewicki, 2006), and attention (Feldman & Friston, 2010; Spratling, 2008) have been demonstrated to involve top-down predictions that are matched against sensory input signals (Wacongne et al., 2011). In the domain of reinforcement learning, reward prediction errors are used to update state-action values, allowing agents to choose optimal actions by updating their internal models of complex environments (Hollerman & Schultz, 1998; Maia, 2009; Schultz, 2000; Schultz, Dayan, & Montague, 1997). The widespread evidence for predictive coding in various domains has led some authors to suggest that forming predictions might be one fundamental principle of neural computation in the brain (Bubic, von Cramon, & Schubotz, 2010; Clark, 2013; Friston, 2010).

More recently, prediction errors have been reconceptualized as general teaching signals (Bar, 2007; Clark, 2013) that may enhance memory for ongoing aversive events (Trapp, O'Doherty, & Schwabe, 2018). This is based on the notion that aversive events, besides the physiological arousal that they induce, can be characterized by their unpredictability (de Berker et al., 2016). Thus, they are linked to prediction errors that may be interpreted as evidence that an agent's present model of the environment is insufficient or that necessary information is missing. Prediction errors may, presumably through their effects on the dopaminergic system (Schultz & Dickinson, 2000; Shohamy & Adcock, 2010), promote a state that enables rapid learning of ongoing events. According to this view, it might be hypothesized that the enhanced memory for stimuli that precede an aversive event is at least partly due to the prediction error associated with this event. Indeed, there is first evidence from reward learning suggesting that prediction errors might promote episodic memory formation in humans (Jang, Nassar, Dillon, & Frank, 2018; Rouhani, Norman, & Niv, 2018). Similarly, a recent study manipulated both participants' prior expectations and following evidence to actively control prediction errors and found that these prediction errors led to improved one-shot declarative learning (Greve, Cooper, Kaula, Anderson, & Henson, 2017).

However, to date it remains completely unknown whether prediction errors may contribute to the enhanced memory for information linked to aversive events and, even more importantly, whether the putative contribution of a prediction error to the superior memory for events encoded shortly before an aversive event goes beyond the impact of physiological arousal on memory.

Thus, we aimed here to determine the role of prediction errors, above and beyond physiological arousal, in the superior memory for stimuli that precede aversive events. In two experiments, we asked participants to predict the occurrence of aversive electric shocks in a combined Pavlovian fear conditioning and incidental memory encoding paradigm. In this task, unique pictures of exemplars from two categories (animals and tools) were presented. Pictures from one of the two categories were followed by an electric shock with a probability of two thirds, while pictures from the remaining category were never followed by a shock. In each trial, participants predicted the occurrence of a shock in a forced-choice fashion, while we measured SCRs as indicators of physiological arousal. Therefore, we collected data on both prediction errors and physiological arousal during encoding. Memory for the previously presented pictures was assessed in a surprise recognition test about 24 hr later. We hypothesized that recognition performance would be enhanced for pictures that were linked with incorrect shock predictions and that these memory advantages could not be fully explained by the increased physiological arousal elicited by the aversive event.

Experiment 1

Experiment 1 was designed to test the role of prediction errors in episodic memory formation in the context of aversive fear conditioning. Specifically, we aimed to investigate whether prediction errors can explain memory advantages for events associated with aversive stimuli beyond the well-known memory effects of physiological arousal on subsequent remembering. To this end, participants completed an incidental memory task in which they were instructed to predict whether a picture would be followed by an electric shock, while we recorded SCRs as a physiological measure of arousal.

Method

Participants. Forty-four healthy men and women between 19 and 33 years of age ($M = 25.05$, $SD = 3.75$) participated in this experiment. This sample size was based on an a priori sample size calculation with the software G*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) to achieve a statistical power of .90 to detect a medium sized effect ($d_z = 0.5$) using a two-tailed dependent means t test at $\alpha = .05$. Exclusion criteria comprised any current physical or mental illness, life-time history of any neurological disorder, electronic medical devices such as pacemakers, and pregnancy in women. Each participant gave written informed consent before testing and received a monetary compensation of 20€. Ethical approval for the study protocol was obtained from the ethics committee of the Faculty of Psychology and Human Movement Sciences of the University of Hamburg.

Materials. Stimuli were 180 color pictures of animals and 180 color pictures of tools isolated on white backgrounds. Pictures were acquired from the Bank of Standardized Stimuli (BOSS; Brodeur, Dionne-Dostie, Montreuil, & Lepage, 2010; Brodeur, Guérard, & Bouras, 2014) as well as from publicly available Internet sources. All pictures were chosen to be of neutral valence, to avoid ceiling effects in memory performance and any interference between stimulus-related arousal on the one hand and prediction errors or arousal induced by the aversive event on the other hand. They were selected to be unique exemplars of their respective category. For example, there were not two pictures of different dogs or two pictures of different hammers. Sixty pictures were used during the learning session on experimental Day 1 and 120 pictures for encoding tasks that were unrelated to the purpose of the present study and took place before or after the learning session. More important, this task did not feature any aversive events, nor were participants asked to make any predictions. The remaining 180 stimuli were used as lures during the recognition test. The order in which individual items were presented was randomized across participants. Likewise, the allocation of each stimulus as either learning item or lure was randomized per participant.

Procedure. The experiment took place on two consecutive days, with encoding session on experimental Day 1 and the test session on experimental Day 2. Upon arrival at the lab on experimental Day 1, participants gave written informed consent and completed a demographic questionnaire. They then received written instructions that they were going to see a series of pictures of animals and tools and that some pictures would be followed by a brief electric shock after the picture had disappeared. Participants were instructed to try to predict whether a shock would be following the current picture. We did not inform participants about the underlying shock contingencies, but participants should learn these by trial and error, using the electric shocks as feedback to improve their predictions (see Figure 1). Participants were not informed about the subsequent memory test for the shown pictures.

To measure SCRs as indicators of physiological arousal and conditioned fear, electrodes were placed on the distal phalanx of the second and third finger of the left hand. Skin conductance was measured using the MP-160 BIOPAC system (BIOPAC systems, Goleta, CA). For electrical stimulation, we used the STM-200 module connected to the MP-160. A stimulation electrode was placed on the right lower leg, approximately 25 cm centrally above the heel. Stimulation intensity was adjusted individually to be unpleasant but not painful using a standardized procedure. More specifically, a total of twelve 200 ms single pulse shocks were administered, with an initial intensity of 20 V. After each trial, participants rated whether the shock had been painful in a forced-choice fashion using the "left" ("not painful") and "right" ("painful") keys. Whenever participants rated the shock as not painful, its intensity for the next trial was increased slightly. Analogous, when participants rated the shock as painful, it was decreased slightly.

During the encoding session, 30 pictures of animals and 30 pictures of tools were presented in a pseudorandomized order so that no more than three pictures from the same category appeared in a row. Each picture was presented only once. In each trial, a picture from one of the two categories was presented centrally on a computer screen for 4.5 s, during which participants were requested to make their binary prediction whether an electric shock was going to follow using the left and right arrow keys on the computer keyboard. A 200 ms electric shock with the intensity determined for a participant before (see above) was presented immediately after the offset of some of the pictures. Critically, shock contingencies were linked to item categories (i.e., tools vs. animals). For each participant, one of the two item categories was randomly determined to be the CS^+ category, while the other served as the CS^- category. Which stimulus category served as CS^+ and CS^-, respectively, was counterbalanced across participants. For each CS^+ trial, the probability of a 200 ms single-pulse shock was two thirds, so that 20 out of 30 CS^+ trials were followed by a shock. In the 30 trials that featured images from the CS^- category, no shocks were administered. Between pictures, a black fixation cross was presented centrally on a white background with a variable duration of 8 ± 2 s, which allowed us to measure the relatively slow SCRs elicited by the pictures and the electric shock. After completion of the conditioning phase, electrodes were removed, and participants rated the intensity of shocks on a scale from 1 (*not unpleasant at all*) to 10 (*extremely unpleasant*).

On experimental Day 2, 22 to 26 hr after the encoding session, participants returned for a surprise recognition test. First, they completed a short questionnaire to assess whether they anticipated a memory test and then rated how surprised they were about the recognition test on a scale from 1 (*not surprised at all*) to 5 (*very surprised*). Next, they received written instructions explaining details of the following recognition test. During the recognition test, participants were presented all pictures they had seen on experimental Day 1 (90 pictures of animals, 90 pictures of tools) as well as 180 "new" pictures (90 pictures of animals, 90 pictures of tools) that had not been presented on the previous day. Each trial started with a central black fixation cross on a white background for 1.5 ± 0.5 s, followed by an "old" or "new" picture presented centrally on the computer screen. For each item, participants made a two-staged forced-choice decision. First, participants had 5 s to indicate whether the currently presented picture was old (presented on the previous day) or new (not presented before) using the left and right arrow keys, respectively. Then, participants had 5 s to indicate how confident they were with this decision by pressing buttons corresponding to "very unsure," "rather unsure," "rather sure," and "very sure."

