

**Declarative memory modulation in threatening
environments: Introducing a cognitive account based
on aversive prediction errors**

Dissertation
zur Erlangung des Doktorgrades
an der Universität Hamburg,
Fakultät für Psychologie und Bewegungswissenschaft,
Institut für Psychologie

vorgelegt von
Felix Kalbe

Hamburg, 2021

Tag der mündlichen Prüfung: 25.11.2021

Promotionsprüfungsausschuss

Vorsitzender: Prof. Dr. Martin Spieß

1. Dissertationsgutachter: Prof. Dr. Lars Schwabe

2. Dissertationsgutachter: Dr. Jan Gläscher

1. Disputationsgutachterin: PD Dr. Kirsten Hötting

2. Disputationsgutachter: Prof. Dr. Ulf Liszkowski

Acknowledgements

Preparing this section made me realize not only how much support I received over the last four years but also the sheer number of people that were, directly or indirectly, involved in this project. First, I would like to thank Prof. Dr. Lars Schwabe for introducing me to the world of academia, for his steady encouragement, and for his exceptional responsiveness to all my questions. Lisa K., Lisa D., Lisa W., Sabrina, Conny, and Mario, thanks for leading the way and sharing your knowledge with me. Franzi, Gundi, Nadine, and Felix, thanks for all the fun times I had the pleasure of sharing with you inside and outside the lab. Anna, I do not think I could have done this without you sitting next to me most of the time. Carlo, Valentina, Stefan, Anna-Maria, Hendrik, Leah, Kaja, and Li, thanks for being part of an awesome team that I always felt welcome in. Anke, Janis, Sandra, and Katharina, thank you for all the support not just regarding organizational matters. I would especially like to acknowledge the help of the many students who made this work possible by assisting with data collection along the way, including Friederike Baier, Stina Bange, Pia D'Agostino, Leandra Feldhusen, Jan-Ole Großmann, Denise Hartwig, Schmaila Kahn, Hülya Keskin, Manuel Krohn, Constantin Kuhlmann, Vincent Kühn, Moana Lamm, Annika Lutz, Fabian Schacht, Felix Schiborn, Celine Schneller, Anne-Sophie Siegel, Elizabeth Sievert, Rosann Stocker, Seher Teymuroglu, Till Thelosen, and Ricarda Vielhauer. Thank you all. It would also be hard to overstate the importance that my friends and family have played in this work. Knowing that I could always rely on you made this a hundred times easier. Thank you, Marcus, for reminding me that sometimes taking breaks from work is important. Sina, Paddi, Emmi, and Mathilda, you are by far the coolest young family in all of Hamburg. Thank you for always being there for me. Chrissi, Alex, and Lia, thanks for sharing your happiness with me. Finally, I am incredibly grateful to my parents for providing me with so much unconditional love and support. Thank you!

Abstract

Declarative memory aids behavioral adaptation by identifying predictors of important consequences. To ensure an efficient retrieval of such self-relevant information considering the overabundance of everyday impressions, long-lasting memory is formed relatively sparsely. The present work aimed to characterize principles under which this memory modulation operates in threatening environments. Prior studies have repeatedly shown an unspecific modulation of memory for stimuli surrounding salient experiences, but more recent evidence additionally points towards a more specific enhancement of memory that only affects items belonging to a motivationally significant category. In Study 1, we successfully replicated a category-specific ‘online’ memory enhancement, which was characterized by the superior memory for stimuli from a shock-associated category that were encoded during fear conditioning. This effect prospectively carried over to stimuli from the same category that were subsequently encoded without the threat of shock. However, our results cast doubt over claims that this category-specific memory modulation also retroactively affects stimuli from the shock-associated category that were encoded prior to fear conditioning. In Study 2, we proposed an even more specific modulation of memory that operates at the level of unique stimuli. In line with established models, we found that greater physiological arousal elicited by individual trial outcomes was linked with better subsequent memory performance. Critically, we present evidence for a novel cognitive account of memory modulation that goes beyond these influences of physiological arousal and is characterized by an improved memory for stimuli associated with surprising outcomes, which were formalized as aversive prediction errors (PEs). In Study 3, we aimed to characterize the neural basis of this PE-driven account using fMRI. Results suggested a mechanism that is distinct from expectancy-congruent modes of memory formation associated with an activation of medial-temporal structures and instead relies on the recruitment of the salience network. Overall, our results paint a nuanced picture of an adaptive memory system that uses multiple complementary strategies to ensure an efficient storage of self-relevant information.

Contents

1	Introduction	1
1.1	Models of memory modulation	2
1.1.1	Physiological approach	4
1.1.2	Cognitive approach	6
1.2	Levels of memory modulation	8
1.2.1	Unspecific	9
1.2.2	Category level	10
1.2.3	Exemplar level	12
1.3	Research goals	12
2	Experimental Studies	14
2.1	Study 1: How reliable is the category-specific retroactive enhancement of memory?	14
2.1.1	Background	14
2.1.2	Methods	14
2.1.3	Results	15
2.1.4	Conclusions	16
2.2	Study 2: Can aversive PEs promote memory for predictive items beyond the effects of arousal?	17
2.2.1	Background	17
2.2.2	Methods	17
2.2.3	Results	18
2.2.4	Conclusions	18
2.3	Study 3: What neural mechanism drives the memory-modulating effects of aversive PEs?	19
2.3.1	Background	19
2.3.2	Methods	19
2.3.3	Results	20
2.3.4	Conclusions	21

3	General Discussion	22
3.1	Unspecific, category-level, and exemplar-level memory modulation	23
3.2	Separating physiological from cognitive accounts of memory modulation . . .	26
3.3	A distinct mechanism underlying effects of aversive PEs on memory	28
3.4	Future directions	30
3.5	Conclusions	32
	References	34
A	Appendix A: Study 1	47
B	Appendix B: Study 2	86
C	Appendix C: Study 3	100

List of Figures

1	Schematic representation of episodic reinforcement learning	3
2	Amygdala-mediated modulation of memory for arousing experiences	4
3	Effects of unsigned reward PEs on memory in Experiment 1 of Rouhani et al. (2018)	7
4	Three levels of specificity in memory modulation	9
5	Modulation of memory in the 24h weak encoding group of Dunsmoor et al. (2015)	11
6	Procedure of Study 1 (Kalbe and Schwabe, in press)	15
7	Pooled results of Study 1 (Kalbe and Schwabe, in press)	16
8	Effects of unsigned PEs on recognition memory in Experiments 1 and 2 of Study 2 (Kalbe and Schwabe, 2020a)	18
9	Main behavioral findings of Study 3 (Kalbe and Schwabe, 2021)	20
10	Neural findings of Study 3 (Kalbe and Schwabe, 2021) that link effects of PEs with subsequent recognition memory	21
11	Integration of our results with the proposed framework of declarative memory modulation	32

List of Abbreviations

AIC	Akaike information criterion
BOLD	Blood oxygen level dependent
CI	Confidence interval
CR	Conditioned response
CS	Conditioned stimulus
dACC	Dorsal anterior cingulate cortex
fMRI	Functional magnetic resonance imaging
GLMM	Generalized linear mixed model
HC	Hippocampus
mPFC	Medial prefrontal cortex
MTL	Medial temporal lobe
MVPA	Multivoxel pattern analysis
PE	Prediction error
PHC	Parahippocampal gyrus
ROC	Receiver operating characteristic
SCR	Skin conductance response
UCS	Unconditioned stimulus

Introduction

Memory allows us to leverage the past to guide future behavior. Particularly episodic memory, a subtype of declarative memory that captures the contextual and spatial details of personal experiences, informs decision-making in complex environments (Anderson and Milson, 1989; Gershman and Daw, 2017; Klein et al., 2010; Shohamy and Adcock, 2010; Tulving, 1972). One specific strength of episodic memory lies in its potential to enable effective learning with only limited experience (Botvinick et al., 2019; Pritzel et al., 2017). Especially in threatening and aversive contexts, this learning from limited data proves useful, as it minimizes the exposure to potential harm. Consider an experienced police officer performing traffic stops as part of her daily routine. Almost always, her working day passes without any remarkable incidents. However, on a single occasion, a stopped driver suddenly became irritated and drew a firearm, quickly escalating the routine stop into a life-threatening situation. Although she was not harmed in the incident, identifying, and subsequently avoiding similarly dangerous situations is critical to long-term survival. Episodic memories are rich in contextual details and can therefore capture subtle cues of dangerous outcomes. Theories of adaptive memory assume that this utility for increasing reproductive fitness was a major evolutionary factor for the development of episodic memory (Klein et al., 2002; Nairne and Pandeirada, 2008a, 2008b). How can such an adaptive memory system inform decisions that ensure long-term survival?

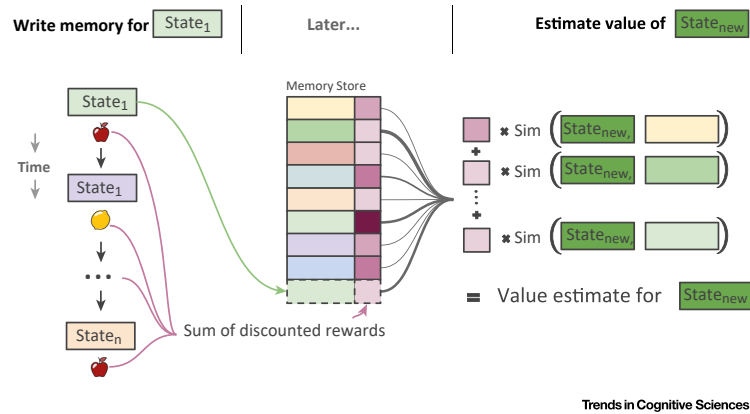
Despite its potential to guide behavior, episodic memory plays no explicit role in dominant computational models of sequential decision making, including model-free and model-based reinforcement learning (Doya et al., 2002; Kaelbling et al., 1996; Sutton et al., 1992). Instead, these models assume that learning about the potential value of a choice option in a given situation (conceptualized as a *state*) occurs incrementally (Botvinick et al., 2019). Specifically, relevant information of any new experience is extracted to update either cached state-action values (in model-free approaches) or parameters of an internal model from which state-action values can be dynamically computed (in model-based approaches). Afterwards, the remaining raw perceptual details of an experience can be discarded (Gershman and Daw, 2017). In behavioral tasks based on a Markov decision process, as commonly applied in experimental studies of reinforcement learning, an agent faces only a few discrete states that they revisit many times (Gershman and Daw, 2017; Maia, 2009). Complex real-life situations pose a more significant challenge. They typically include continuous states that are only partially observable and might only be visited once (Gershman and Daw, 2017; Niv et al., 2015). For example, every traffic stop is unique regarding its combination of potentially predictive

stimuli, such as the time of the day, the model of the stopped car, its condition, or the initial response of its driver. Further, while knowing whether a driver is concealing a gun heavily influences the threat-level estimate of the situation, the police officer can only learn about this fact after further investigation (i.e., it is initially unobservable; Whitehead and Lin, 1995).

Non-parametric learning based on episodic memory provides one solution to approximate values of actions despite these challenges (Gershman and Daw, 2017; Lengyel and Dayan, 2008; Figure 1). In contrast to conventional models of reinforcement learning, these models assume only a minimal processing of experiences at the time of encoding. Instead, both perceptual information of an experience, as well as its associated consequences (formalized as a sum of discounted rewards), are stored in episodic memory. Therefore, the rich information surrounding the event is retained and accessible later. At the same time, this approach relegates the computational burden to the time of decision: When a novel situation is encountered and a decision is to be made, relevant traces of past experiences are first retrieved from episodic memory. Then, each sampled memory trace is weighted based on its similarity to the current situation. An overall value estimate for any possible action in the given situation can finally be obtained by summing the similarity-weighted associated consequences of the retrieved experiences (Gershman and Daw, 2017). A direct result of this model's mode of value computation is that the set of retrievable memories at decision time has a critical impact on subsequent decisions. Therefore, it might seem desirable to non-selectively store any personal experience in long-term memory, which would ensure that subsequent decision making can be informed by the broadest possible data basis. At the same time, long-lasting episodic memory is known to only be formed for selective events. How can this apparent contradiction be resolved?