Data analysis. For each trial, we derived a binary unsigned prediction error, which was calculated as the absolute value of the difference between participants' explicit binary shock expectancy ratings (coded 0 when no shock was expected and coded 1 when a shock was expected) and the actual outcome of the trial (coded 0 when no shock occurred and 1 when a shock occurred in the current trial). The resulting prediction error is, therefore, also binary, attaining 0 for any correct prediction (i.e., either an expected shock or an expected shock omission) and 1 for any incorrect prediction (i.e., either an unexpected shock or an unexpected shock omission). It is important to note differences in this conceptualization of prediction errors from other common learning models, such as the Rescorla-Wagner model (Rescorla & Wagner, 1972), that assume prediction errors to be continuous.

SCRs were analyzed using Continuous Decomposition Analysis in Ledalab Version 3.4.9 (Benedek & Kaernbach, 2010). Specifically, we derived the average phasic driver within the specified response window. First, skin conductance data were downsampled to a resolution of 50 Hz and optimized using four sets of

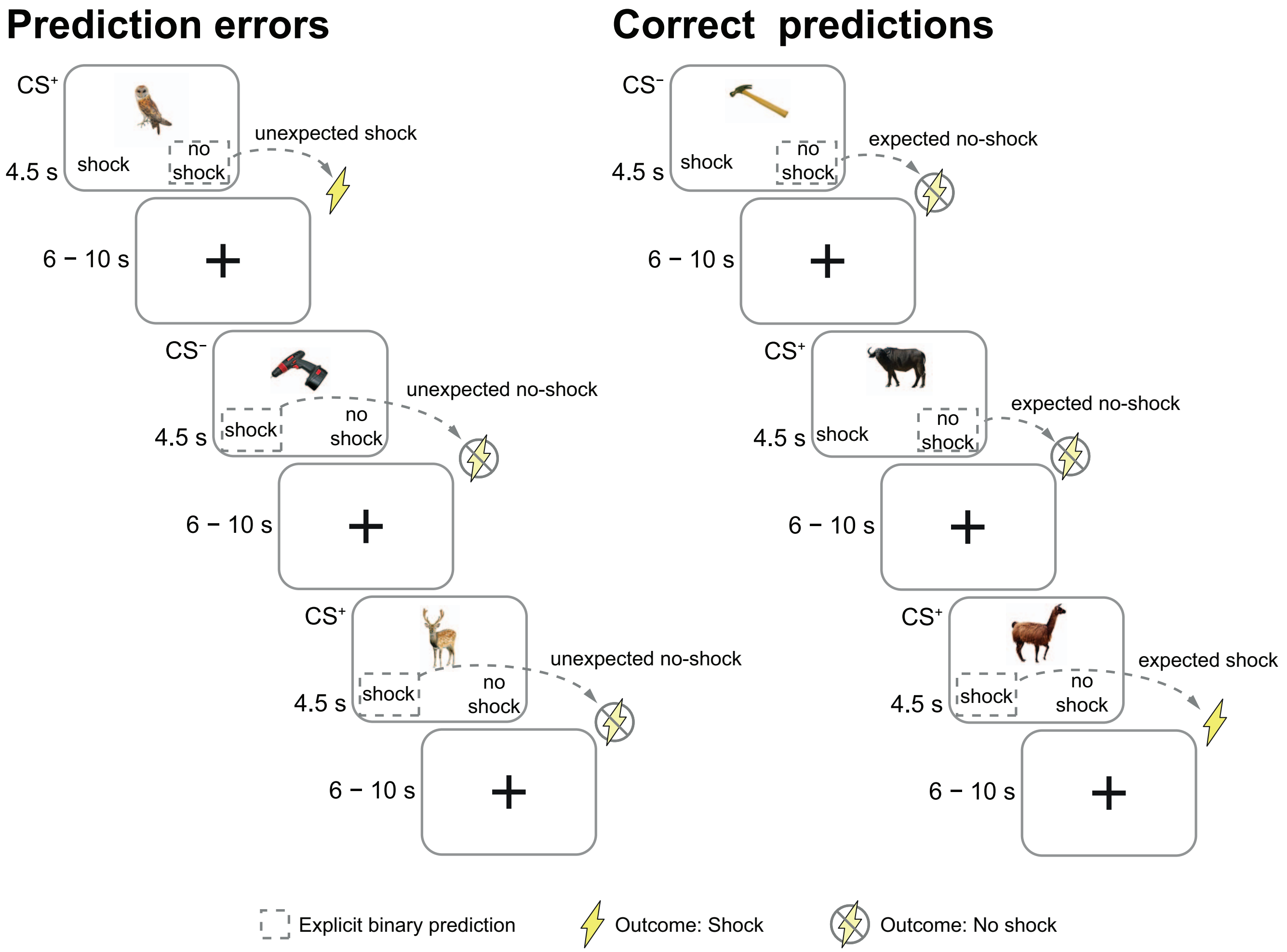

Figure 1. Task design. In each of the 60 trials of a combined incidental learning and fear conditioning task, participants saw a picture of an animal or a tool and predicted whether they would receive an aversive electric shock or not. One of the two stimulus categories (animals or tools) was randomly selected as the conditional stimulus (CS^+) category, while the other served as the CS^- category. Twenty out of the 30 CS^+ pictures were followed by a mild electric shock, while the 30 CS^- pictures were never followed by a shock. Participants were not instructed about these contingencies but had to learn them by trial and error. Memory for the pictures was tested in a surprise recognition test about 24 hr after encoding. See the online article for the color version of this figure.

initial values. For the anticipatory SCR, the response window was set from 0.5 to 4.5 s after stimulus onset. For the outcome-related SCR, the response window was set from 4.5 to 7.9 s after stimulus onset. More important, aversive electrodermal stimulation always occurred exactly 4.5 s after stimulus onset; thus, leaving the anticipatory SCR unaffected by the shock itself. The minimum amplitude threshold was set to 0.01 μS for both the anticipatory and the outcome-related SCR. Resulting estimates of average phasic driver within each response window were returned in μS. It should be noted that these estimates are sensitive to interindividual baseline skin conductance differences because of physiological factors such as the thickness of the corneum (Figner & Murphy, 2011). To account for these interindividual baseline differences, we standardized both the anticipatory and the outcome-related SCR by dividing the average phasic driver estimated in each trial by the maximum average phasic driver for each participant observed in any of the 60 trials.

To investigate how prediction errors and physiological arousal impacted the ability to recognize pictures presented during incidental encoding on the next day, we fitted generalized linear mixed models (GLMMs) with a logit link function using the lme4 R package (Bates, Mächler, Bolker, & Walker, 2015). Compared with a "classic" analysis of proportions of binary recognition per condition and per participant, GLMMs have several advantages, such as increased statistical power and being less prone to spurious results (Dixon, 2008; Jaeger, 2008). Following guidelines to maximize the generalizability of these models, we included the maximal random effects structure, treating subjects as random effects for both the intercept and all slopes of the fixed effects included in the model (Barr, Levy, Scheepers, & Tily, 2013). The recognition of an individual item was treated as the binary dependent variable, coded ‘0’ for misses and ‘1’ for hits. In line with previous research on episodic memory (Bartlett, Till, & Levy, 1980), our analysis focused on high-confidence responses, that is, only trials in which participants indicated that they were either rather sure or very sure were considered. Such high-confidence recognitions have been linked to a hippocampus-based recollection rather than only familiarity with an item, which is assumed to depend on the perirhinal cortex (Eichenbaum, Yonelinas, & Ranganath, 2007). We fitted models using different sets of independent variables, such as prediction errors and measures of arousal and compared their goodness of fits using likelihood ratio tests to select the most appropriate model, indicating which factors drive episodic memory formation most strongly.

Results and Discussion

Anticipation of the memory test. To assess whether participants had expected a recognition test on the second experimental day, they gave ratings from 1 (*not surprised at all*) to 5 (*very surprised*). Questionnaire data from six participants were missing. In the remaining sample of 38 participants, the mean response was 2.92 ($SD = 0.97$), indicating that, on average, participants were moderately surprised. Only four participants indicated that they had anticipated the recognition test by choosing the *not surprised at all* option. These four participants were still included in the analysis and excluding them did not change the pattern of results.

General memory performance. On average, participants correctly recognized 69.5% ($SD = .12$) of all pictures that they had seen on the previous day (*hit rate*). When counting only high-confidence recognitions (i.e., responses with rather sure and very sure confidence ratings) as hits and low-confidence recognitions as misses, the hit rate decreased slightly to 54.5% ($SD = .15$). In comparison, the *false alarm rate* (i.e., incorrectly classifying a new picture as old) was overall low to moderate at 24.4% ($SD = .09$). More important, the false alarm rate for items from the CS^+ category ($M = .25$, $SD = .10$) was comparable with the false alarm rate for items from the CS^- category ($M = .24$, $SD = .13$), $t(43) = 0.35$, $p = .72$, $d_{av} = 0.06$.