1.1 Models of memory modulation

Considering the many experiences that people make every day, only a small proportion of them will be recallable over long periods of time. For example, unless anything extraordinary happened, the police officer will typically not remember any routine traffic stop that she performed several weeks ago. On the other hand, she still vividly remembers the single incident in her career that developed into a life-threatening situation when an armed driver aggressively confronted her. Why is stable memory formed only for certain experiences? Intuitively, one explanation could lie in the limited storage capacities of long-term declarative memory. Although estimating the exact capacity of long-term memory is challenging and depends on various assumptions, the finite number of possible synapses in the brain must ultimately pose an upper limit (Dudai, 1997). On the other hand, prior studies have demonstrated the enormous capacity of human declarative memory (Brady et al., 2008; Standing, 1973). An alternative explanation could lie in limited capacities to



Trends in Cognitive Sciences

Figure 1. Schematic representation of episodic reinforcement learning (Gershman and Daw, 2017). Experiences including both perceptual details as well as associated consequences (formalized as a sum of discounted rewards) are stored in episodic memory. In a novel situation, value estimates can be computed by summing previous consequences of stored episodic memories, each weighted by an estimate of their closeness to the current situation. As only such experiences that have been stored in long-term memory can inform value estimates, this model emphasizes the role of selective memory formation for subsequent decisions based on episodic memory. Reprinted from Botvinick et al. (2019). Licensed under CC BY 4.0.

form new memories. Indeed, the sensory systems (e.g., the visual system) act as an early filter to memory formation by only attending to a subset of the potential sensory input of any experience (Carrasco, 2011; Humphreys et al., 1998). Available bandwidths in both encoding and consolidation further restrict the amount of new information that can enter long-term memory over any period (Feld et al., 2016; Fukuda and Vogel, 2019). Interestingly, existing memories can sometimes even be selectively forgotten if they prove unreliable (Kim et al., 2014). Such pruning might serve adaptive memory by removing irrelevant or incorrect information and therefore give more weight to memories that proved reliable. This provides an intriguing perspective where the modulation of memory formation is not just because of limited encoding and consolidation resources but reflects the need for an efficient retrieval of self-relevant and reliable information.

As decision-making based on episodic memory requires cognitive resources to compute optimal choices dynamically at decision time, this process associates a cost with each additional episode that needs to be processed, which should optimally be outweighed by its informational gain (Anderson and Milson, 1989; Anderson and Schooler, 2000; Kuhl et al., 2007). Sampling only a few experiences from memory for each decision seems to be one way by which the brain satisfies this requirement (Bornstein et al., 2017; Bornstein and Norman, 2017). While this process optimizes computational efficiency, it also dictates that each sampled episode contains useful information while avoiding redundancy. Distinguishing such useful experiences from those with little relevance poses a significant challenge. We here present two mechanisms that have empirically been shown to modulate memory formation:

one based on physiological arousal elicited by an experience, the other based on a cognitive evaluation of informational measures associated with an experience.

1.1.1 Physiological approach

It has long been known that details surrounding highly emotional events, such as the police officer being confronted by an armed driver, are exceptionally well remembered (Christianson and Loftus, 1987; LaBar and Cabeza, 2006; Schwabe et al., 2012). This promotion of memory formation is not just limited to the emotional event itself, but also extends to stimuli encoded within the temporal context of an emotional event, even in the absence of a causal link between these two. At the neurobiological level, the amygdala, and particularly its basolateral subregion, have been identified as central hubs for the modulatory effects of emotional arousal on long-term memory consolidation (McIntyre et al., 2003; Roozendaal et al., 2006).

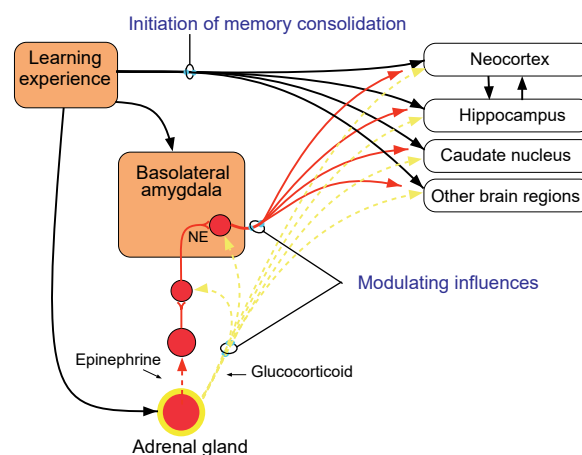


Figure 2. Amygdala-mediated modulation of memory for arousing experiences. An emotional learning experience (upper left corner) induces arousal, which prompts the adrenal gland to release glucocorticoids (e.g., cortisol) and catecholamines (e.g., (nor-)epinephrine). The interaction of glucocorticoids with norepinephrine in the basolateral amygdala modulates memory consolidation through its projections to areas involved in long-term memory formation such as the hippocampus and prefrontal cortex. Reprinted from McGaugh (2000). Reprinted with permission from AAAS.

One popular model of arousal-mediated memory modulation focuses on concurrent glucocorticoid and noradrenergic activity in the basolateral amygdala in reaction to emotional experiences (Cahill and McGaugh, 1998; McGaugh and Roozendaal, 2002). The basolateral amygdala then strengthens consolidation processes through its projections to memory-critical regions (McGaugh, 2000; Paré, 2003; Roozendaal et al., 2006; Figure 2). While adrenaline and noradrenaline are released rapidly via the sympathetic nervous system in response to an emotional experience, the slower release of corticosteroids (including cortisol) from the adrenal cortex takes several minutes to reach significantly elevated levels (de Kloet et al.,

2005; Joëls et al., 2012). Accordingly, findings of trial-by-trial differences in memory due to varying emotional responses to individual stimuli have been attributed to a fast noradrenergic modulation rather than the slower action of corticosteroids (Bergt et al., 2018; Christianson et al., 1991; Richardson et al., 2004; Strange and Dolan, 2004). At the neurobiological level, this alternative view emphasizes arousal-induced locus coeruleus activity that leads to the secretion of noradrenaline (Mather et al., 2016). Key target regions for the memory-modulating effects of arousal include the hippocampus and adjacent structures of the medial temporal lobe (MTL), which have since the classic case of the neurosurgical patient H.M. been regarded as central for the formation of declarative memory (Scoville and Milner, 1957; Squire, 2009; Squire and Zola-Morgan, 1991). Modern neuroimaging techniques have since corroborated this critical role of the MTL in declarative memory formation. This includes studies that were able to directly link MTL activation during encoding to subsequent memory performance (Davachi and Wagner, 2002; Eichenbaum, 2004; Fernández et al., 1999). In response to emotional events, the basolateral amygdala exerts further modulatory effects on the medial prefrontal cortex (mPFC), which is involved in higher-order cognitive and affective processing (Barsegyan et al., 2010; Frith and Dolan, 1996), but also interacts with the hippocampus in memory consolidation (Preston and Eichenbaum, 2013). Previous studies have further shown that the effects of glucocorticoids on memory consolidation depend on bidirectional interactions between the basolateral amygdala and mPFC (Barsegyan et al., 2010; Roozendaal, McReynolds, et al., 2009).

A critical issue when testing the physiological account of declarative memory modulation is the operationalization of arousal. Commonly, studies make use of skin conductance as the peripheral measure of physiological arousal, which reflects the regulation of sweat gland activity by the sympathetic nervous system (Dawson et al., 2011). The continuous skin conductance signal can further be decomposed into a tonic, slow-varying component and a phasic, fast-varying component (Benedek and Kaernbach, 2010a, 2010b). The slower, tonic component reflects overall states of physiological arousal. The faster, phasic component reflects skin conductance responses (SCRs) to individual stimuli. These phasic SCRs have been linked to amygdala-based processing that is typically associated with the improved memory for emotional stimuli (Williams et al., 2001). Therefore, the physiological approach predicts stronger memory enhancement for stimuli that elicit greater phasic SCRs. Such physiological responses are typical for emotional experiences, but not their sole defining feature (Lench et al., 2011; Mauss and Robinson, 2009). Particularly, this arousal-centered view largely ignores the cognitive component of emotional processing. From such an alternative cognitive perspective, emotional events can often be further characterized by their unpredictability and subsequent experiences of surprise (Trapp et al., 2018).

1.1.2 Cognitive approach

An influential formalization of the concepts surrounding surprise and uncertainty was introduced by Shannon (1948) in his works founding the field of information theory (Atick, 1992). Although originally developed as a theory of information transmission in communication systems, the underlying principles have since been applied to a wider field of disciplines, including psychology and neuroscience (Berlyne, 1957; Borst and Theunissen, 1999). It assumes a system of discrete states, each associated with a known probability. The amount of information contained within the occurrence of a single state is then defined as the negative logarithm of its associated probability (Lombardi et al., 2016; Shannon, 1948). Therefore, rare, or unexpected events are associated with greater surprise and carry more information. While information theory links surprise to the observation of individual events, the Shannon entropy, or uncertainty, reflects the expected value of surprise over all possible events and is a property of the random process rather than individual outcomes (Strange et al., 2005).

A striking observation in human cognition is that identical outcomes can subjectively elicit different levels of surprise when an individual's beliefs about the underlying contingencies change (Itti and Baldi, 2009; Knill and Pouget, 2004). For example, a seemingly harmless bumper sticker would evoke different predictions about the outcome of a traffic stop depending on the police officer's knowledge that its symbolism is used by a violent political movement. Cognitive theories of predictive coding account for this fact by assuming a mutable internal model of the world from which top-down predictions of expected sensory input are generated (Friston, 2018; Rao and Ballard, 1999). Comparing these predictions against observed sensory data in a bottom-up fashion allows the internal model to be updated and generate increasingly accurate predictions, which minimizes future surprise and increases coding efficiency. Evidence for such error-driven learning has been found in various domains of perception and cognition, including visual perception (Hosoya et al., 2005; Rao and Ballard, 1999) and auditory perception (Baldeweg, 2006; Heilbron and Chait, 2018), but also attention (Spratling, 2008) and reward preferences (O'Doherty et al., 2006). Some authors have therefore argued that this process of contrasting generated predictions against perceptual evidence in order to improve an internal model might be a unifying principle of neural computation (Clark, 2013; Friston, 2010).

At the core of the predictive coding account lie so-called *prediction errors (PEs)*, which provide a numerical measure of the difference between an expected signal and the actual outcome. Prominently, they enable learning incrementally about associative relationships in temporal difference approaches to reinforcement learning (Sutton, 1988) or the Rescorla-Wagner model of Pavlovian conditioning (Miller et al., 1995; Rescorla and Wagner, 1972). For example, when applied to Pavlovian fear conditioning, the Rescorla-Wagner model can explain how an individual incrementally learns to associate an initially neutral stimulus with a probabilistically linked unconditioned stimulus (UCS; e.g., an aversive shock). Over the

course of the conditioning procedure, the neutral stimulus turns into a conditioned stimulus (CS), whose presentation alone induces a conditioned response (CR). To explain this learning process, the Rescorla-Wagner model assumes for any given trial t , that the associative value V_{x_t} of a CS x is updated through a trial-unique PE (σ_t) weighted by a constant learning rate (α ; Miller et al., 1995; Niv and Schoenbaum, 2008):

$$V_{x_{t+1}} = V_{x_t} + \alpha \sigma_t \quad (1.1)$$

The PE σ_t in trial t is defined as the difference between the value of the outcome of this trial (R_t) and the sum of predictions from all available stimuli in this trial (V_{total_t}):

$$\sigma_t = R_t - V_{total_t} \quad (1.2)$$

Intuitively, in the Rescorla-Wagner model, outcomes that exceed predictions therefore produce positive PEs, which in turn increase the associative strength between the CS and UCS. Predictions that exceed outcomes produce negative PEs, which decrease the associative strength between the CS and UCS. Notably, the idea of using discrepancy signals to drive learning is not bound to one specific conceptualization of PEs (Niv and Schoenbaum, 2008). For example, other learning models might implement an unsigned PE, where, unlike in the Rescorla-Wagner model, only the magnitude, but not the direction of the discrepancy between predictions and outcomes is relevant.

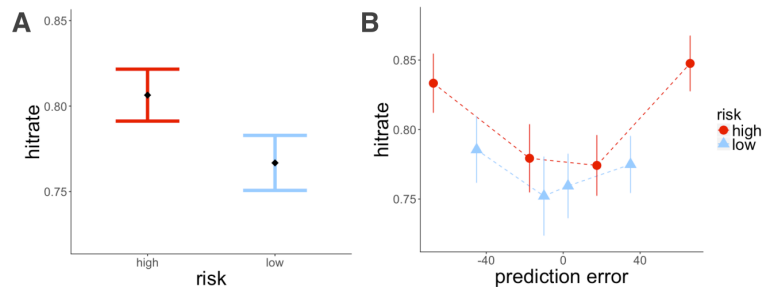


Figure 3. Memory formation scaled with risk and quadratic unsigned reward PEs in Experiment 1 of Rouhani et al. (2018). Participants learned to predict rewards associated with unique stimuli in either a low- or high-variance condition. When stimuli were encoded in a high-variance environment, the hit rate in a subsequent memory test was overall increased compared with stimuli encoded in a low-variance environment. Importantly, greater discrepancies between predicted and received rewards (i.e., unsigned reward PEs) increased recognition performance at the exemplar-level. The observed U-shaped relationship between PEs and hit rates implies that both under- and overestimating rewards had similar memory-promoting effects. Reprinted from Rouhani et al. (2018). Reprinted with permission from APA.

Despite their prominent role in many cognitive learning models, PEs have historically received little attention in the field of declarative memory modulation. Only recently, and mostly in the reward domain, have the memory-modulating effects of novelty and PEs been explored in more detail (Ergo et al., 2020; Greve et al., 2017; Jang et al., 2019; Quent et al.,

2021). In one such study, participants saw a series of unique pictures and learned to predict rewards associated with them (Rouhani et al., 2018; Figure 3). The authors found that greater discrepancies between predicted and received rewards were associated with improved memory for the associated pictures in a subsequent recognition test. This memory-promoting effect of PEs was unsigned: Both the under- and overestimation of rewards enhanced memory in a similar (quadratic) fashion.