Successful fear conditioning. Physiological data from anticipatory skin conductance responses confirmed that fear conditioning was successful. On average, participants showed significantly greater anticipatory SCRs to CS^+ items compared with CS^- items, $t(43) = 4.79$, $p < .001$, $d_{av} = 0.52$. To further analyze when participants first began to show signs of conditioned fear, we divided the task into six consecutive blocks, each consisting of 10 trials (Figure 2A). As expected, in the first 10 trials of the task, participants did not yet show increased anticipatory SCRs to CS^+ items compared with CS^- items, $t(43) = 1.32$, $p = .19$, $d_{av} = 0.13$. Starting from the second block (Trials 11–20); however, we consistently found that anticipatory SCRs were greater for CS^+ items than for CS^- items in all five remaining blocks (all $ps < .004$). This shows that conditioned fear was acquired relatively fast and lasted over the whole encoding phase. An analysis of variance (ANOVA) with block and condition as within-subject factors revealed that anticipatory SCRs were affected by both the condition, $F(1, 43) = 22.04$, $p < .001$, $\eta^2_G = .042$, as well as the block, $F(5, 215) = 13.71$, $p < .001$, $\eta^2_G = .072$. There was no significant interaction between these two factors, $F(5, 215) = 1.49$, $p = .19$, $\eta^2_G = .004$. The lack of a significant Condition × Block interaction is not necessarily surprising, given the fact that anticipatory SCRs differentiated very quickly between CS^+ and CS^- items. An ANOVA might, therefore, not have enough power to detect such small differences within the first few trials, as SCRs were clearly distinct for CS^+ and CS^- stimuli in all following trials. At the descriptive level, however, we found that mean anticipatory SCRs were almost identical in the first five trials for CS^+ items ($M = 0.44$ µS, $SD = 0.21$ µS) versus CS^- items ($M = 0.42$ µS, $SD = 0.24$ µS), providing additional evidence that anticipatory responses for both conditions were initially comparable.

Improved memory for CS^+ items compared with CS^- items. As expected, the average hit rate for items from the CS^+ category ($M = .73$, $SD = .14$) was significantly higher than for items from

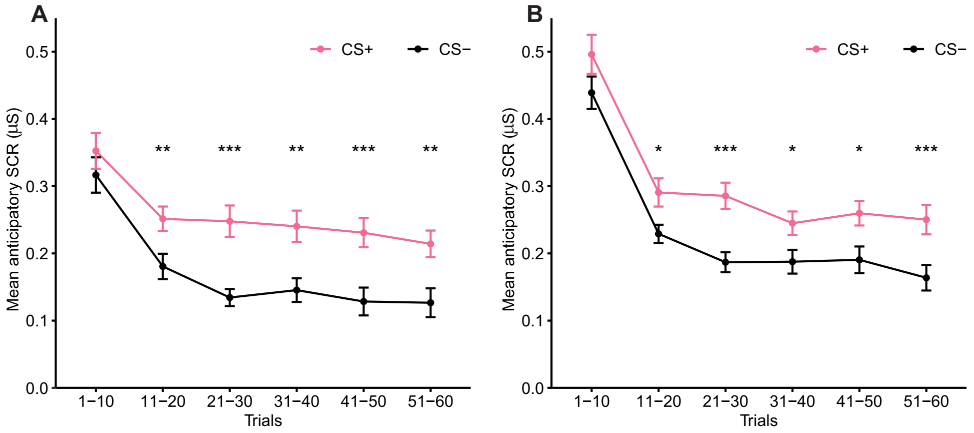

Figure 2. Average anticipatory skin conductance responses by block and condition. Apart from the first 10 trials, anticipatory skin conductance responses were always significantly higher for items from the conditional stimulus (CS^+) category compared with items from the CS^- category in both Experiment 1 (A) and Experiment 2 (B), confirming that the fear conditioning procedure was successful. Error bars represent *SEM*. * $p < .05$. ** $p < .01$. *** $p < .001$. See the online article for the color version of this figure.

the CS^- category (M = .66, SD = .15), $t(43) = 2.35$, $p = .023$, $d_{av} = 0.42$. This finding is generally in line with the classic model that attributes memory advantages for CS^+ items to increased physiological arousal associated with these items.

Prediction errors. Recent evidence suggests that episodic memory formation might not only be driven by physiological arousal during encoding, but also by errors made in predicting future outcomes (Jang et al., 2018; Rouhani et al., 2018). We requested participants to make explicit binary predictions about shock outcomes in each trial. On average, participants made incorrect predictions in 27.8% (SD = .10) of all trials. As expected, the average number of prediction errors decreased as the task progressed, $r(58) = -.34$, $p = .008$. This finding indicates that participants learned the contingency between picture category and shock very well. Because of the partial reinforcement schedule, however, prediction errors occurred also after the contingency was learned. Notably, participants made substantially more prediction errors for CS^+ items (M = .45, SD = .09) compared with CS^- items (M = .11, SD = .14), $t(43) = 18.11$, $p < .001$, $d_{av} = 2.96$. On the other hand, it should be noted that prediction errors were still conceptually different enough from the CS^+/CS^- categories so that their effects could be differentiated. This was reflected by an only moderate association, at item level, between binary prediction errors and the binary category membership of an item (CS^- vs. CS^+), $\varphi = .38$, $p < .001$. This significant moderate association is likely because of prediction errors occurring far more often in CS^+ trials. On the other hand, participants made prediction errors in less than half of CS^+ trials, leaving enough variance in prediction errors even if only CS^+ trials are considered.

Similarly, prediction errors exhibited a small but significant point-biserial correlation with standardized anticipatory SCRs, $r(2{,}624) = .10$, $p < .001$. The same was true for the outcome-related SCRs, $r(2{,}624) = .12$, $p < .001$. Again, these findings are not at all surprising, as SCRs might partly reflect uncertainty and surprise, two concepts that are also linked to prediction errors, and it has been demonstrated before that prediction errors may lead to a certain state of arousal (de Berker et al., 2016). On the other hand, as correlation coefficients were small, we still expected that effects of these two concepts (i.e., arousal and prediction errors) would be separable in a GLMM.

Effects of encoding order on memory performance. The serial position of an individual item within the encoding session could potentially influence memory performance for this item in the following recognition test. For example, participants might show greater attention to items that appear early in the encoding task, leading to better recognition of these early items (i.e., a primacy effect). Awareness of such an effect would be critical, as it might be confounded with other measures that have varying frequencies over the course of the task, such as prediction errors, which become less frequent as the encoding session progresses. To investigate whether the probability of correctly recognizing an item in the memory test depends on the relative position of the item within the task, we fitted a GLMM with the position of each item within the encoding session (i.e., the trial number) as the sole independent variable to explain differences in item recognition on the following day. This revealed no effect of the serial position of an item during encoding on memory formation, $z = 0.56$, $p = .57$, $\beta = 0.002$.

Modeling recognition at item level. So far, we have shown that, on average, items from the CS^+ category were better recognized after 24 hr than items from the CS^- category. Two plausible underlying mechanisms have been identified. First, we showed that CS^+ items provoked increased anticipatory SCRs compared with CS^- items, suggesting that physiological arousal may promote episodic memory. In addition, however, we showed that CS^+ items were also associated with a substantially increased rate of prediction errors for aversive electric shocks, providing initial evidence for an intriguing alternative model in which the observed memory advantage for CS^+ items is linked to an increased prediction error for this category. To test these two models, we fitted GLMMs at item level, treating the binary recognition of an item presented on Day 1 as the dependent variable.

First, to test the model of arousal-induced memory enhancements at item level, we treated the standardized anticipatory SCR in each trial as the sole independent variable to predict the binary recognition of an item. As we expected this model to best reflect fear conditioning-induced memory effects, we treated it as a baseline model for later comparisons. Surprisingly, estimates obtained after fitting the model revealed no significant effect of the anticipatory SCR on item recognition, $z = 0.81$, $p = .41$, $\beta = 0.22$. Next, we added the standardized outcome-related SCR as an additional predictor that reflects physiological arousal after the outcome in a trial has become apparent (i.e., either a shock or no shock). This additional variable showed the expected positive relationship with item-specific recognition performance, indicating that higher SCRs were associated with improved recognition, $z = 2.82$, $p = .005$, $\beta = 0.76$, in line with models of arousal-induced memory enhancement.

In a first minimal model of prediction error-induced memory enhancements, we added the unsigned binary prediction error as the sole independent variable. This revealed that episodic memory was indeed enhanced for trials in which an incorrect shock prediction was made, $z = 2.20$, $p = .027$, $\beta = 0.46$.

To investigate the possibility that the effects of physiological arousal and the effects of prediction errors on memory might reflect distinct mechanisms, we added both measures of arousal (i.e., anticipatory and outcome related SCRs) to the previously defined minimal model that featured only the binary prediction error as the sole independent variable. Again, this revealed no significant effect of anticipatory SCRs on item recognition, $z = 0.24$, $p = .81$, $\beta = -0.07$. Larger outcome-related SCRs, on the other hand, were again associated with better item recognition, $z = 2.52$, $p = .012$, $\beta = 0.72$. For prediction errors, there was a strong trend in the direction that recognition was improved in trials with incorrect predictions, yet this trend did not reach statistical significance, $z = 1.83$, $p = .067$, $\beta = 0.36$.

Using likelihood ratio tests, we next compared the previously introduced models to identify which of them is best suited to describe the mechanisms underlying episodic memory formation in this task (Figure 3A). Critically, the combined model with the anticipatory and outcome-related SCRs as well as prediction errors best reflected the observed recognition performance. As such, its model fit was significantly better compared with the model that only featured the anticipatory and outcome-related SCR as independent variables, $\chi^2(5) = 15.52$, $p = .008$. This shows that prediction errors play a role beyond physio-

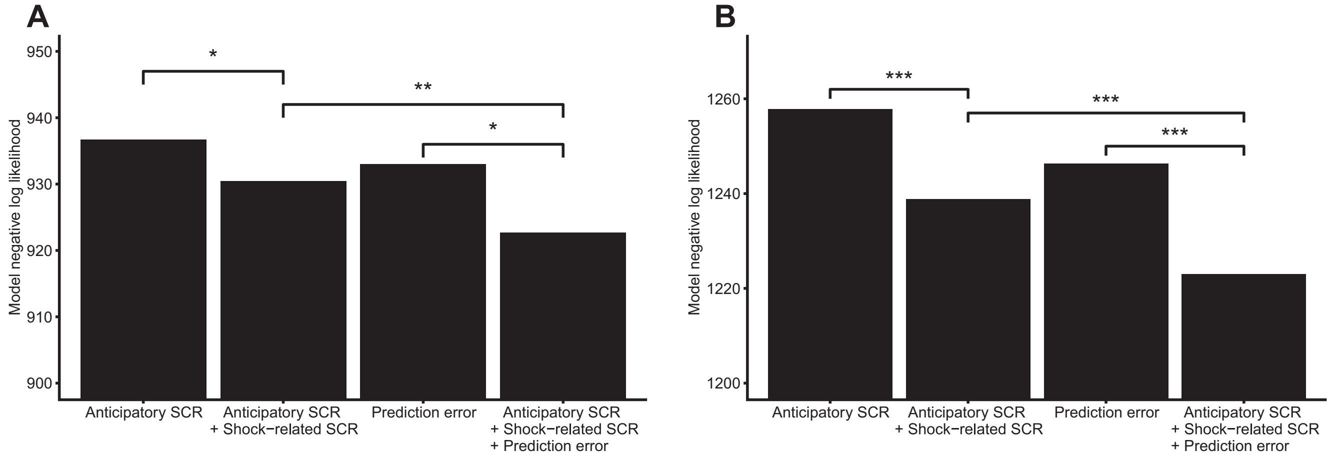

Figure 3. Generalized linear mixed-model (GLMM) fit indices for models with different sets of independent variables to predict the binary recognition of an item in the incidental learning paradigm in Experiment 1 (A) and Experiment 2 (B). Smaller values indicate a better model fit. Notably, the best-fitting model in both experiments combines both measures of physiological arousal and prediction errors, suggesting that both processes contribute to episodic memory formation. Comparisons between models refer to results from likelihood ratio tests. * $p < .05$. ** $p < .01$. *** $p < .001$.

logical arousal in episodic memory formation. On the other hand, adding the two measures of physiological arousal (i.e., anticipatory and outcome-related SCRs) also improved the model fit compared with a model that only relies on prediction errors as the single independent variable, $\chi^2(9) = 20.63$, $p = .014$. Thus, both physiological arousal and prediction errors seem to be important factors in episodic memory formation, each contributing to improve predictions in a combined model.

One potential alternative explanation for our results could be that our measurement of arousal through SCRs does not capture all aspects of physiological arousal. If this was the case, then it would be possible that prediction errors only seemingly predict memory formation beyond arousal because they reflect aspects of arousal that are not fully captured through SCRs. We assumed that such an effect would be particularly strong in the case of unexpected shocks, which should elicit larger outcome-related physiological responses. To test whether the putative contribution of prediction errors to memory formation is mainly driven by such unexpected shocks, we disregarded all trials in which participants incorrectly predicted that no shock would follow; thus, leaving only trials with either correct predictions or unexpected no shocks. Again, we fitted a GLMM to explain the binary recognition of an item with prediction errors, anticipatory and outcome-related SCRs as independent variables. In this model, we found that unexpected no shocks, which include a prediction error but low outcome-related arousal, were associated with an improved recognition on the following day, even after controlling for both SCR measures, $z = 2.00$, $p = .045$, $\beta = 0.43$. Thus, we find a positive link between prediction errors and item recognition even after controlling for arousal and when excluding trials that featured unexpected shocks.

So far, we have shown that prediction errors improved memory for items encoded shortly before the associated aversive outcome became apparent or not, resulting in a possible prediction error. In other words, the effects of prediction errors identified so far have been retroactive in nature. To investigate whether prediction errors might also promote memory for unrelated items in the opposite, proactive direction, we fitted a model with the binary unsigned prediction error of the previous trial as the sole independent variable to explain memory for the current item. We found no effect of prediction errors from the previous trial on the probability of recognizing the item from the current trial, $z = 1.57$, $p = .12$, $\beta = -0.24$. Therefore, memory advantages associated with prediction errors seem to be mainly retroactive and specific to related items, rather than also proactive and generalizable to unrelated items.

Finally, we hypothesized that prediction errors might improve memory independent of the fear conditioning-based memory difference between CS^+ and CS^- items. If this was the case, we should be able to find memory advantages induced by prediction errors even within a conditioned stimulus category. Because of characteristics of our task, prediction errors were rare in CS^- trials (11%), but much more prevalent in CS^+ trials (45%). Therefore, we fit a model with binary prediction errors as the sole independent variable to predict the recognition of an individual item, but this time only included CS^+ trials. Even though the parameter estimate for prediction errors was only slightly diminished compared with the same model fit on all trials and in the expected direction, its effect did not reach significance, $z = 1.34$, $p = .18$, $\beta = 0.33$. We suspected that this might have been because of insufficient statistical power, as including only CS^+ item removed half of all trials from this analysis in a generally rather small sample.

Nonetheless, Experiment 1 overall provided evidence that prediction errors for aversive events were associated with improved item recognition in a surprise memory test on the following day. Critically, these effects of prediction errors on episodic memory could not be fully explained by traditional models based on physiological arousal during encoding.

Experiment 2

Experiment 2 was designed to replicate and clarify the findings of Experiment 1. Specifically, in Experiment 1 we observed two effects of prediction errors at descriptive level that did not reach statistical significance. First, we hypothesized that prediction errors would improve episodic memory even when controlling for measures of physiological arousal. Second, we hypothesized that prediction errors would influence memory even if only CS$^+$ trials are considered. To ensure an appropriate statistical power to detect these possible effects, we almost doubled the sample size compared with Experiment 1 while keeping the procedure largely identical.

Method

Participants. Eighty-four healthy men and women between 18 and 35 years of age (M = 25.23, SD = 4.08) participated on two consecutive days. Four of these participants were excluded from analysis because they either did not complete the task or because of experimenter error. The target sample size was determined a priori in G*Power 3 to achieve a power of .95 to detect an effect size obtained for the memory advantage for CS$^+$ compared with CS$^-$ items observed in Experiment 1 ($d_z \approx 0.4$) using a two-tailed dependent means t test at α = .05. Our decision to increase the statistical power compared with Experiment 1 was based on the observation of some statistical trends in the previous experiment that we aimed to clarify. None of the participants from Experiment 1 participated in Experiment 2. Again, participants received a monetary compensation of 20€ for the completion of the experiment, which was approved by the ethics committee of the Faculty of Psychology and Human Movement Science at the University of Hamburg.

Materials. To rule out the possibility that our results in Experiment 1 could be item specific, we used a new set of stimuli in Experiment 2. These had previously been utilized in a similar incidental learning procedure (Dunsmoor et al., 2015). Again, the stimulus set consisted of 180 color pictures of animals and 180 color pictures of tools on white backgrounds. As in Experiment 1, all stimuli were of neutral valence.