Applying this idea to the aversive domain, we propose that, in addition to driving incremental learning processes that associate a predictive stimulus (CS) with a conditioned response (UCS; Miller et al., 1995; Rescorla and Wagner, 1972), a secondary function of aversive PEs lies in prioritizing stimuli with surprising outcomes for preferential storage in long-term declarative memory (Trapp et al., 2018). A proposed mechanism behind this surprise-driven memory modulation lies in PEs' ability to create event boundaries (Rouhani et al., 2020). Rather than simply enhancing 'standard' processes of memory formation associated with the MTL, PEs indicate that new information is incongruent with existing knowledge structures represented by the schema network, which includes the mPFC, angular gyrus, and precuneus (van Kesteren et al., 2012; Vogel et al., 2018). Therefore, events associated with high PEs might create a separate memory trace rather than being integrated into existing schema memory. Particularly the neurotransmitter dopamine might play a key role in this PE-driven memory enhancement (Shohamy and Adcock, 2010). It is well established that activation patterns of dopaminergic midbrain neurons in the ventral tegmental area and substantia nigra track signed reward PEs (Lak et al., 2014; Schultz et al., 1997). Further studies demonstrated PE-related signaling in a variety of brain regions, most prominently in the striatum (Gläscher et al., 2010; O'Doherty et al., 2003; Pagnoni et al., 2002; Pessiglione et al., 2006), which receives dense dopaminergic input, but also the frontal cortex, especially the dorsal anterior cingulate cortex (dACC; Schultz, 2016; Schultz et al., 1998; Seo and Lee, 2007; Silvetti et al., 2011). While the mesolimbic dopaminergic system has also been theorized to be involved in the signaling of aversive PEs (Brooks and Berns, 2013; Matsumoto and Hikosaka, 2009), its neuroendocrinological substrate is overall more controversial (Fiorillo, 2013; Schultz, 2019). It should also be noted that no study has yet directly demonstrated the critical involvement of dopamine in the PE-driven modulation of declarative memory.

1.2 Levels of memory modulation

Irrespective of the driving forces behind declarative memory modulation, these effects can emerge at different levels of specificity. After her life-threatening encounter with an armed driver, which contents of the experience should the police officer preserve in long-term memory? Should the memory promotion be limited to details of the armed driver? Or should it also cover impressions preceding the violent encounter, even if they seem inconsequential?

Here, we propose three levels of specificity, reaching from a completely unspecific memory modulation, over such memory modulation that is limited to stimuli from a specific category, to a highly specific memory modulation that only applies to unique exemplars (Figure 4). As these levels of specificity are orthogonal to the question of underlying mechanisms, both the physiological and the cognitive models of memory modulation are in principle compatible with all three levels of specificity.

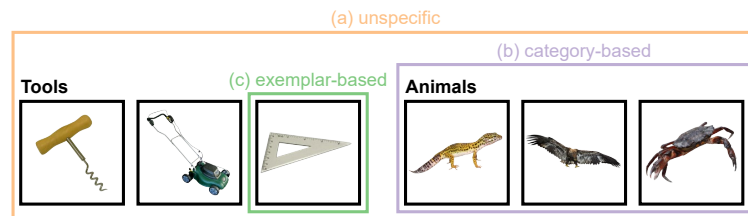


Figure 4. Three levels of specificity in memory modulation. In an exemplary encoding task, participants see unique pictures of animals and tools. (a) In an unspecific modulation of memory, no semantic link between a memory-promoting event and the promoted learning material is necessary. Instead, the mere temporal proximity of a stimulus to a salient event produces superior memory. For example, a group of participants undergoing an experimental stress induction after encoding might show, relative to an unstressed control group, an overall improved memory for previously encoded stimuli of both animals and tools. (b) In a category-based memory modulation, memory for some stimuli, but not others, will be enhanced based on the motivational significance of their respective category. For example, memory for pictures from a category associated with aversive electric shocks might be enhanced compared with pictures from a control category that is not associated with any aversive shocks (Dunsmoor et al., 2015). (c) The most specific form of memory modulation is exemplar-based. Here, memory for individual pictures is modulated through trial-unique learning signals, such as phasic physiological arousal or associated PEs. Example stimuli stem from the Bank of Standardized Stimuli (Brodeur et al., 2010), licensed under CC BY-SA 3.0.

1.2.1 Unspecific

In its most encompassing form, a modulation of memory can be semantically unspecific. A typical experiment investigating the effects of stress on memory consolidation involves participants viewing a series of stimuli for which memory will be tested at a later stage. If they are confronted with a stress induction procedure in a critical time window before or after encoding, their memory for these stimuli will typically be improved compared with a non-stressed control group (Nater et al., 2007; Roozendaal, 2002; Schwabe et al., 2008; Schwabe et al., 2012; Wolf, 2012). Even though this memory promotion seems to be more pronounced for emotional than for neutral material (Cahill et al., 2003; Smeets et al., 2008), the effect is still unspecific in the sense that no semantic link between enhanced stimuli and the stress procedure that triggers the memory modulation is necessary. To explain these findings, the neuroendocrinergic model of stress effects on memory consolidation focuses on the effects of glucocorticoids and adrenal stress hormones in the basolateral amygdala, as detailed in the physiological model above (Roozendaal, McEwen, et al., 2009).

A similar line of research focuses on the level of individual neurons, where long-term potentiation poses the dominant cellular model of long-term memory (Frey and Morris, 1997; Nicoll and Roche, 2013). According to the *synaptic tagging and capture hypothesis* (Frey and Morris, 1997, 1998; Martin and Kosik, 2002; Rogerson et al., 2014), this requires two separable processes to occur in hippocampal neurons that lead to prolonged increases in synaptic strength. First, a transient tag is set in a synapse after the stimulation of a neuron. While relatively weak stimulation of the neuron can be sufficient for this step, associated physiological changes are only short-lived (i.e., not lasting longer than a few hours). To achieve long-lasting changes in synaptic signaling, a learning tag set in the previous step needs to capture additional plasticity-related proteins that are only synthesized in the cell body of the neuron after stronger stimulation. An important implication of this model is that weak stimuli that would by themselves not trigger protein synthesis can still be transformed into long-lasting memories when they capture necessary plasticity-related proteins produced by a separate, stronger stimulus before their learning tag has decayed. Applying these principles to the behavioral level, studies reporting *behavioral tagging* have shown in both rodents and humans that memory for neutral stimuli could be unspecifically promoted through a subsequent salient event that has no semantic link to the promoted stimuli (Ballarini et al., 2013; Ballarini et al., 2009; Moncada et al., 2015).

1.2.2 Category level

Although an unspecific modulation of memory formation minimizes the chance that important indicators of self-relevant outcomes are missed, it will also promote memory for many non-informative stimuli. For example, after a traffic stop that turned violent, the police officer might also experience a promotion of memory for inconsequential details, such as the brand of coffee she was drinking before initiating the stop. To reduce the number of experiences in long-term memory that bear little predictive value, an alternative mode of memory modulation can be limited to a category of stimuli that proved to signal motivationally significant outcomes. In three separate encoding phases, Dunsmoor et al. (2015) presented healthy participants a series of unique pictures of animals and tools. Only in the second intermediate phase, pictures from one category (e.g., animals) were paired with a mild electric shock in two-thirds of all trials (CS^+) while the remaining category (e.g., tools) was never paired with a shock (CS^-). In both encoding phases before fear conditioning (pre-conditioning) and after fear conditioning (post-conditioning), there was no threat of shock, as no shock leads were attached. Participants' recognition memory for pictures from all three encoding phases was probed either immediately after the last encoding phase, 6h, or 24h later. This design illustrates the three temporal directions in which a memory modulation can operate: In a *retroactive* promotion of memory, previously encoded memories are enhanced through subsequent experiences. In an *online* memory promotion, encoding and memory

promotion occur within the same temporal context. Finally, in a *prospective* promotion of memory, the memory-promoting event precedes the encoding of the promoted stimuli.

Regardless of the interval between encoding and retrieval, the authors found a memory advantage for pictures from the CS⁺ category that were encoded during fear conditioning over pictures from the CS⁻ category that were encoded in the same phase (i.e., an online effect). In the 24h delay group, the authors further found both a category-specific prospective and a category-specific retroactive promotion of memory for CS⁺ over CS⁻ items (Figure 5). Particularly the latter retroactive effect is remarkable, as during the encoding of stimuli in the pre-conditioning phase, participants had no way of knowing which item category would be subsequently linked with aversive shocks. Based on the additional observation that the 6h delay group showed evidence for a category-specific *retroactive* memory enhancement, but, unlike the 24h delay group, not for the category-specific *prospective* memory enhancement, the authors further speculated that both effects might depend on separate mechanisms. For the specific retroactive effect, they referred to the tag-and-capture hypothesis, although such behavioral tagging effects had previously only been observed in the form of an unspecific memory enhancement.

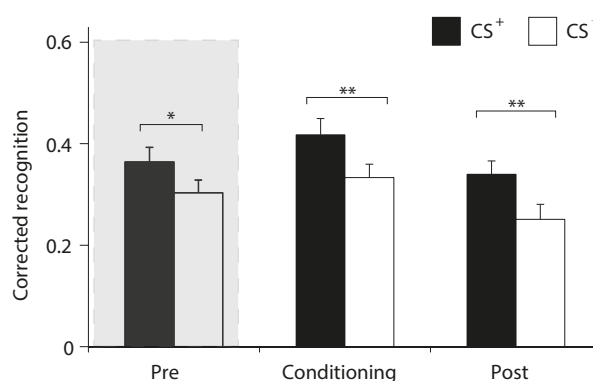


Figure 5. Modulation of memory in the 24h weak encoding group of Dunsmoor et al. (2015). In each of three encoding phases, participants saw unique pictures of animals and tools. In the intermediate fear conditioning phase, pictures from one category (CS⁺) were paired with a mild electric shock, but not pictures from the other category (CS⁻). Results showed an ‘online’ memory enhancement for items from the CS⁺ category over items from the CS⁻ category that were encoded during fear conditioning. Critically, this memory enhancement for items from the CS⁺ category carried over to both the pre-conditioning phase (i.e., category-specific retroactive memory enhancement) and the post-conditioning phase (i.e., category-specific prospective memory enhancement) even though these items were themselves never paired with any shocks. * $p < .05$, ** $p < .01$. Reprinted from Dunsmoor et al. (2015). Reprinted with permission from Springer Nature Customer Service Centre GmbH.

Since Dunsmoor et al. (2015) provided first evidence for the category-specific retroactive memory effect, several follow-up studies have been published, including two conceptual replications aiming to translate the effect into the reward domain. Replacing aversive shocks with monetary rewards but sticking otherwise close to the classical conditioning procedure of Dunsmoor et al. (2015) produced both an online category-specific memory modulation as

well as the category-specific prospective memory effect after a 24h consolidation period, but provided no evidence for the category-specific retroactive modulation of memory (Oyarzún et al., 2016). Utilizing instead a delayed match-to-sample task with low- versus high rewards in an intermediate reward phase, Patil et al. (2017) showed a category-specific retroactive memory enhancement for items from the subsequently highly rewarded category, even when those items were encoded before any rewards were introduced. To date, no independent group of researchers has tried to closely replicate these reports of category-specific retroactive memory enhancement.

1.2.3 Exemplar level

At the highest level of specificity, the memory modulation can be limited to a single stimulus. Here, it is important to distinguish modulating effects that were evoked by the stimulus itself from those that are caused by a separate memory modulating event. For example, the occurrence of a particular stimulus might by itself induce physiological arousal or surprise during encoding. In this case, there is no separation between the triggering event and the target of the memory modulation. Such effects have been observed as a superior memory for surprising stimuli in oddball paradigms, where a sequence of uniform stimuli is occasionally interrupted by another deviant (i.e., surprising) stimulus (Cycowicz and Friedman, 2007; Strange and Dolan, 2004; Strange and Dolan, 2001). Compared with this simple memory enhancement for the surprising occurrence of rare stimuli, a memory modulation during associative learning presents a more significant challenge. In the case of the police officer faced with a violent traffic offender, an adaptive memory system would not only call for a promotion of the violent behavior itself but should also ensure that predictors of these threatening outcomes are stored in long-term memory. Therefore, predictive cues of surprising or arousing outcomes are the focus of this mode of memory modulation, rather than the outcome itself. Typically, this mode of memory modulation can also be characterized by a temporal separation, where the target of the memory modulation precedes the triggering event. Such effects have recently been investigated for stimuli associated with reward PEs (Ergo et al., 2020; Jang et al., 2019; Rouhani et al., 2018). Whether similar effects occur for stimuli associated with surprising outcomes in the aversive context is currently unknown. In addition to this cognitive perspective based on PEs, the physiological approach to memory modulation hypothesizes that phasic physiological arousal to individual outcomes drives the memory modulation for associated cues at item level.