Procedure. The procedure for Study 2 was mostly identical as in Study 1. We changed the location where the stimulation electrode was placed from the right lower leg to the back of the right hand near the wrist to make results more comparable with studies utilizing a similar fear conditioning paradigm (Dunsmoor et al., 2015). As this area tends to be more sensitive to electrical stimulation, the initial intensity in the procedure to determine the pain threshold was reduced to 10 V instead of 20 V. We also replaced the two-step forced-choice decision in the surprise recognition test with a single-step decision that included both whether participants regarded the currently presented picture as old or new as well as participants' confidence with this decision. Thus, on each trial, participants performed a single button press on either the '1,' '2,' '3,' or '4' key at the upper left of the keyboard, indicating that the current item was "definitely old," "maybe old," "maybe new," or "definitely new," respectively.

Data analysis. The statistical analysis was identical to Experiment 1.

Results and Discussion

Anticipation of the memory test. Overall, participants were moderately surprised by the recognition test on the second experimental day, as indicated by a mean rating of 2.89 (SD = 1.12) on a scale from 1 (*not surprised at all*) to 5 (*very surprised*). A total of nine participants answered that they were not surprised at all. As in Experiment 1, these nine participants were still included in the following analyses and excluding them did not affect the pattern of results.

General memory performance. The average hit rate in Experiment 2 was 63.9% (SD = .14) and, therefore, comparable with Experiment 1. Treating only high-confidence recognitions (i.e., correct definitely old responses) as hits reduced the hit rate to 39.3% (SD = .17), considerably lower than in Experiment 1. We suspected that this difference was because of changes in the procedure how confidence was assessed in Experiment 2, which, unlike Experiment 1, did not include a *rather sure* rating. We found a similar false alarm rate as in Experiment 1 at 25.2% (SD = .10). The false alarm rate for items from the CS$^+$ category (M = .25, SD = .11) was comparable with the false alarm rate for CS$^-$ items (M = .26, SD = .14), $t(79) = 0.51$, $p = .61$, $d_{av} = 0.07$.

Successful fear conditioning. As in Experiment 1, anticipatory SCRs provided physiological evidence that our procedure was successful in inducing conditioned fear. More specifically, average anticipatory SCRs to items from the CS$^+$ category were significantly larger than anticipatory SCRs to items from the CS$^-$ category, $t(79) = 4.32$, $p < .001$, $d_{av} = 0.35$. Analogous to Experiment 1, we further divided the task into six consecutive blocks, each consisting of 10 trials, to identify when participants started to show first signs of conditioned fear (Figure 2B). Again, in the first 10 trials of the task, participants did not yet show a significantly increased anticipatory SCRs to CS$^+$ items compared with CS$^-$ items, although a trend was already visible, $t(79) = 1.84$, $p = .07$, $d_{av} = 0.16$. In all five remaining blocks representing Trials 11 to 60, we consistently found that anticipatory SCRs were greater for CS$^+$ items than for CS$^-$ items (all $ps < .02$). This demonstrates that conditioned fear emerged relatively fast and lasted over the whole encoding session. As in Experiment 1, a repeated measures ANOVA revealed that the anticipatory SCRs depended on both the condition, $F(1, 79) = 19.10$, $p < .001$, $\eta^2_G = .018$, as well as the block, $F(5, 395) = 40.10$, $p < .001$, $\eta^2_G = .105$. There was no significant interaction between condition and block, $F(5, 395) = 0.62$, $p = .68$, $\eta^2_G = .0001$. At descriptive level, however, we found that within the first five trials, there was almost no difference in mean anticipatory SCRs between CS$^+$ items (M = 0.60 μS, SD = 0.33 μS) compared with CS$^-$ items (M = 0.61 μS, SD = 0.28 μS), providing additional evidence that anticipatory responses for both conditions were initially comparable.

Improved memory for CS$^+$ items compared with CS$^-$ items. As expected, we could replicate the previous finding of improved recognition for items from the CS$^+$ category. More specifically, the average hit rate for CS$^+$ items (M = .68, SD = .18) was significantly higher than for CS$^-$ items (M = .60, SD = .18), $t(79) = 3.53$, $p < .001$, $d_{av} = 0.47$.

Prediction errors. On average, participants made incorrect predictions in 26.0% (SD = .06) of all trials. They learned the underlying picture-shock contingencies very well, as reflected in the observation that the average proportion of prediction errors decreased as the task progressed, $r(58) = -.62$, $p < .001$. As in Experiment 1, participants made substantially more prediction errors in trials in which CS$^+$ pictures were displayed ($M = .45$, $SD = .09$) compared with trials that displayed CS$^-$ pictures ($M = .07$, $SD = .09$), $t(79) = 29.14$, $p < .001$, $d_{av} = 4.44$. Still, prediction errors were conceptually differentiable from the CS$^+$/CS$^-$ categories as indicated by only a medium-sized association between binary prediction error and category membership (CS$^-$ vs. CS$^+$) at item level, $\varphi = .44$, $p < .001$. Again, we assumed that this significant association mostly reflects that prediction errors were far more common in CS$^+$ trials. On the other hand, prediction errors did occur in less than half of all CS$^+$ trials, leaving enough differential variance to separate these two concepts. As in Experiment 1, we found an only small but significant point-biserial correlation between prediction errors and the standardized anticipatory SCR, $r(4{,}858) = .10$, $p < .001$. This was paralleled by a small to moderate significant point-biserial correlation between prediction errors and the standardized outcome-related SCR, $r(4{,}858) = .21$, $p < .001$. Corroborating findings from Experiment 1, this likely reflects how uncertainty and surprise might be connected to both arousal measures and prediction errors. More important, however, as correlation coefficients were only small, we expected that effects of these two concepts on episodic memory formation could be differentiated in a GLMM.

Effects of encoding order on memory performance. To explore possible effects of the serial position of an item within the encoding session, we fitted a GLMM with the trial number of each item as the sole independent variable to explain differences in item recognition on the following day. As in Experiment 1, this revealed that memory formation was not influenced by the serial position of an item during encoding, $z = 1.46$, $p = .14$, $\beta = -0.005$.

Modeling recognition at item level. For a more precise analysis of mechanisms underlying episodic memory formation, we fitted the same GLMMs as in Experiment 1 to predict the binary recognition of individual items. We started with the same baseline model as in Experiment 1 with the standardized anticipatory SCR as the sole independent variable. As in Experiment 1, this revealed no significant effect of the anticipatory SCR on recognition performance, $z = 1.47$, $p = .14$, $\beta = -0.38$. Next, we added the standardized outcome-related SCR as an additional independent variable to the model. In this model, surprisingly, we found that anticipatory SCRs were linked to a decreased chance that an item would be recognized, $z = 2.01$, $p = .039$, $\beta = -0.54$. We could, however, replicate the finding from Experiment 1 that the outcome-related SCR was associated with better item recognition, $z = 3.29$, $p < .001$, $\beta = 1.02$.

Fitting a simple model with binary prediction errors as the single independent variable to predict item recognition, we replicated the finding from Experiment 1 that prediction errors were linked to improved recognition performance, $z = 4.32$, $p < .001$, $\beta = 0.73$. In a combined model, we added both measures of physiological arousal (i.e., anticipatory and outcome-related SCR) together with prediction errors as independent variables. In this model, the anticipatory SCR was again associated with reduced recognition, $z = 2.75$, $p = .006$, $\beta = -0.72$. Congruent with all previous findings, there was also a positive effect of outcome-related SCRs on item recognition, $z = 2.93$, $p = .003$, $\beta = 0.90$. Most important, however, in this combined model, we found a significant positive effect of prediction errors on recognition even when accounting for measures of physiological arousal through SCRs, $z = 4.19$, $p < .001$, $\beta = 0.67$. This demonstrates that prediction errors influence item recognition through other mechanisms than the well-known arousal-based effects.

Next, we compared all previously introduced models using likelihood ratio test to identify the model that best reflects underlying mechanisms of episodic memory formation in Experiment 2 (Figure 3B). The results mimicked the pattern observed in Experiment 1. Again, the model combining physiological arousal measures (i.e., anticipatory and outcome-related SCRs) with prediction errors showing the best fit to predict the recognition of individual items. This combined model fit our recognition data significantly better than the model that only featured measures of physiological arousal, $\chi^2(5) = 31.79$, $p < .001$, demonstrating that the role of prediction errors in episodic memory formation goes beyond arousal. Likewise, the combined model also had a significantly better fit than the model that only included the prediction error to explain recognition differences, $\chi^2(9) = 46.64$, $p < .001$. In line with Experiment 1, these findings demonstrate that episodic memory formation is influenced by both arousal and prediction errors.

As in Experiment 1, we considered the possibility that the putative positive effect of prediction errors on memory formation beyond arousal might be because of the way we measure arousal through SCRs, which might not capture every aspect of physiological arousal. To investigate this possibility, we again excluded all trials with unexpected shocks, for which we assumed a particularly pronounced physiological response should follow. Including only the remaining trials, which featured either correct predictions or unexpected no shocks, we fit a GLMM with the binary recognition of an item on the following day as the dependent variable and prediction errors, anticipatory and outcome-related SCRs as the independent variables. As in Experiment 1, prediction errors were still associated with an improved item recognition even after controlling for arousal and excluding all trials with unexpected shocks, $z = 3.97$, $p < .001$, $\beta = 0.90$.

The results from Experiment 2 so far provide evidence that prediction errors retroactively promote memory for related items. As in Experiment 1, we further investigated whether prediction errors also affected memory for subsequent unrelated pictures, in a proactive manner. We fitted a model with the unsigned binary prediction error in the previous trial as a single independent variable to explain memory for the current picture. Although not significant, there was a tendency indicating that prediction errors might also have a proactive, memory-promoting effect for directly following pictures, $z = 1.88$, $p = .06$, $\beta = 0.34$.

Like in Experiment 1, prediction errors were rare for items of the CS$^-$ category (7%), but common for items of the CS$^+$ category (45%) because of task characteristics. We hypothesized that, in this larger sample, we might be able to identify memory improvements through prediction errors even when analyzing only trials from the CS$^+$ category. This finding would be particularly interesting, as it would indicate that the effects of prediction errors on memory formation cannot solely be attributed to the increased number of prediction errors for CS$^+$ items. It would, therefore, point to a general role of prediction errors for aversive events in memory formation.

To test whether prediction errors may account for variability in memory for CS$^+$ items, we again fitted a model with the binary prediction error as a single independent variable to predict the recognition of an item, including only items from the CS$^+$ category. With the increased sample size in this experiment, we found a positive effect of prediction errors on the recognition performance for CS$^+$ items only, $z = 2.95$, $p = .003$, $\beta = 0.58$. In other words, when two pictures were both from the CS$^+$ category, but for one an incorrect prediction was made, this item was more likely to be recognized later than the item for which a correct prediction was made. This finding provides striking evidence that prediction errors for aversive events generally improve memory formation.

General Discussion

Classic models of emotional memory formation have attributed the enhanced memory for information linked to aversive events to increased physiological arousal during encoding (Cahill et al., 1994; McGaugh, 2018; McGaugh & Roozendaal, 2002). Based on the assumption that aversive events are often characterized by their unpredictability (de Berker et al., 2016; Trapp et al., 2018), we hypothesized that the memory enhancement for stimuli linked to aversive events might additionally be driven by an element of surprise (i.e., prediction errors) that has not been accounted for by purely arousal based models. To test this hypothesis, we exposed participants to a combined fear conditioning and incidental learning paradigm that featured partially predictable aversive shocks while we collected data on both physiological arousal and prediction errors to predict 24 hr delayed memory performance. In line with the model of arousal enhanced memory formation, we found that outcome-related arousal predicted, on a trial-by-trial basis, whether an item was later recognized. Most important, however, our data show that, in addition to arousal, binary unsigned prediction errors derived from participants' explicit shock predictions were associated—on a trial-by-trial basis—with enhanced recognition. In support of the idea that the impact of a prediction error on memory goes beyond the mere effect of arousal, a model that included both measures of physiological arousal and the unsigned prediction error to explain recognition significantly outperformed models featuring only one of these measures. This pattern of results was replicated in a second experiment in a larger sample. In addition, we showed in this second experiment the memory facilitating effect of prediction errors when only items from the CS$^+$ category were included; thus, demonstrating the robustness of this effect and that the facilitating effect of prediction errors on memory remained stable even after controlling for the influence of arousal. Together, these findings provide strong evidence that prediction errors promote, above and beyond physiological arousal, memory formation for stimuli linked to aversive events.

While our findings point to a new mechanism involved in the formation of episodic memories for stimuli linked with emotional events, they provide also further evidence for the well-established model of arousal-based memory enhancement (McGaugh, 2018). In particular, SCRs, a common indicator of autonomic arousal, elicited by the outcome in each trial (i.e., either a shock or no shock) were linked to enhanced item recognition. Somewhat surprisingly, anticipatory SCRs, reflecting arousal in anticipation of a possible shock, had either no effect on item recognition (Experiment 1) or were even associated with a decreased recognition performance (Experiment 2). These divergent findings between anticipatory SCRs, associated with either no effect (Experiment 1) or even a negative effect on memory encoding (Experiment 2), and outcome-related SCRs, linked to enhanced memory formation, might be explained through different processes underlying these measures of physiological arousal. Outcome-related SCRs have been demonstrated to partly reflect surprise (i.e., prediction errors), while anticipatory SCRs have been associated with concepts such as uncertainty and fear (de Berker et al., 2016). Therefore, it is tempting to speculate that fear-related anticipatory arousal during the encoding might, unlike surprise, act as a distractor and hence have negative effects on memory formation. However, the negative effect of anticipatory SCRs on memory formation was not consistent across our two experiments and, therefore, remains to be interpreted with caution.

The key finding of our experiments, however, is that the enhanced memory for stimuli paired with aversive events is not exclusively because of the associated physiological arousal, as measured through SCRs, but also due to a violation of expectations. These prediction errors facilitated recognition memory independent from the beneficial effects of arousal. In line with models of adaptive memory (Anderson & Milson, 1989; Nairne & Pandeirada, 2008; Nairne et al., 2007; Shohamy & Adcock, 2010) proposing that memory is essential to guide future behavior, the impact of prediction errors was inherently retroactive in nature. Prediction errors enhanced memory for preceding stimuli that were linked to the incorrect prediction but not for stimuli that followed the prediction error, suggesting that the prediction error does not open a "bidirectional" window of enhanced memory formation but selectively favors memory for preceding events. To explain these findings, we propose that prediction errors might transiently put agents into a state of enhanced information processing (Trapp et al., 2018), which also extends to the recently encoded stimulus that the prediction error originated from. At the neural level, the dopaminergic system is a likely candidate to be involved in the observed effects. Rouhani et al. (2018) explained memory promoting effects of prediction errors in reward learning through dopaminergic modulation of the hippocampus. This is plausible because the coding of reward prediction errors through dopamine is well established (Schultz & Dickinson, 2000). Which neurotransmitter system is carrying the aversive prediction error, however, is less clear (Delgado, Li, Schiller, & Phelps, 2008).

Prediction errors may indeed be a driving force that promotes adaptive memory, allowing the efficient storage selectively of those memories that are relevant to guide future behavior (Nairne & Pandeirada, 2008; Nairne et al., 2007; Shohamy & Adcock, 2010). The enhanced storage of information linked to previously unexpected events, makes especially this information more available in memory that may help to make more accurate predictions in the future. In accordance with this assumption, prediction errors became less frequent as the task progressed. This finding might be problematic if it was interpreted as an indicator of task disengagement in later trials. However, it is important to note that, even in later stages of the task, participants' mean shock expectancy ratings for CS$^+$ items were clearly below 80%. One explanation for this finding could be that pictures from the CS$^+$ category were not continuously paired with the UCS (rate of 66%), which likely kept participants more alert and made task disengagement less likely.

The neural underpinnings of arousal-induced memory changes are very well documented: emotional events activate β-adrenergic receptors in the basolateral amygdala that then modulates the consolidation of memories in other areas such as the hippocampus (Cahill & McGaugh, 1996, 1998; McGaugh, 2018; McGaugh & Roozendaal, 2002). Neural signatures for aversive prediction errors, on the other hand, have mainly been localized in the striatum (Li, Schiller, Schoenbaum, Phelps, & Daw, 2011; Robinson, Frank, Sahakian, & Cools, 2010; Robinson, Overstreet, Charney, Vytal, & Grillon, 2013; Seymour, Daw, Dayan, Singer, & Dolan, 2007; Seymour et al., 2004). Thus, it appears likely that prediction errors promote memory for aversive events through a different neural pathway focusing around the striatum compared with the amygdala-based effect of arousal.

Although we argue that physiological arousal and prediction errors exert separable influences on memory, arousal and prediction errors may not necessarily be independent of one another. In particular, there is first evidence that prediction errors might be reflected in outcome-related SCRs (Spoormaker et al., 2012; but see Bach & Friston, 2012) and a recent study suggested that physiological arousal could be tuned by environmental uncertainty (de Berker et al., 2016). This evidence points to the intriguing possibility that arousal is, at least partly, the result of a prediction error. In line with this observation, we found small, but significant positive correlations between prediction errors and outcome-related SCRs in both of our experiments, which might suggest that outcome-related SCRs were partially driven by prediction errors. Nevertheless, it is important to note that a combined model of physiological arousal and prediction errors could explain memory performance significantly better than models that relied solely on physiological arousal or prediction errors alone. Furthermore, in Experiment 2, we showed that prediction errors were associated with enhanced recognition even after controlling for arousal. These data suggest that the effects of arousal and prediction error are at least partly independent of each other.