1.3 Research goals

Although it has long been known that salient events in temporal proximity to an experience modulate declarative memory formation, most previous research focused on unspecific

effects. Only recently, evidence has emerged that such memory modulation can specifically prioritize memory for stimuli from a motivationally significant category (Dunsmoor et al., 2015; Patil et al., 2017). These specific effects extended in all three temporal directions: to stimuli from the relevant category that were encoded *during* salient events (i.e., online), *after* salient events (i.e., prospectively), and even to stimuli that were encoded minutes *before* salient events (i.e., retroactively). Particularly the category-specific retroactive memory promotion would have drastic implications for our understanding of memory consolidation. However, behavioral evidence for this phenomenon is so far based on only a few studies from a single group of researchers. In *Study 1*, we initially set out to closely replicate the first study providing evidence for a category-specific retroactive strengthening of memory (Dunsmoor et al., 2015). Overall, results from four adequately powered close replications cast doubt on the reliability of this effect.

In contrast, the category-specific online modulation of memory proved reliable across all four experiments. Established physiological models of memory modulation attribute this effect to increased arousal associated with salient outcomes. Inspired by recent theoretical advances that emphasized the role of expectancy violations for the superior memory of emotional events (Trapp et al., 2018), in *Study 2* we asked whether an alternative cognitive model based on aversive PEs can explain these findings. To this end, we reanalyzed data from *Study 1* for potential physiological and cognitive drivers of memory formation at the exemplar level. Based on the large body of evidence for arousal-based memory modulation, we predicted that increased phasic SCRs close to the encoding of an item would be associated with greater recognition performance. Following a cognitive approach to memory modulation, we further hypothesized that unsigned PEs derived from participants' explicit shock expectancy ratings could explain memory formation beyond arousal-based effects.

Study 2 provided evidence for exemplar-level effects of both physiological arousal and unsigned PEs on memory formation in an aversive learning task. Interestingly, we found that both effects were statistically dissociable. While effects of arousal on memory formation have been linked to the amygdala and its modulatory influences on memory-critical regions, including the MTL and the mPFC, the neural mechanism behind the declarative memory promotion triggered by aversive PEs is still largely unknown. In *Study 3*, we addressed this issue by recording brain activity during memory encoding using functional magnetic resonance imaging (fMRI). We further improved upon the behavioral task such that it allowed us to assess participants' prediction uncertainty as well as derive continuous, rather than binary, PEs. To explain the neural underpinnings of modulating influences of aversive PEs on memory formation, we considered two competing models. The first model hypothesizes that PEs strengthen pathways associated with schema-congruent memory formation, which prominently include the MTL and mPFC. Alternatively, PEs might promote memory by inducing a qualitative shift in mnemonic processing, which might even lead to a decreased activation of regions classically associated with declarative memory formation.

Experimental Studies

2.1 Study 1: How reliable is the category-specific retroactive enhancement of memory?

Kalbe, F., & Schwabe, L. (in press). On the search for a selective and retroactive strengthening of memory: Is there evidence for category-specific behavioral tagging? *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001075> – (Appendix A)

2.1.1 Background

When seemingly mundane experiences relate to important consequences, an adaptive memory system should enable their preferential storage in long-term memory (Klein et al., 2010; Shohamy and Adcock, 2010). The synaptic tag-and-capture hypothesis provides a neurobiological framework for such a mechanism (Frey and Morris, 1997, 1998; Martin and Kosik, 2002; Rogerson et al., 2014). Applications to the behavioral level have shown such retroactive memory enhancement for stimuli preceding salient or stressful events (Ballarini et al., 2013; Ballarini et al., 2009; Moncada et al., 2015). Critically, this retroactive memory enhancement was exclusively unspecific, in the absence of a causal or semantic link. However, recent evidence suggests that the retroactive enhancement of memory can also be limited to stimuli belonging to a category that gains motivational significance by its association with aversive shocks (Dunsmoor et al., 2015) or monetary rewards (Patil et al., 2017). While such a specific retroactive memory enhancement would have considerable implications for our understanding of declarative memory formation, these effects have not yet been replicated by an independent group of researchers. Therefore, we aimed here to closely replicate the study that provided the initial evidence for a category-specific retroactive enhancement of memory for neutral stimuli from a category that later predicted the occurrence of aversive shocks (Dunsmoor et al., 2015).

2.1.2 Methods

Across four experiments including data from 285 unique participants, we closely replicated Dunsmoor et al. (2015) regarding aspects such as the procedure, stimuli, and statistical analysis. In short, participants saw a series of unique pictures of animals and tools across three separate encoding phases (Figure 6). Only the second, intermediate phase featured

a fear conditioning procedure where most stimuli of one category (CS^+), but not the other (CS^-), were followed by a mild shock. Consistently across all four experiments, we used a 24h interval between encoding and the following surprise recognition test, which had previously shown the most robust evidence for the category-specific retroactive memory effect.

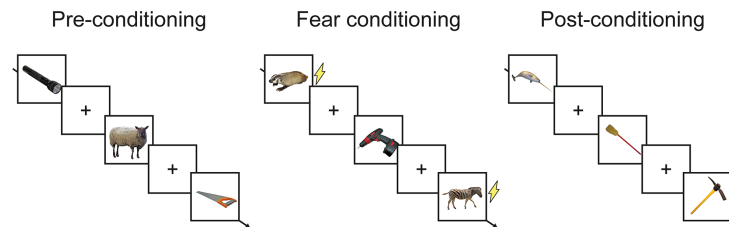


Figure 6. Procedure of Study 1, which closely replicated Dunsmoor et al. (2015). Participants saw unique pictures of animals and tools across three separate encoding phases. Only in the intermediate fear conditioning phase were pictures from one category (CS^+), but not the other category (CS^-), paired with a mild electric shock in two-thirds of all trials. Whether pictures of animals or tools were paired with shocks was counterbalanced across subjects. We probed recognition memory for pictures from all three phases 24h after encoding, which had previously produced the most consistent evidence of category-specific retroactive memory enhancement. Reprinted from Kalbe and Schwabe (in press).

For three out of four experiments, an a priori power analysis indicated a statistical power of $> 95\%$ to detect a category-specific retroactive memory enhancement based on the effect size reported by Dunsmoor et al. (2015). Experiment 4 was additionally pre-registered and pre-reviewed. Our statistical analysis closely replicated Dunsmoor et al. (2015), but also extended it substantially by including alternative measures of memory performance, Bayesian statistics, and a pooled analysis across all four replications.

2.1.3 Results

In all four experiments, we could consistently replicate the online memory advantage for CS^+ items over CS^- items that participants encoded during fear conditioning. However, when we strictly replicated the analysis strategy from Dunsmoor et al. (2015), none of our four experiments provided any evidence of a category-specific retroactive enhancement of memory. Parallel Bayesian analyses consistently favored the null hypothesis rejecting category-specific retroactive memory enhancement, with substantial evidence for the null hypothesis in two of the four experiments. In an exploratory analysis focusing only on high confidence recognition memory, we found a small but significant category-specific retroactive effect in Experiment 2, a trend towards this effect in Experiment 4 ($p = .088$), and no evidence for this effect in Experiments 1 and 3. Parallel analyses on the alternative recognition measure of d' from signal detection theory (Macmillan and Creelman, 2004; Wickens, 2002) provided no evidence for category-specific retroactive memory enhancement in any of the four experiments, even when only high confidence recognition was included. Similarly, in a pooled analysis using linear

mixed models (Figure 7), only when analyzing high confidence memory based on corrected recognition scores, we found a small but significant category-specific retroactive memory effect.

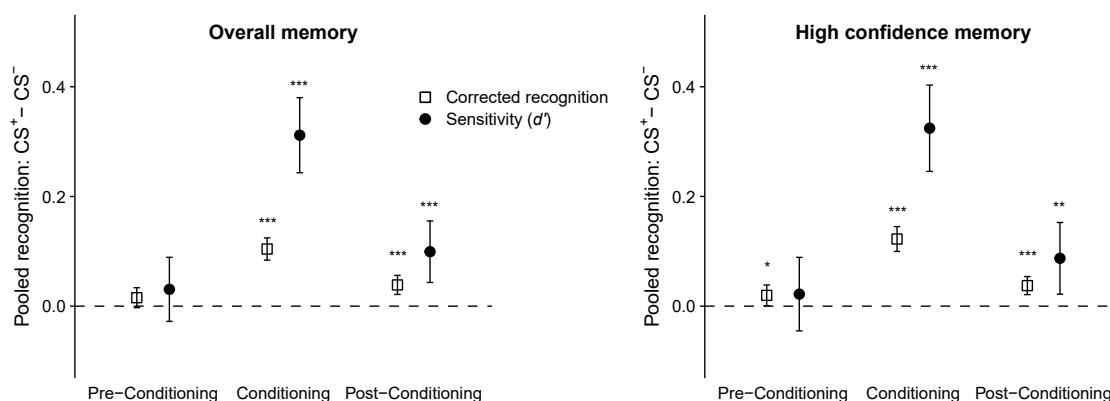


Figure 7. Results of a pooled analysis of data from all four experiments of Study 1 using linear mixed models. The left panel shows results when recognition memory was collapsed across confidence, while the right panel shows parallel results when only high confidence recognitions were treated as hits. Both *corrected recognition* (parallel to the analysis in Dunsmoor et al., 2015) and d' from signal detection theory are displayed as measures of recognition memory performance. Results provided support for both a category-specific online memory enhancement, as well as a category-specific prospective memory enhancement. Significant evidence of the critical category-specific retroactive effect was only found for high confidence memory and only in corrected recognition, but not in d' . Therefore, these results cast doubt on the reliability and generalizability of the specific retroactive enhancement of memory. Data show the fixed effect with error bars reflecting the 95% confidence interval (CI). Dashed lines show expected effects under the null hypothesis. * $p < .05$, ** $p < .01$, *** $p < .001$. Reprinted from Kalbe and Schwabe (in press).

2.1.4 Conclusions

Across four close replications, we found no evidence for the category-specific retroactive memory effect when we strictly replicated the original analysis strategy from Dunsmoor et al. (2015). Only for high confidence memory, uncorrected for multiple comparisons, and only in corrected recognition scores, we found a significant category-specific retroactive memory enhancement in one of four experiments and in a pooled analysis across experiments. That this effect was not detectable in parallel analyses on memory sensitivity (d'), which is the empirically better supported measure of recognition memory (Dube and Rotello, 2012; Pazzaglia et al., 2013; Wixted, 2007), raises further doubts about the generalizability of the putative retroactive category-specific memory effect. On the other hand, we found reliable evidence of both the category-specific online enhancement of memory and the category-specific prospective enhancement of memory.

2.2 Study 2: Can aversive PEs promote memory for predictive items beyond the effects of arousal?

Kalbe, F., & Schwabe, L. (2020a). Beyond arousal: Prediction error related to aversive events promotes episodic memory formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 234-246. <https://doi.org/10.1037/xlm0000728> – (Appendix B)

2.2.1 Background

In Study 1, we observed a robust memory modulation for items from the CS⁺ category over items from the CS⁻ category that were encoded online during fear conditioning. Dunsmoor et al. (2015) provided an arousal-based account of this effect (Cahill and McGaugh, 1998; McGaugh, 2018). However, outcomes of CS⁺ trials were not only associated with higher levels of arousal, but their outcomes were also partly unpredictable, while CS⁻ trials never resulted in a shock. From a cognitive perspective, CS⁺ items were therefore associated with higher levels of uncertainty and a greater number of PEs (Trapp et al., 2018). We aimed here to investigate whether these aversive PEs drive memory formation at the level of individual items. Evidence for such a mechanism would parallel recent results from the reward domain (Jang et al., 2019; Rouhani and Niv, 2021; Rouhani et al., 2018). In line with arousal-based models, we further predicted that both anticipatory arousal (in reaction to the stimulus and the anticipation of a shock), as well as outcome-related arousal (in reaction to the shock or its omission), would enhance subsequent memory. Finally, we tested whether these putative effects of PEs and arousal on memory are statistically separable.

2.2.2 Methods

We reanalyzed data from Experiment 1 ($N = 44$ healthy participants) and Experiment 2 ($N = 84$ healthy participants) of Study 1. As we were interested in the effects of aversive PEs on subsequent memory formation, our analyses focused only on recognition memory for the 60 pictures of animals and tools that were encoded in the intermediate fear conditioning phase. PEs occurred when participants either indicated that they expected a shock, but none was administered (unexpected shock omission) or when they indicated that they did not expect a shock, but one was administered (unexpected shock). Arousal was operationalized separately through anticipatory and outcome-related SCRs. To explain memory formation at the level of individual items, we fitted generalized linear mixed models (GLMMs) with a binary response function coding whether participants recognized an item in the surprise memory test that followed 24h after encoding.

2.2.3 Results

In line with arousal-based models, we found that outcome-related SCRs, but not anticipatory SCRs, were at item level associated with better recognition performance in both experiments. Critically, our results further showed that memory formation was improved in both experiments for items associated with unsigned PEs compared with items for which correct predictions were made (Figure 8). When controlling for arousal in a joint model, we found an additional effect of unsigned PEs at trend-level in Experiment 1 ($p = .067$) and a significant effect in Experiment 2. Model comparisons for both experiments further confirmed that a full model comprising unsigned PEs, anticipatory and outcome-related arousal explained memory formation significantly better than any model featuring only one of these predictors. In Experiment 2, we further found that the memory-promoting effects of PEs were detectable even when only items from the CS⁺ category were analyzed. This finding implies that the PE-driven memory enhancement was not a mere artifact of their confounding with conditioning categories.

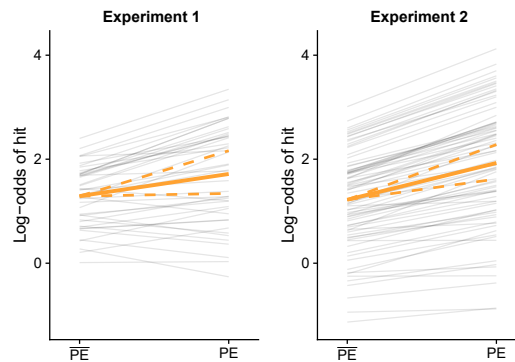


Figure 8. Effects of unsigned PEs on recognition memory in Experiments 1 and 2 of Study 2. Compared with trials in which participants correctly predicted the outcome (\overline{PE}), items from trials featuring incorrect predictions (PE) had a higher probability of being subsequently recognized. Grey lines show estimates from individual participants. Thick orange lines show the fixed effect across all participants, while dashed orange lines show its 95% CI. Adapted from Kalbe and Schwabe (2020a).

2.2.4 Conclusions

The present study provides support for a cognitive model of memory modulation that operates at the exemplar level by enhancing memory for stimuli associated with aversive PEs. Our data were also in line with the more traditional model of arousal-based memory modulation which was supported by consistently improved memory for stimuli followed by higher outcome-related arousal. Together, these findings indicate that both physiological and cognitive models can successfully explain memory formation for individual items. Our results further suggest that both mechanisms might modulate memory formation in an additive fashion.

2.3 Study 3: What neural mechanism drives the memory-modulating effects of aversive PEs?

Kalbe, F., & Schwabe, L. (2021). Prediction errors for aversive events shape long-term memory formation through a distinct neural mechanism. *bioRxiv*. <https://doi.org/10.1101/2021.03.19.436177> – (Appendix C)

2.3.1 Background

Study 2 provided behavioral evidence for the long-term memory-promoting effects of aversive PEs. The neural basis of this effect, on the other hand, is still largely unknown. Based on the previous literature, we considered two contrasting models to explain the superior memory for aversive PE-related stimuli. One model assumes that aversive PEs enhance standard processes of long-term memory formation revolving around the MTL (particularly the hippocampus and parahippocampal gyrus) and the mPFC (Davachi and Wagner, 2002; Eichenbaum, 2004; Preston and Eichenbaum, 2013). Alternatively, large aversive PEs might induce a qualitative shift in mnemonic processing that is characterized by the creation of a separate memory trace for PE-associated stimuli rather than their integration into existing schemata (Rouhani et al., 2020; van Kesteren et al., 2012). In this study, we aimed to test these alternative accounts by recording neural activity using fMRI while participants encoded unique pictures associated with aversive PEs in a further optimized behavioral task.

2.3.2 Methods

Fifty healthy participants completed an incidental memory task in an MRI scanner while we additionally recorded skin conductance as a measure of arousal. This task featured a series of 120 unique pictures, of which some were followed by mild electric shocks that were partially predictable based on picture categories. In each trial, participants estimated the probability that a shock would follow, which allowed us to calculate the explicit continuous PE. To reach a more even distribution of positive and negative PEs, we introduced a third picture category, which was followed by a shock in only one-third of all trials. In a surprise recognition test 24h after encoding, participants saw all 120 pictures from the previous day intermixed with an equal number of previously unseen pictures and categorized them as old or new. To gain insights into neural mechanisms behind the memory modulating effects of PEs, we used a combination of univariate fMRI analyses, large-scale network analyses, and multivoxel pattern analyses (MVPA).

2.3.3 Results

Results confirmed previous findings that PEs modulate memory formation at the item level, but specific effects depended on their sign. Whereas greater negative PEs (associated with unexpected shock omissions) were associated with improved recognition performance, greater positive PEs (associated with unexpected shocks) seemed to impair recognition performance (Figure 9). Again, a model including both arousal measures and PEs explained subsequent memory performance best. At the neural level, negative PEs activated the bilateral anterior insula and the dACC, which are key areas of the salience network (Ham et al., 2013; Menon and Uddin, 2010). Structures of the medial-temporal encoding network (including the hippocampus and parahippocampal gyrus), as well as the mPFC, precuneus, and angular gyrus, which form a schema network (van Kesteren et al., 2012; Vogel et al., 2018), showed decreased activation in response to negative PEs. In line with the established role of the MTL in episodic memory formation, greater activation of the (para-)hippocampus during stimulus presentation was linked with improved subsequent memory. However, improved memory in response to greater negative PEs was linked with even decreased activation of these regions (Figure 10A). An additional follow-up analysis of large-scale brain networks showed that between-network connectivities of the salience and schema network and of the salience and medial temporal encoding network were increased for larger negative PEs, the former of which was linked to better recognition performance (Figure 10B).

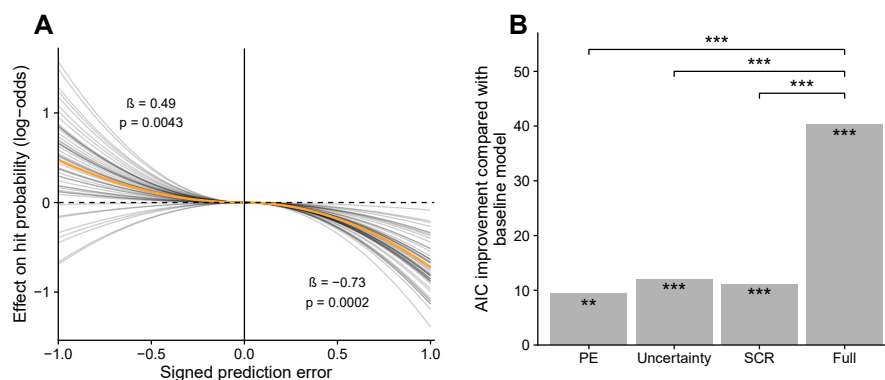


Figure 9. Main behavioral findings of Study 3. (A) Signed effects of quadratic PEs on recognition memory. Whereas negative PEs (reflecting the surprising omission of shocks) were associated with improved recognition performance, positive PEs (reflecting the surprising delivery of shocks) were associated with decreased memory performance. Individual grey lines show estimates for individual participants, while the orange line reflects the estimated fixed effect across participants. (B) Model comparisons using likelihood-ratio tests revealed that a model combining cognitive measures (i.e., uncertainty and PEs) and arousal measures (i.e., anticipatory and outcome-related SCRs) explained memory formation significantly better than any model including only a single of these predictors (markings between bars). A substantial improvement in the Akaike information criterion (AIC) for the full model also confirmed this finding. Furthermore, any single predictor significantly improved the model fit compared with a baseline model estimating only a random intercept per participant (markings within each bar). ** $p < .01$, *** $p < .001$. Adapted from Kalbe and Schwabe (2021).

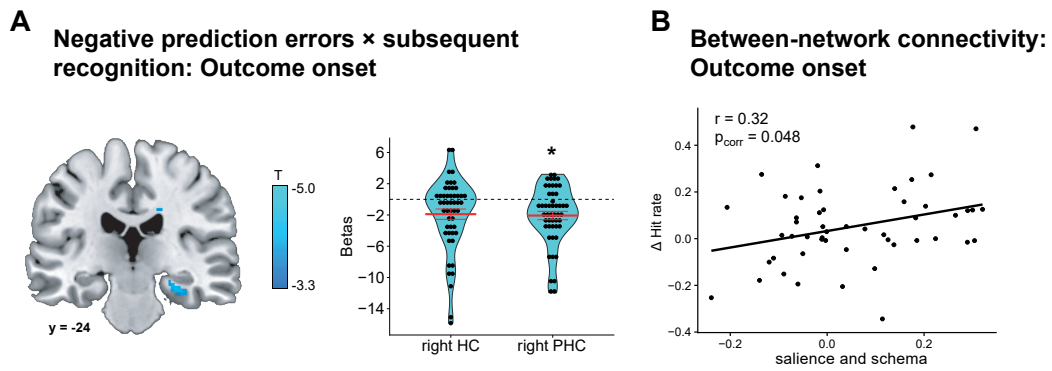


Figure 10. Neural findings of Study 3 (Kalbe and Schwabe, 2021) that linked effects of PEs with subsequent recognition memory. (A) Items associated with larger negative PEs that were subsequently recognized were associated with decreased BOLD responses in the right posterior parahippocampal gyrus (PHC), and at trend level, in the right hippocampus (HC). Black dots show data from individual participants. Thick red lines show mean betas, while the thin red lines mark ± 1 standard error of the mean. (B) Larger increases in between-network connectivity for the salience and schema network to negative PEs were associated with greater memory enhancement. p_{corr} indicates that p -values were Bonferroni corrected for the number of sequential tests. * $p_{\text{SVC}} < .05$ (small volume corrected (SVC), family-wise-error- and Bonferroni-corrected for the number of regions). Adapted from Kalbe and Schwabe (2021).

2.3.4 Conclusions

Together, these results indicate that memory-promoting effects of negative PEs are driven by a neural mechanism that is distinct from standard modes of memory formation. This mechanism might be orchestrated through the salience network, which has been shown to modulate other large-scale brain networks including the medial temporal encoding network (Menon and Uddin, 2010; Sridharan et al., 2008). While we also replicated previous findings showing that arousal moderated memory formation at the item level, these effects were separable from those of PEs at both the behavioral and neural levels. Therefore, these results further point towards a distinct cognitive mechanism that relies on aversive PEs to shape long-term declarative memory formation in threatening environments.

General Discussion

Negative experiences are sometimes perceived as inevitable, but they do not always come unforeseen. Theories of adaptive memory posit that a major phylogenetic driver in the development of long-term declarative memory was its potential to detect and subsequently avoid threatening situations (Klein et al., 2002; Nairne and Pandeirada, 2008a, 2008b). Consequently, if initially neutral stimuli relate to important outcomes, then not only these salient outcomes themselves should be kept in long-term memory but also stimuli that predict their occurrence. To achieve this goal, the literature has so far mostly focused on an unspecific modulation of memory, where no semantic link between a promoted stimulus and the memory-promoting event is necessary. Instead, these effects non-selectively promote memory for stimuli within a limited time frame around the salient experience. More recently, it has been shown that this memory modulation can also be specific to stimuli from a category that gains motivational significance, while leaving memory for stimuli from another category unaffected (Dunsmoor et al., 2015; Oyarzún et al., 2016; Patil et al., 2017). Such category-specific effects were identified for stimuli encoded within the same temporal context as the salient experiences (i.e., online), for stimuli encoded after salient experiences (i.e., prospectively), and, most remarkably, even for stimuli encoded before salient experiences (i.e., retroactively). Based on the very limited evidence for this phenomenon so far, in Study 1, we aimed to closely replicate the reported category-specific memory enhancement through aversive shocks (Dunsmoor et al., 2015). Overall, our results cast doubt on the reliability and generalizability of the proposed category-specific retroactive effect but provided substantial support for both a prospective and an online promotion of memory that was specific to items from the category predictive of electric shocks during fear conditioning. In Study 2, we proposed an alternative exemplar-level interpretation of this effect. Specifically, we hypothesized that trial-unique learning signals based on two separate models would explain differential memory formation. In line with the first, more traditional approach that focuses on the memory-promoting effects of physiological arousal, we found that trial-level outcome-related SCRs were linked with improved memory. In addition to this arousal-based account, we built upon recent theoretical advances (Trapp et al., 2018) and introduced a novel cognitive model that predicts a memory promotion for stimuli associated with surprising outcomes (i.e., aversive PEs). Results not only confirmed that aversive PEs were associated with improved memory, but model comparisons further suggested that these effects were separable from those based on physiological arousal. Therefore, these results pointed to a novel cognitive account of memory formation that is separable from long-established arousal-based effects. Study 3 was designed to shed light

on the neural underpinnings of this novel mechanism. Besides partially replicating effects behaviorally, the distinction of arousal- versus PE-based memory modulation was also reflected neurally. Our results suggest that (negative) PEs promote memory through a specific neural mechanism that is associated with increased activity in the salience network but decreased activity in networks associated with expected outcomes including both the schema network as well as the MTL. In contrast to standard modes of memory formation, these decreases in the activation of medial-temporal structures were associated with the PE-induced memory improvements, suggesting a distinct neural mechanism.