It should be noted that, although SCRs are commonly used to measure physiological arousal in studies concerned with both fear conditioning (Beckers, Krypotos, Boddez, Effting, & Kindt, 2013; Dengerink & Taylor, 1971; Epstein & Clarke, 1970) and stress (Fowles, Roberts, & Nagel, 1977; Jacobs et al., 1994; Lazarus, Speisman, & Mordkoff, 1963), there might be certain components of arousal responses that are not fully captured by SCRs. This is demonstrated by the finding that different indices of physiological arousal do not always correlate (Neiss, 1988). SCRs have also been found to measure concepts beyond physiological arousal, such as the anticipation of cognitive demand (Botvinick & Rosen, 2009). Therefore, it is possible that prediction errors enhance memory through an aspect of physiological arousal that cannot be measured through SCRs. Similarly, it is possible that SCRs were linked to an improved memory formation not exclusively because of arousal, but also because of other factors that they measure, such as cognitive demand. Future research should address this limitation by using a wider array of arousal measures such as pupil diameter and subjective stress ratings. One consistent finding across both experiments, however, was that prediction errors were associated with an improved item recognition beyond arousal as measured through SCRs, even if we excluded any trials featuring unexpected shocks. As we assumed greater physiological arousal for unexpected shocks compared with unexpected shock omissions, this finding could be interpreted as evidence against the possibility that our results were biased by an imperfect arousal measurement through SCRs.

While numerous studies have demonstrated predictive coding in a variety of cognitive domains (Feldman & Friston, 2010; Hollerman & Schultz, 1998; Hosoya et al., 2005; Maia, 2009; Rangel, Camerer, & Montague, 2008; Rao & Ballard, 1999; Smith & Lewicki, 2006; Spratling, 2008), prediction errors were related to the formation of human long-term memory only very recently.

Two recent studies showed that surprise during reward learning may promote episodic memory formation (Jang et al., 2018; Rouhani et al., 2018). Our findings are generally in line with these studies but extend them significantly. We demonstrate for the first time that prediction errors are critical in memory formation related to aversive events and that this impact of prediction errors goes beyond the effect of physiological arousal, which is at the heart of traditional models on emotional memory formation. While it cannot be ruled out that, in the context of these prior studies, some participants perceived receiving a smaller than expected monetary reward as aversive, outcomes were always positive, meaning they never had to fear losing any money. Our study, on the other hand, used aversive electric shocks, which have been extensively used to induce conditioned fear in experimental contexts as a model for psychopathology.

It is also important to note conceptual differences between our findings and classic learning models that rely on prediction errors, such as the Rescorla-Wagner model (Rescorla & Wagner, 1972). In the Rescorla-Wagner model, each stimulus is typically presented several times and the associative strength between UCS and CS is updated after each episode through a weighted prediction error. In other words, the prediction error facilitates learning to a stimulus that is presented repeatedly. We, on the other hand, show here that the prediction error promotes episodic memory for an individual stimulus that is presented only once during encoding.

Demonstrating the relevance of prediction errors in memory formation related to aversive events is particularly relevant because episodic memories for aversive events play a key role in several psychopathologies, including phobia or posttraumatic stress disorder (de Quervain et al., 2017; Dunsmoor & Paz, 2015; Pitman, 1989).

In summary, we show here that superior memory for information paired with aversive events is, at least partly, driven by prediction errors. While classical models of emotional memory formation focused largely on emotional arousal, the present findings point to a cognitive mechanism that contributes to memory formation related to aversive events. Taking this cognitive side of emotional memory formation into account may enhance our understanding of adaptive emotional memory and might ultimately have relevant implications for treating psychopathologies that are characterized by aberrant memory for emotional events.

References

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96,* 703–719. http://dx.doi.org/10.1037/0033-295X.96.4.703

Bach, D. R., & Friston, K. J. (2012). No evidence for a negative prediction error signal in peripheral indicators of sympathetic arousal. *NeuroImage, 59,* 883–884. http://dx.doi.org/10.1016/j.neuroimage.2011.08.091

Baldeweg, T. (2006). Repetition effects to sounds: Evidence for predictive coding in the auditory system. *Trends in Cognitive Sciences, 10,* 93–94. http://dx.doi.org/10.1016/j.tics.2006.01.010

Bar, M. (2007). The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Sciences, 11,* 280–289. http://dx.doi.org/10.1016/j.tics.2007.05.005

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68,* 255–278. http://dx.doi.org/10.1016/j.jml.2012.11.001

Bartlett, J. C., Till, R. E., & Levy, J. C. (1980). Retrieval characteristics of complex pictures: Effects of verbal encoding. *Journal of Verbal Learning & Verbal Behavior, 19,* 430–449. http://dx.doi.org/10.1016/S0022-5371(80)90303-5

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. http://dx.doi.org/10.18637/jss.v067.i01

Beckers, T., Krypotos, A.-M., Boddez, Y., Effting, M., & Kindt, M. (2013). What's wrong with fear conditioning? *Biological Psychology, 92,* 90–96. http://dx.doi.org/10.1016/j.biopsycho.2011.12.015

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods, 190,* 80–91. http://dx.doi.org/10.1016/j.jneumeth.2010.04.028

Botvinick, M. M., & Rosen, Z. B. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research, 73,* 835–842. http://dx.doi.org/10.1007/s00426-008-0197-8

Brodeur, M. B., Dionne-Dostie, E., Montreuil, T., & Lepage, M. (2010). The Bank of Standardized Stimuli (BOSS), a new set of 480 normative photos of objects to be used as visual stimuli in cognitive research. *PLoS ONE, 5,* e10773. http://dx.doi.org/10.1371/journal.pone.0010773

Brodeur, M. B., Guérard, K., & Bouras, M. (2014). Bank of Standardized Stimuli (BOSS) phase II: 930 new normative photos. *PLoS ONE, 9,* e106953. http://dx.doi.org/10.1371/journal.pone.0106953

Bubic, A., von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in Human Neuroscience, 4,* 25.

Cahill, L., & McGaugh, J. L. (1996). Modulation of memory storage. *Current Opinion in Neurobiology, 6,* 237–242. http://dx.doi.org/10.1016/S0959-4388(96)80078-X

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences, 21,* 294–299. http://dx.doi.org/10.1016/S0166-2236(97)01214-9

Cahill, L., Prins, B., Weber, M., & McGaugh, J. L. (1994). β-adrenergic activation and memory for emotional events. *Nature, 371,* 702–704. http://dx.doi.org/10.1038/371702a0

Christianson, S. A., & Loftus, E. F. (1987). Memory for traumatic events. *Applied Cognitive Psychology, 1,* 225–239. http://dx.doi.org/10.1002/acp.2350010402

Christianson, S. A., Loftus, E. F., Hoffman, H., & Loftus, G. R. (1991). Eye fixations and memory for emotional events. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 693–701. http://dx.doi.org/10.1037/0278-7393.17.4.693

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36,* 181–204. http://dx.doi.org/10.1017/S0140525X12000477

de Berker, A. O., Rutledge, R. B., Mathys, C., Marshall, L., Cross, G. F., Dolan, R. J., & Bestmann, S. (2016). Computations of uncertainty mediate acute stress responses in humans. *Nature Communications, 7,* 10996. http://dx.doi.org/10.1038/ncomms10996

Delgado, M. R., Li, J., Schiller, D., & Phelps, E. A. (2008). The role of the striatum in aversive learning and aversive prediction errors. *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences, 363,* 3787–3800. http://dx.doi.org/10.1098/rstb.2008.0161

Dengerink, H. A., & Taylor, S. P. (1971). Multiple responses with differential properties in delayed galvanic skin response conditioning: A review. *Psychophysiology, 8,* 348–360. http://dx.doi.org/10.1111/j.1469-8986.1971.tb00465.x

de Quervain, D., Schwabe, L., & Roozendaal, B. (2017). Stress, glucocorticoids and memory: Implications for treating fear-related disorders. *Nature Reviews Neuroscience, 18,* 7–19. http://dx.doi.org/10.1038/nrn.2016.155

Dixon, P. (2008). Models of accuracy in repeated-measures designs. *Journal of Memory and Language, 59,* 447–456. http://dx.doi.org/10.1016/j.jml.2007.11.004

Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature, 520,* 345–348. http://dx.doi.org/10.1038/nature14106

Dunsmoor, J. E., & Paz, R. (2015). Fear generalization and anxiety: Behavioral and neural mechanisms. *Biological Psychiatry, 78,* 336–343. http://dx.doi.org/10.1016/j.biopsych.2015.04.010

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience, 30,* 123–152. http://dx.doi.org/10.1146/annurev.neuro.30.051606.094328