3.1 Unspecific, category-level, and exemplar-level memory modulation

Our initial plan for this series of experiments was to explore the cognitive and neural basis of the putative category-specific retroactive enhancement of memory, as first introduced by Dunsmoor et al. (2015) and later conceptually replicated in the reward domain by Patil et al. (2017). Their findings introduced a promising new direction in the field of memory modulation, which had previously mostly reported unspecific effects. In a typical report of such unspecific memory modulation, a group of participants experiencing a salient or stressful event showed, relative to a control group without such an experience, improved memory for stimuli encoded in temporal proximity (Moncada et al., 2015; Roozendaal, 2002; Schwabe et al., 2012). Category-specific memory modulation, on the other hand, implies that only specific stimuli from a category linked with salient outcomes receive this memory promotion. Reportedly, these effects could be observed in all three temporal directions, based on the timing of the promoted stimuli relative to the memory-promoting event (Dunsmoor et al., 2015). In an online enhancement of memory, promoted stimuli and memory-promoting events occurred in the same temporal context. In a prospective enhancement of memory, participants first experienced that one category of stimuli predicted salient events and then showed improved memory for stimuli from the same category encountered at a later stage, even though these stimuli were themselves never directly paired with the salient events. Finally, in the retroactive enhancement of memory, initially neutral stimuli were first encoded, and only later was their respective category linked with salient events, which still prompted a specific retroactive promotion of these stimuli. This specific retroactive effect is remarkable, as shock contingencies linking picture categories with aversive shocks could only be inferred once the subsequent conditioning phase began (Dunsmoor et al., 2015). Therefore, differential attention between stimulus categories during encoding (Baddeley et al., 1984; Chun and Turk-Browne, 2007; Craik et al., 1996) could not explain this effect. Instead, Dunsmoor et al. (2015) argued that these findings would be in line with consolidation effects according to the synaptic tag and capture hypothesis (Frey and Morris, 1997, 1998; Martin and Kosik,

2002; Rogerson et al., 2014). However, the exact neurophysiological mechanisms by which synaptic tagging would be limited to a specific category of stimuli is unclear. Rather, previous applications of synaptic tagging to the behavioral level in both rodents and humans had only reported an unspecific retroactive memory modulation (Ballarini et al., 2013; Ballarini et al., 2009; Moncada et al., 2015).

In Study 1, across four experiments aimed to closely replicate the category-specific retroactive memory enhancement through aversive fear conditioning (Dunsmoor et al., 2015), we gathered data from 285 unique participants but found only very limited evidence for the phenomenon. Specifically, none of the four experiments provided evidence for category-specific retroactive memory enhancement when we strictly replicated the analysis strategy from the original study that collapsed recognition memory across confidence (Dunsmoor et al., 2015). Exploratory high confidence memory analyses showed significant category-specific retroactive memory enhancement in one of our four experiments, with one additional experiment showing the effect at trend level. However, it should be noted that we refrained from correcting for the number of exploratory statistical tests and that none of the reported effects would have survived such a correction. Even though a pooled analysis across all four experiments should be able to detect even tiny effects, only the model focusing on high confidence memory showed a significant category-specific retroactive memory effect. Interestingly, parallel Bayesian tests on pooled high confidence recognition data provided even substantial evidence for the null hypothesis rejecting any category-specific retroactive memory modulation. Together, these findings challenge the reliability of the category-specific retroactive memory effect.

Our results also cast doubt on the validity of claims that the reported category-specific retroactive memory effect reflects true memory modulation. Even the limited evidence in favor of this effect was restricted to corrected recognition as the measure of recognition memory, which is based on the two-high threshold model (Bröder et al., 2013; Snodgrass and Corwin, 1988). This model can be differentiated from signal detection theory (Macmillan and Creelman, 2004; Wickens, 2002) and its measure of recognition memory (d') based on the assumed shape of the receiver operating characteristic (ROC), which describes the relationship of hit rates and false alarm rates across variations in the response criterion (Yonelinas and Parks, 2007). Critically, signal detection theory, which assumes that ROCs are curvilinear, has much more empirical support than the two-high threshold model, which assumes that ROCs are linear (Dube and Rotello, 2012; Pazzaglia et al., 2013; Wixted, 2007). Choosing a recognition measure based on the correct model is critical as only when its assumptions are met, resulting estimates of recognition memory are independent of response criteria (Snodgrass and Corwin, 1988). Therefore, if evidence for category-specific retroactive memory enhancement only emerges in corrected recognition based on the problematic two-high threshold model, one might hypothesize that these effects reflect changes in the response criterion rather than actual memory modulation. Indeed, at least in one of the four

experiments, we found evidence of a more liberal response criterion for pre-conditioning items from the CS⁺ category over those from the CS⁻ category.

In this context, it is also important to consider the so far only successful conceptual replication of Dunsmoor et al. (2015), which reported category-specific retroactive memory enhancement for stimuli from a category that was later linked with higher reward opportunities (Patil et al., 2017). Originally, in Study 1, we also reported results from an attempt to closely replicate these findings from the reward domain, although we later dropped this experiment from Study 1 to focus entirely on the replication of Dunsmoor et al. (2015). Notably, our attempt to closely replicate Patil et al. (2017) not only failed to provide evidence for the specific retroactive memory enhancement through rewards but could not even replicate the online effect for stimuli directly linked with larger rewards (Kalbe and Schwabe, 2020b). As these data might also show that our experimental manipulation was overall unsuccessful, some caution is warranted when interpreting our results as a failure to replicate the specific retroactive memory effect in the reward context.

Overall, our pattern of results is more in line with a previous conceptual replication of Dunsmoor et al. (2015) that investigated category-specific retroactive memory enhancement through subsequent rewards that participants encountered in a Pavlovian conditioning paradigm (Oyarzún et al., 2016). Parallel to our results, these authors found improved memory for stimuli from a rewarded category over unrewarded stimuli that were encoded during an intermediate reward phase (i.e., an online memory promotion) and during a post-reward phase, but no evidence for any specific retroactive memory enhancement. Taken together, these consistent findings of an online and prospective memory enhancement across multiple studies, recognition measures, and research groups confirm that memory can be modulated selectively based on category membership, albeit not retroactively.

For purely category-based effects, all stimuli from a given category should experience, on average, similar levels of memory promotion. Contrarily, if there are systematic differences in memory within the same category that cannot be attributed merely to the inherent memorability of the stimulus, this would indicate that an additional mechanism at least supplements category-level effects. We hypothesized in Study 2 that memory formation during the fear conditioning phase would be modulated by both cognitive- and arousal-based measures at the level of individual stimuli (i.e., exemplar-based). Indeed, results showed that stimuli associated with binary aversive PEs and greater outcome-related arousal were more likely to be recognized in the memory test 24h after encoding. In isolation, these results would not necessarily prove an exemplar-level memory modulation. As items from the CS⁺ category were, relative to items from the CS⁻ category, associated with both an increased number of PEs and greater outcome-related SCRs, these item-level predictors are partially confounded with conditioning categories. Can this confounding explain our findings of exemplar-level effects? In Study 2, we found that CS⁺ items associated with PEs were recognized significantly better than CS⁺ items associated with correct outcome predictions. In other words, we found

a PE-based memory modulation within the same stimulus category, which provides compelling evidence that our exemplar-level effects were not purely artifacts of confounding with conditioning categories. Particularly for the online modulation of memory, exemplar-level effects could underlie apparent category-level effects. At first glance, findings of a prospective memory enhancement seem to stem from a category- rather than exemplar-based memory modulation, as participants were no longer required to make any predictions of outcomes, nor was there the threat of receiving shocks in this encoding phase. At the same time, it cannot be excluded that the beginning of the post-conditioning phase still served as an extinction training of previously acquired fear, which would lead to an increased exemplar-level arousal particularly for early CS⁺ trials (Dunsmoor et al., 2018; Hermans et al., 2006).

3.2 Separating physiological from cognitive accounts of memory modulation

Beyond identifying different levels at which the modulation of memory operates, a central goal of this series of experiments was to gain insight into the underlying mechanisms that select stimuli for preferential storage in long-term memory. Here, we contrasted two models based on a physiological versus a cognitive approach. In principle, memory phenomena of all three levels of specificity (i.e., unspecific, category-specific, and exemplar-based) are compatible with either approach. A typical finding of memory modulation in the stress literature is that a stressful experience retroactively and unspecifically enhances memory for stimuli previously encoded under neutral valence (Nater et al., 2007; Schwabe et al., 2008; Schwabe et al., 2012). The stress literature has often favored a physiological account of these stress effects by focusing on the well-established effects of adrenal and glucocorticoid stress hormones in the basolateral amygdala, which modulates activity in memory-relevant regions such as the MTL and thereby aids memory consolidation (McGaugh, 2000; McIntyre et al., 2003; Roozendaal et al., 2006; Schwabe et al., 2012). Conceptually similar effects of unspecific retroactive memory enhancement have been described in the behavioral tagging framework. Here, memory for neutral stimuli is unspecifically and retroactive enhanced through subsequent salient experiences, which are often characterized by their novelty (Balarini et al., 2009; Chen et al., 2020; Moncada et al., 2015; Moncada and Viola, 2007). Whereas behavioral tagging has its roots in the physiological model of synaptic tagging, the concept of novelty has a strong cognitive notion and is conceptually similar to PEs (Clark, 2018; Kiverstein et al., 2019; Wessel et al., 2012). Inspired by a theoretical account that links stressful events with expectancy violations (Trapp et al., 2018), we recently showed that the memory boost for stimuli surrounding stressful events can be reduced by providing participants with detailed information about the upcoming stressor and therefore reduce subsequent surprise associated with the stressful situation (Kalbe et al., 2020). While one

might hypothesize that our cognitive manipulation also reduced the physiological response to the stressor, both subjective and psychophysiological data indicated similar stress responses regardless of prior information. Overall, these results support a cognitive model that emphasizes the role of expectancy violations in the unspecific stress-driven modulation of memory.

For the more specific category-based memory modulation, investigations of the underlying mechanisms are still rare. Dunsmoor et al. (2015) proposed that their findings of category-specific memory enhancement in the aversive context were mediated by arousal, but provided no explicit evidence for this hypothesis. A congruent finding with this arousal-based account were the increased anticipatory SCRs to items from the CS⁺ category over items from the CS⁻ category, which reflected successful fear conditioning. On the other hand, if category-specific effects critically depend on the extent of acquired fear, then one might expect a significant correlation between the induced arousal during fear conditioning and the overall memory performance, which we found no evidence for in any of our four close replications.

Another study with a very similar design recorded fMRI during encoding and linked the specific retroactive memory enhancement to increased activity of the ventral tegmental area and substantia nigra during the conditioning phase (Clewett et al., 2020). The authors interpret this finding as an arousal-related neuromodulatory effect on the dopaminergic system, which supposedly drives the category-specific retroactive effect. However, their failure to demonstrate any significant category-specific retroactive memory enhancement at the behavioral level means that these results must be interpreted with caution. Furthermore, these findings fit similarly within a cognitive perspective of category-specific memory enhancement, as neurons of the ventral tegmental area and substantia nigra are known for their coding of PEs (Lak et al., 2014; Schultz et al., 1997). At the behavioral level, outcomes of CS⁺ trials were further much less predictable as they were probabilistically linked with shocks. Meanwhile, outcomes of CS⁻ trials were, at least after participants learned that they were never followed by a shock, perfectly predictable. However, our data from Study 3 suggest that this overall difference in the entropy of outcomes alone cannot explain the superior memory for items from the CS⁺ category. Specifically, we found in Study 3 that items from the CS^{a+} category, associated with a two-thirds shock probability, were recognized significantly better than items from the CS^{b+} category, associated with a one-third shock probability. Since shock probabilities of one-third and two-thirds lead to an equal Shannon entropy, this factor alone cannot explain these memory differences. Moreover, we found no evidence that the greater number of shocks in the CS^{a+} category could explain these differences in memory. Rather, our analyses at the exemplar level suggested that different types of PEs predominantly associated with these conditioning categories contributed to these findings.

Identifying and characterizing such exemplar-level memory effects were our primary goals in Studies 2 and 3. In line with established models focusing on physiological arousal (McGaugh, 2000; McIntyre et al., 2003; Roozendaal et al., 2006), we found in both studies that phasic SCRs predicted subsequent memory formation. However, only those SCRs elicited

by the outcome of the trial (i.e., whether a shock occurred) were reliably and positively associated with subsequent memory. For SCRs during stimulus encoding and in anticipation of shocks, we found either no association with subsequent memory or even a negative association with subsequent memory in Experiment 2 of Study 2. Our key focus in Studies 2 and 3 was the exemplar-level memory modulation based on the cognitive measures of aversive PEs. Across slightly different conceptualizations of aversive PEs in Studies 2 and 3, we found memory modulating effects even after accounting for several control variables. Interestingly, effects of outcome-related SCRs and PEs share a similar temporal profile. Both effects emerge only after the end of the stimulus presentation, meaning that their effects cannot be easily explained by enhanced encoding due to increased attention to the stimulus (Baddeley et al., 1984; Chun and Turk-Browne, 2007; Craik et al., 1996). Both effects are also retroactive, albeit on a much narrower time scale compared with the concept of behavioral tagging (Dunsmoor et al., 2015; Moncada et al., 2015; Patil et al., 2017). This similarity in the temporal dynamics of arousal- and PE-based effects might indicate that both are different manifestations of the same underlying mechanism. Consistent with this hypothesis, outcome-related SCRs have previously been shown to reflect PEs in aversive environments, particularly the unexpected omission of shocks (de Berker et al., 2016; Spoomaker et al., 2012; but see Bach and Friston, 2012). Contrarily, our data from Studies 2 and 3 consistently showed that arousal and aversive PE-based effects were statistically separable. The best fitting model in both studies was always based on a combination of arousal with aversive PEs, suggesting that both measures contributed uniquely to memory formation. However, while skin conductance is one of the most established measures of arousal in fear learning experiments (Bach and Melinscak, 2020; Lonsdorf et al., 2017), it does not necessarily capture all aspects of arousal (Neiss, 1988) and also tracks other constructs such as cognitive demand (Botvinick and Rosen, 2009). Still, these results suggest that aversive PEs modulate memory through a mechanism that goes beyond the prominent effects of physiological arousal. In this case, one would also expect the neural mechanism behind the PE-driven memory enhancement to deviate from the amygdala-associated pathways that are assumed to underlie arousal-based effects on memory (McGaugh, 2000; McIntyre et al., 2003; Roozendaal et al., 2006).

3.3 A distinct mechanism underlying effects of aversive PEs on memory

We specifically designed Study 3 to investigate neural mechanisms of the aversive PE-driven memory modulation by recording fMRI during encoding. Overall, results provided little support for the hypothesis that a simple promotion of standard memory processes revolving around the MTL is behind this effect (Eichenbaum, 2004; Fernández et al., 1999; Squire and Wixted, 2011), as theorized by neurobiological models focusing on physiological arousal

(McGaugh, 2000; Roozendaal et al., 2006). During stimulus presentation, increased activity in the MTL was indeed associated with greater subsequent memory performance, replicating earlier findings (Davachi and Wagner, 2002; Eichenbaum, 2004; Fernández et al., 1999). Critically, this relationship between MTL activity and subsequent memory seemed to reverse with large negative PEs, which only emerged after the stimulus presentation. Here, we found even decreased activation in response to larger negative PEs in both the MTL and the mPFC. Greater decreases in MTL activation in response to large negative PEs were further linked with an improved subsequent memory. At first glance, this finding appears to be not only in stark contrast to the vast literature linking increased hippocampal activity to improved episodic memory formation (Davachi and Wagner, 2002; Eichenbaum, 2004; Fernández et al., 1999), but also seems to contradict our earlier findings that showed a positive link between hippocampal activity and subsequent memory during stimulus presentation. However, it should be noted that the negative link between (para-)hippocampal activity and subsequent memory in response to large negative PEs only emerged as the outcome of a trial was revealed and therefore when the stimulus had already disappeared. Still, these findings suggest that a qualitatively distinct neural mechanism drives the memory-modulating effects of aversive PEs.

This mechanism was characterized by a markedly increased activity in the anterior insula and the dACC in response to negative PEs. Both areas are assumed to be involved in error monitoring (Bastin et al., 2016; Botvinick et al., 2004; Preuschoff et al., 2008), but they also form key regions of the salience network, which responds to behaviorally salient events, such as outcomes that conflict with prior expectations (Ham et al., 2013; Menon and Uddin, 2010). The salience network has further been identified as a modulator of other large-scale brain networks (Menon and Uddin, 2010; Sridharan et al., 2008). In a subsequent analysis of large-scale brain networks, we found increased between-network connectivity in response to large negative PEs between the salience network and the medial temporal encoding network (at trend level), as well as between the salience network and the schema network. The schema network, which encompasses the angular gyrus, precuneus, and mPFC, integrates new information into existing memory structures (van Kesteren et al., 2012; Vogel et al., 2018). When predictive stimuli are not followed by their expected outcome, this signals that the particular exemplar stands out and cannot be easily integrated into an existing memory schema, leading instead to a separate memory trace (Rouhani et al., 2020; van Kesteren et al., 2012). In line with this notion, we found that greater between-network connectivity between the salience network and the schema network to large PEs was linked with improved subsequent memory performance.

One might speculate whether this distinction between the cognitive PE-based approach and the physiological arousal-based approach to memory modulation can also be translated to the neuroendocrinological level. Reward PEs are strongly linked with the action of dopamine in the mesolimbic system, particularly in the ventral tegmental area, substantia nigra, and ventral striatum (Lak et al., 2014; O'Doherty et al., 2003; Pagnoni et al., 2002; Pessiglione

et al., 2006; Schultz et al., 1997). Although the neuroendocrinological substrate of aversive PEs is overall more controversial (Fiorillo, 2013; Schultz, 2019), a critical involvement of the mesolimbic dopamine system has still been suggested (Brooks and Berns, 2013; Matsumoto and Hikosaka, 2009). On the other hand, the memory-promoting effects of physiological arousal have been linked to the rapid secretion of noradrenaline in response to arousing stimuli (Hauser et al., 2019; Joëls et al., 2011; Schwabe et al., 2012). Future studies could directly test this account by using pharmacological manipulations that modulate dopaminergic or noradrenergic receptors. In case of a double dissociation, manipulating the dopaminergic system should selectively affect the PE-based memory modulation, while manipulating the noradrenergic system should selectively affect the memory-modulating effects of arousal.

3.4 Future directions

Overall, our findings of a cognitive memory modulation that uses PE-signals to select stimuli with unexpected consequences for storage in long-term memory introduce a new perspective on declarative memory formation in threatening environments. Even though this aversive PE-mediated memory modulation resembles recent findings in the reward domain and beyond (Calderon et al., 2021; Clark and Chang, 2021; Rouhani and Niv, 2021; Rouhani et al., 2018; Rouhani et al., 2020), some key aspects need further clarification by future research. One such issue is the exact shape of the relationship between PEs and subsequent memory formation. Both Studies 2 and 3 showed consistent support for the link of negative aversive PEs (associated with unexpected shock omissions) with improved memory formation. For positive PEs (associated with unexpected shock deliveries), results were less consistent. The data from Study 2 favored a memory-promoting effect of positive PEs, although their low rate of occurrence made this finding less reliable. In contrast to these previous results, after modifying the behavioral task to balance the distribution of PEs in Study 3, we found opposite, memory decreasing effects of positive PEs. Based on neural data from Study 3 showing that larger positive PEs were associated with clusters of increased activity in areas linked to sensorimotor functions, we speculated that their memory decreasing effect might be due to distracting influences of the aversive shock on mnemonic processing. Whether these memory-decreasing effects of positive PEs also generalize beyond the specific encoding context inside the MRI scanner remains a critical issue for future research.

Another issue that has been identified by our research group and is currently under investigation concerns the scope of memory modulation observed in this series of experiments. The associative learning literature often distinguishes learning based on temporal contiguity versus contingency (Hammond, 1980; Schultz, 2006). Applied to the present research question, if the observed effects of PEs were contiguity-based, then the mere temporal proximity of the promoted stimulus relative to the surprising outcome would be the critical

component that enables the modulation of memory formation. Contrarily, if this effect depends on contingency, then a surprising outcome should modulate memory only for such stimuli that contributed to the (incorrect) prediction, possibly even when a larger temporal gap separates these two. In the experiments presented in this work, predictive stimuli were always immediately followed by their respective outcome. Therefore, both contiguity and contingency were equally present in each trial. To dissociate influences of both components, our research groups currently runs an experiment that includes additional neutral stimuli after participants completed their prediction in each trial, but before the respective outcome is revealed. If temporal contiguity were sufficient, then these neutral stimuli should also be modulated by subsequent surprising outcomes, even when they are completely non-predictive. If the effect critically depends on contingency, then only memory for the stimulus that gave rise to the incorrect prediction should be modulated, but not memory for the interleaved non-predictive stimulus.

The exact nature of memory that can be modulated by aversive PEs is a more general issue. In all three of our studies, we assessed memory modulation through recognition performance. Although the correct recognition of an item implies at least familiarity, we cannot infer that an individual could also recollect the broader context a stimulus was encountered in. According to dual-process theories, recognition memory is supported by both familiarity- and recollection-based processes (Diana et al., 2006; Rugg and Yonelinas, 2003; Yonelinas, 1994). However, if our observed memory modulation does indeed serve behavioral adaptation by identifying and avoiding aversive outcomes, then the mere familiarity with a previously encountered predictive stimulus is insufficient. Instead, to form adaptive decisions based on such memories, an individual must also be able to recall the specific outcome that was associated with the predictive stimulus (Gershman and Daw, 2017), as enabled only by recollection-based memory. Future studies should address this issue by not only probing the recognition of predictive stimuli but also of their respective outcomes.

Finally, our findings might improve interventions in fear- and trauma-related disorders, which are often characterized by aberrant episodic memory functioning (Airaksinen et al., 2005; Berntsen et al., 2003; Sartory et al., 2013). Regarding empirically well-supported treatments based on exposure therapy (Böhlelein et al., 2020; Foa et al., 1991; Rauch et al., 2012), our consistent findings of memory-promoting effects of negative aversive PEs from Studies 2 and 3 suggest that therapists should activate patients' fear-related expectations prior to the confrontation with the feared stimulus. The subsequent observation that feared consequences stay out constitutes a large negative PE and should therefore lead to stable fear-incongruent declarative memory. Whether these considerations are indeed beneficial for the therapeutic practice would have to be clarified by future studies.

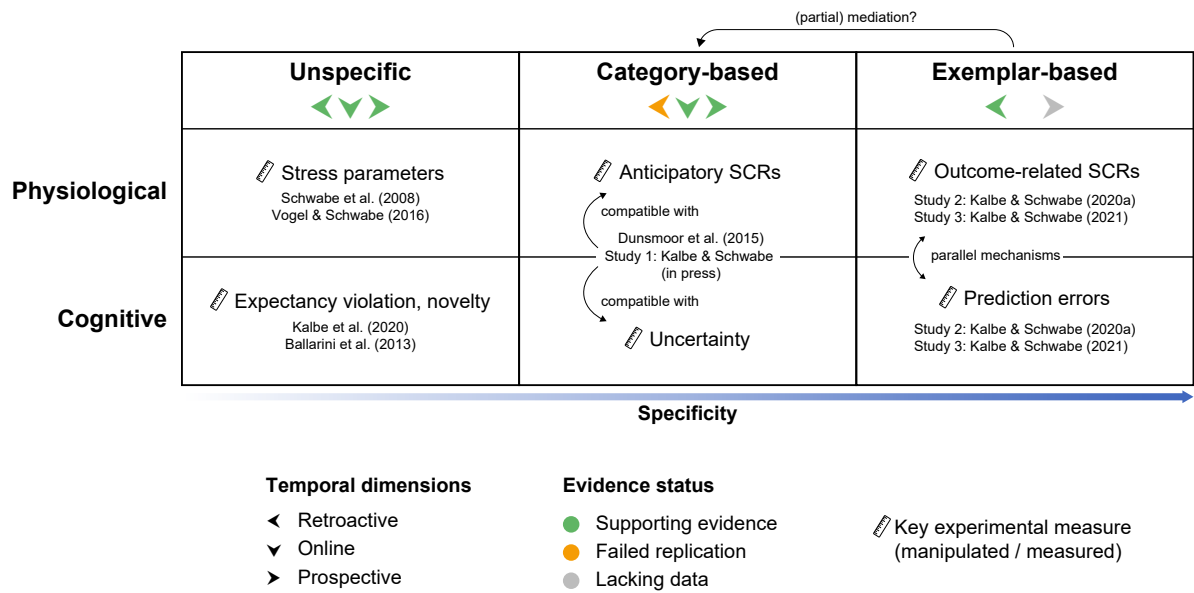


Figure 11. Integration of our results with the proposed framework of declarative memory modulation. A common feature of all the effects captured by this framework is that memory for a stimulus is modulated by a separate event. Three displayed dimensions of this memory modulation include its specificity (in ascending order: unspecific, category-based, or exemplar-based), the assumed underlying mechanisms (physiological or cognitive), and the temporal directionality of effects (retroactive, online, and prospective). In summary, any form of memory modulation could be attributed to both physiological and cognitive variables. In the case of unspecific effects, stressful events have long been known to modulate memory for surrounding stimuli in a retroactive, online, and prospective fashion. Dominant accounts attribute this effect to adrenal stress hormones, but more recent research has begun to additionally explore the role of cognitive factors, including expectancy violation and novelty. The key focus of our series of studies were the more specific category-based and the most specific exemplar-based modulation of memory. In line with previous reports (Dunsmoor et al., 2015; Patil et al., 2017), in Study 1 we found evidence for category-specific memory enhancement both online and prospectively, but not retroactively. These effects were compatible with both a physiological and a cognitive account. Our Studies 2 and 3 focused on a retroactive exemplar-based memory modulation that might partially underlie these apparent category-level effects. Results suggested that physiological and cognitive drivers reflect separable mechanisms that both contribute to the exemplar-based memory modulation. Whether this parallelism is similarly present in the category-specific and the unspecific modulation of memory needs to be clarified by future research. Note that provided references are non-exhaustive examples of experimental studies in support of each specific effect.