Epstein, S., & Clarke, S. (1970). Heart rate and skin conductance during experimentally induced anxiety: Effects of anticipated intensity of noxious stimulation and experience. *Journal of Experimental Psychology, 84,* 105–112. http://dx.doi.org/10.1037/h0028929

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39,* 175–191. http://dx.doi.org/10.3758/BF03193146

Feldman, H., & Friston, K. J. (2010). Attention, uncertainty, and free-energy. *Frontiers in Human Neuroscience, 4,* 215. http://dx.doi.org/10.3389/fnhum.2010.00215

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

Fowles, D. C., Roberts, R., & Nagel, K. E. (1977). The influence of introversion/extraversion on the skin conductance response to stress and stimulus intensity. *Journal of Research in Personality, 11,* 129–146. http://dx.doi.org/10.1016/0092-6566(77)90012-5

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11,* 127–138. http://dx.doi.org/10.1038/nrn2787

Greve, A., Cooper, E., Kaula, A., Anderson, M. C., & Henson, R. (2017). Does prediction error drive one-shot declarative learning? *Journal of Memory and Language, 94,* 149–165. http://dx.doi.org/10.1016/j.jml.2016.11.001

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature Neuroscience, 1,* 304–309. http://dx.doi.org/10.1038/1124

Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature, 436,* 71–77. http://dx.doi.org/10.1038/nature03689

Jacobs, S. C., Friedman, R., Parker, J. D., Tofler, G. H., Jimenez, A. H., Muller, J. E., . . . Stone, P. H. (1994). Use of skin conductance changes during mental stress testing as an index of autonomic arousal in cardiovascular research. *American Heart Journal, 128,* 1170–1177. http://dx.doi.org/10.1016/0002-8703(94)90748-X

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language, 59,* 434–446. http://dx.doi.org/10.1016/j.jml.2007.11.007

Jang, A., Nassar, M., Dillon, D., & Frank, M. J. (2018, May 21). Positive reward prediction errors strengthen incidental memory encoding. *bioRxiv, 327445*. http://dx.doi.org/10.1101/327445

Joëls, M., & Baram, T. Z. (2009). The neuro-symphony of stress. *Nature Reviews Neuroscience, 10,* 459–466. http://dx.doi.org/10.1038/nrn2632

LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience, 7,* 54–64. http://dx.doi.org/10.1038/nrn1825

Lazarus, R. S., Speisman, J. C., & Mordkoff, A. M. (1963). The relationship between autonomic indicators of psychological stress: Heart rate and skin conductance. *Psychosomatic Medicine, 25,* 19–30. http://dx.doi.org/10.1097/00006842-196301000-00004

Li, J., Schiller, D., Schoenbaum, G., Phelps, E. A., & Daw, N. D. (2011). Differential roles of human striatum and amygdala in associative learning. *Nature Neuroscience, 14,* 1250–1252. http://dx.doi.org/10.1038/nn.2904

Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective & Behavioral Neuroscience, 9,* 343–364. http://dx.doi.org/10.3758/CABN.9.4.343

Maren, S. (2001). Neurobiology of Pavlovian fear conditioning. *Annual Review of Neuroscience, 24,* 897–931. http://dx.doi.org/10.1146/annurev.neuro.24.1.897

McGaugh, J. L. (2018). Emotional arousal regulation of memory consolidation. *Current Opinion in Behavioral Sciences, 19,* 55–60. http://dx.doi.org/10.1016/j.cobeha.2017.10.003

McGaugh, J. L., & Roozendaal, B. (2002). Role of adrenal stress hormones in forming lasting memories in the brain. *Current Opinion in Neurobiology, 12,* 205–210. http://dx.doi.org/10.1016/S0959-4388(02)00306-9

Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin, 117,* 363–386. http://dx.doi.org/10.1037/0033-2909.117.3.363

Nairne, J. S., & Pandeirada, J. N. S. (2008). Adaptive memory: Remembering with a stone-age brain. *Current Directions in Psychological Science, 17,* 239–243. http://dx.doi.org/10.1111/j.1467-8721.2008.00582.x

Nairne, J. S., Thompson, S. R., & Pandeirada, J. N. S. (2007). Adaptive memory: Survival processing enhances retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33,* 263–273. http://dx.doi.org/10.1037/0278-7393.33.2.263

Neiss, R. (1988). Reconceptualizing arousal: Psychobiological states in motor performance. *Psychological Bulletin, 103,* 345–366. http://dx.doi.org/10.1037/0033-2909.103.3.345

Pape, H.-C., & Pare, D. (2010). Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. *Physiological Reviews, 90,* 419–463. http://dx.doi.org/10.1152/physrev.00037.2009

Phelps, E. A. (2004). Human emotion and memory: Interactions of the amygdala and hippocampal complex. *Current Opinion in Neurobiology, 14,* 198–202. http://dx.doi.org/10.1016/j.conb.2004.03.015

Pitman, R. K. (1989). Post-traumatic stress disorder, hormones, and memory. *Biological Psychiatry, 26,* 221–223. http://dx.doi.org/10.1016/0006-3223(89)90033-4

Rangel, A., Camerer, C., & Montague, P. R. (2008). A framework for studying the neurobiology of value-based decision making. *Nature Reviews Neuroscience, 9,* 545–556. http://dx.doi.org/10.1038/nrn2357

Rao, R. P. N., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2,* 79–87. http://dx.doi.org/10.1038/4580

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Robinson, O. J., Frank, M. J., Sahakian, B. J., & Cools, R. (2010). Dissociable responses to punishment in distinct striatal regions during reversal learning. *NeuroImage, 51,* 1459–1467. http://dx.doi.org/10.1016/j.neuroimage.2010.03.036

Robinson, O. J., Overstreet, C., Charney, D. R., Vytal, K., & Grillon, C. (2013). Stress increases aversive prediction error signal in the ventral striatum. *Proceedings of the National Academy of Sciences of the United States of America, 110,* 4129–4133. http://dx.doi.org/10.1073/pnas.1213923110

Rouhani, N., Norman, K. A., & Niv, Y. (2018). Dissociable effects of surprising rewards on learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44,* 1430–1443. http://dx.doi.org/10.1037/xlm0000518

Schultz, W. (2000). Multiple reward signals in the brain. *Nature Reviews Neuroscience, 1,* 199–207. http://dx.doi.org/10.1038/35044563

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593–1599. http://dx.doi.org/10.1126/science.275.5306.1593

Schultz, W., & Dickinson, A. (2000). Neuronal coding of prediction errors. *Annual Review of Neuroscience, 23,* 473–500. http://dx.doi.org/10.1146/annurev.neuro.23.1.473

Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. *Neuroscience and Biobehavioral Reviews, 36,* 1740–1749. http://dx.doi.org/10.1016/j.neubiorev.2011.07.002

Schwarze, U., Bingel, U., & Sommer, T. (2012). Event-related nociceptive arousal enhances memory consolidation for neutral scenes. *The Journal of Neuroscience, 32,* 1481–1487. http://dx.doi.org/10.1523/JNEUROSCI.4497-11.2012

Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *The Journal of Neuroscience, 27,* 4826–4831. http://dx.doi.org/10.1523/JNEUROSCI.0400-07.2007

Seymour, B., O'Doherty, J. P., Dayan, P., Koltzenburg, M., Jones, A. K., Dolan, R. J., . . . Frackowiak, R. S. (2004). Temporal difference models describe higher-order learning in humans. *Nature, 429,* 664–667. http://dx.doi.org/10.1038/nature02581

Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences, 14,* 464–472. http://dx.doi.org/10.1016/j.tics.2010.08.002

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature, 439,* 978–982. http://dx.doi.org/10.1038/nature04485

Spoormaker, V. I., Blechert, J., Goya-Maldonado, R., Sämann, P. G., Wilhelm, F. H., & Czisch, M. (2012). Additional support for the existence of skin conductance responses at unconditioned stimulus omission. *NeuroImage, 63,* 1404–1407. http://dx.doi.org/10.1016/j.neuroimage.2012.08.050

Spratling, M. W. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research, 48,* 1391–1408. http://dx.doi.org/10.1016/j.visres.2008.03.009

Trapp, S., O'Doherty, J. P., & Schwabe, L. (2018). Stressful events as teaching signals for the brain. *Trends in Cognitive Sciences, 22,* 475–478. http://dx.doi.org/10.1016/j.tics.2018.03.007

Wacongne, C., Labyt, E., van Wassenhove, V., Bekinschtein, T., Naccache, L., & Dehaene, S. (2011). Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences of the United States of America, 108,* 20754–20759. http://dx.doi.org/10.1073/pnas.1117807108

Walkenbach, J., & Haddad, N. F. (1980). The Rescorla-Wagner theory of conditioning: A review of the literature. *The Psychological Record, 30,* 497–509. http://dx.doi.org/10.1007/BF03394701

Received January 14, 2019
Revision received March 28, 2019
Accepted April 16, 2019 ■