3.5 Conclusions

Looking back at the introductory example of the police officer facing potentially threatening situations while performing traffic stops, results from our three studies paint a nuanced picture of an adaptive memory system that operates at multiple levels of specificity and uses different strategies to prioritize relevant over irrelevant information for storage in long-term memory. It has long been known that stressful or dangerous events, such as a stopped driver

suddenly drawing a firearm, produce arousal that not only ensures that impressions of the firearm itself are likely to be remembered, but that this memory enhancement also extends to inconsequential details that were encoded in the same temporal context. However, based on our cognitive PE-based approach, an early warning that the driver might draw a weapon would reduce associated levels of surprise and therefore the unspecific memory enhancement (Kalbe et al., 2020). Could the escalating traffic stop also promote memory only for a specific category of stimuli? For example, if the perpetrator drove a pickup truck, would surrounding memories of encounters with other pickup trucks also receive a memory enhancement? Our data suggest that this might be the case, but only for those trucks that are encountered afterward (i.e., prospectively) not beforehand (i.e., retroactively). At the highest level of specificity, we demonstrated that a modulation of memory can be limited to unique stimuli that relate to unexpected consequences. For example, a car with aggressive bumper stickers depicting guns would raise suspicion that the driver could be armed. If this suspicion turns out to be true, but the situation stays manageable, this experience will not necessarily be remembered over long periods. On the other hand, if the car with the aggressive bumper stickers turns out to belong to a peaceful old lady, this would constitute a large negative PE, leading to a higher chance that the police officer will form stable memory for this schema-incongruent experience. Conversely, if all signs point to a routine traffic stop, but the situation unexpectedly turns violent, will this positive PE lead to stable memory for the possibly missed indicators of the violent outcome? As our findings regarding this question were overall inconsistent, future research will need to further address this question.

In conclusion, an adaptive declarative memory system needs to prioritize self-relevant over irrelevant information. One major challenge is to build stable memory not only for salient events themselves, but also for predictors of such events. This memory modulation was the common focus of all three of our studies. It can occur at multiple levels of specificity, including unspecific, category-specific, and, most centrally, exemplar-based effects (Figure 11). Besides the established contributions of physiological arousal to the superior memory for emotional experiences, we introduce a PE-based cognitive approach that is separable from physiological models both behaviorally and neurally. Our findings demonstrating a memory modulation of stimuli linked with unexpected outcomes reveal new connections with similar effects in reward-based learning, bridge the gap between the traditionally separate fields of declarative and associative learning, and broaden our understanding of how declarative memory enables us to leverage past experiences in threatening environments.

References

- Airaksinen, E., Larsson, M., & Forsell, Y. (2005). Neuropsychological functions in anxiety disorders in population-based samples: Evidence of episodic memory dysfunction. *Journal of Psychiatric Research*, *39*(2), 207–214. <https://doi.org/10.1016/j.jpsychires.2004.06.001>
- Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*(4), 703–719. <https://doi.org/10.1037/0033-295X.96.4.703>
- Anderson, J. R., & Schooler, L. J. (2000). The adaptive nature of memory. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford Handbook of Memory*. (pp. 557–570). Oxford University Press.
- Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, *3*(2), 213–251. https://doi.org/10.1088/0954-898X_3_2_009
- Bach, D. R., & Friston, K. J. (2012). No evidence for a negative prediction error signal in peripheral indicators of sympathetic arousal. *NeuroImage*, *59*(2), 883–884. <https://doi.org/10.1016/j.neuroimage.2011.08.091>
- Bach, D. R., & Melinscak, F. (2020). Psychophysiological modelling and the measurement of fear conditioning. *Behaviour Research and Therapy*, *127*, 103576. <https://doi.org/10.1016/j.brat.2020.103576>
- Baddeley, A., Lewis, V., Eldridge, M., & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General*, *113*(4), 518–540. <https://doi.org/10.1037/0096-3445.113.4.518>
- Baldeweg, T. (2006). Repetition effects to sounds: Evidence for predictive coding in the auditory system. *Trends in Cognitive Sciences*, *10*(3), 93–94. <https://doi.org/10.1016/j.tics.2006.01.010>
- Ballarini, F., Martínez, M. C., Díaz Perez, M., Moncada, D., & Viola, H. (2013). Memory in elementary school children is improved by an unrelated novel experience. *PLoS ONE*, *8*(6), 1–7. <https://doi.org/10.1371/journal.pone.0066875>
- Ballarini, F., Moncada, D., Martinez, M. C., Alen, N., & Viola, H. (2009). Behavioral tagging is a general mechanism of long-term memory formation. *Proceedings of the National Academy of Sciences*, *106*(34), 14599–14604. <https://doi.org/10.1073/pnas.0907078106>
- Barsegyan, A., Mackenzie, S. M., Kurose, B. D., McGaugh, J. L., & Roozendaal, B. (2010). Glucocorticoids in the prefrontal cortex enhance memory consolidation and impair working memory by a common neural mechanism. *Proceedings of the National Academy of Sciences*, *107*(38), 16655–16660. <https://doi.org/10.1073/pnas.1011975107>
- Bastin, J., Deman, P., David, O., Gueguen, M., Benis, D., Minotti, L., Hoffman, D., Combrisson, E., Kujala, J., Perrone-Bertolotti, M., Kahane, P., Lachaux, J.-P., & Jerbi, K. (2016). Direct recordings from human anterior insula reveal its leading role within the error-monitoring network. *Cerebral Cortex*, bhv352. <https://doi.org/10.1093/cercor/bhv352>

- Benedek, M., & Kaernbach, C. (2010a). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*(1), 80–91. <https://doi.org/10.1016/j.jneumeth.2010.04.028>
- Benedek, M., & Kaernbach, C. (2010b). Decomposition of skin conductance data by means of non-negative deconvolution. *Psychophysiology*, *47*(4), 647–658. <https://doi.org/10.1111/j.1469-8986.2009.00972.x>
- Bergt, A., Urai, A. E., Donner, T. H., & Schwabe, L. (2018). Reading memory formation from the eyes. *European Journal of Neuroscience*, *47*(12), 1525–1533. <https://doi.org/10.1111/ejn.13984>
- Berlyne, D. E. (1957). Uncertainty and conflict: A point of contact between information-theory and behavior-theory concepts. *Psychological Review*, *64*(6), 329–339. <https://doi.org/10.1037/h0041135>
- Berntsen, D., Willert, M., & Rubin, D. C. (2003). Splintered memories or vivid landmarks? Qualities and organization of traumatic memories with and without PTSD. *Applied Cognitive Psychology*, *17*(6), 675–693. <https://doi.org/10.1002/acp.894>
- Böhnlein, J., Altegoer, L., Muck, N. K., Roesmann, K., Redlich, R., Dannlowski, U., & Leehr, E. J. (2020). Factors influencing the success of exposure therapy for specific phobia: A systematic review. *Neuroscience & Biobehavioral Reviews*, *108*, 796–820. <https://doi.org/10.1016/j.neubiorev.2019.12.009>
- Bornstein, A. M., Khaw, M. W., Shohamy, D., & Daw, N. D. (2017). Reminders of past choices bias decisions for reward in humans. *Nature Communications*, *8*, 15958. <https://doi.org/10.1038/ncomms15958>
- Bornstein, A. M., & Norman, K. A. (2017). Reinstated episodic context guides sampling-based decisions for reward. *Nature Neuroscience*, *20*(7), 997–1003. <https://doi.org/10.1038/nn.4573>
- Borst, A., & Theunissen, F. (1999). Information theory and neural coding. *Nature Neuroscience*, *2*(11), 947–957. <https://doi.org/10.1038/14731>
- Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*(12), 539–546. <https://doi.org/10.1016/j.tics.2004.10.003>
- Botvinick, M. M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. <https://doi.org/10.1016/j.tics.2019.02.006>
- Botvinick, M. M., & Rosen, Z. B. (2009). Anticipation of cognitive demand during decision-making. *Psychological Research*, *73*(6), 835–842. <https://doi.org/10.1007/s00426-008-0197-8>
- Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329. <https://doi.org/10.1073/pnas.0803390105>
- Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, *21*(8), 916–944. <https://doi.org/10.1080/09658211.2013.767348>

- 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*(5), e10773. <https://doi.org/10.1371/journal.pone.0010773>
- Brooks, A. M., & Berns, G. S. (2013). Aversive stimuli and loss in the mesocorticolimbic dopamine system. *Trends in Cognitive Sciences*, *17*(6), 281–286. <https://doi.org/10.1016/j.tics.2013.04.001>
- Cahill, L., Gorski, L., & Le, K. (2003). Enhanced human memory consolidation with post-learning stress: Interaction with the degree of arousal at encoding. *Learning & Memory*, *10*(4), 270–274. <https://doi.org/10.1101/lm.62403>
- Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, *21*(7), 294–299. [https://doi.org/10.1016/S0166-2236\(97\)01214-9](https://doi.org/10.1016/S0166-2236(97)01214-9)
- Calderon, C. B., De Loof, E., Ergo, K., Snoeck, A., Boehler, C. N., & Verguts, T. (2021). Signed reward prediction errors in the ventral striatum drive episodic memory. *Journal of Neuroscience*, *41*(8), 1716–1726. <https://doi.org/10.1523/JNEUROSCI.1785-20.2020>
- Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, *51*(13), 1484–1525. <https://doi.org/10.1016/j.visres.2011.04.012>
- Chen, N., Tsai, T.-C., & Hsu, K.-S. (2020). Exposure to novelty promotes long-term contextual fear memory formation in juvenile mice: Evidence for a behavioral tagging. *Molecular Neurobiology*, *57*(9), 3956–3968. <https://doi.org/10.1007/s12035-020-02005-1>
- 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*(4), 693–701. <https://doi.org/10.1037/0278-7393.17.4.693>
- Christianson, S.-å., & Loftus, E. F. (1987). Memory for traumatic events. *Applied Cognitive Psychology*, *1*(4), 225–239. <https://doi.org/10.1002/acp.2350010402>
- Chun, M. M., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology*, *17*(2), 177–184. <https://doi.org/10.1016/j.conb.2007.03.005>
- Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. <https://doi.org/10.1017/S0140525X12000477>
- Clark, A. (2018). A nice surprise? Predictive processing and the active pursuit of novelty. *Phenomenology and the Cognitive Sciences*, *17*(3), 521–534. <https://doi.org/10.1007/s11097-017-9525-z>
- Clark, M. D., & Chang, L. J. (2021). Surprise signals changing affective experiences in naturalistic sports spectating. *Neuron*, *109*(2), 199–201. <https://doi.org/10.1016/j.neuron.2020.12.022>
- Clewett, D., Dunsmoor, J., Bachman, S., Phelps, E., & Davachi, L. (2020). Survival of the salient: Emotion rescues otherwise forgettable memories via neural reactivation and post-encoding hippocampal connectivity. *bioRxiv*. <https://doi.org/10.1101/2020.07.07.192252>
- Craik, F. I. M., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*(2), 159–180. <https://doi.org/10.1037/0096-3445.125.2.159>

- Cycowicz, Y. M., & 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, A. D. (2002). Hippocampal contributions to episodic encoding: Insights from relational and item-based learning. *Journal of Neurophysiology*, *88*(2), 982–990. <https://doi.org/10.1152/jn.2002.88.2.982>
- Dawson, M., Schell, A., & Courtney, C. (2011). The skin conductance response, anticipation, and decision-making. *Journal of Neuroscience, Psychology, and Economics*, *4*, 111–116. <https://doi.org/10.1037/a0022619>
- 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*(1), 10996. <https://doi.org/10.1038/ncomms10996>
- de Kloet, E. R., Joëls, M., & Holsboer, F. (2005). Stress and the brain: From adaptation to disease. *Nature Reviews Neuroscience*, *6*(6), 463–475. <https://doi.org/10.1038/nrn1683>
- Diana, R. A., Reder, L. M., Arndt, J., & Park, H. (2006). Models of recognition: A review of arguments in favor of a dual-process account. *Psychonomic Bulletin & Review*, *13*(1), 1–21. <https://doi.org/10.3758/BF03193807>
- Doya, K., Samejima, K., Katagiri, K.-i., & Kawato, M. (2002). Multiple model-based reinforcement learning. *Neural Computation*, *14*(6), 1347–1369. <https://doi.org/10.1162/089976602753712972>
- Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(1), 130–151. <https://doi.org/10.1037/a0024957>
- Dudai, Y. (1997). How big is human memory, or on being just useful enough. *Learning & Memory*, *3*(5), 341–365. <https://doi.org/10.1101/lm.3.5.341>
- Dunsmoor, J. E., Kroes, M. C. W., Moscatelli, C. M., Evans, M. D., Davachi, L., & Phelps, E. A. (2018). Event segmentation protects emotional memories from competing experiences encoded close in time. *Nature Human Behaviour*, *2*(4), 291–299. <https://doi.org/10.1038/s41562-018-0317-4>
- Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, *520*(7547), 345–348. <https://doi.org/10.1038/nature14106>
- Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron*, *44*(1), 109–120. <https://doi.org/10.1016/j.neuron.2004.08.028>
- Ergo, K., De Loof, E., & Verguts, T. (2020). Reward prediction error and declarative memory. *Trends in Cognitive Sciences*, *24*(5), 388–397. <https://doi.org/10.1016/j.tics.2020.02.009>
- Feld, G. B., Weis, P. P., & Born, J. (2016). The limited capacity of sleep-dependent memory consolidation. *Frontiers in Psychology*, *7*, 1368. <https://doi.org/10.3389/fpsyg.2016.01368>

- Fernández, G., Effern, A., Grunwald, T., Pezer, N., Lehnertz, K., Dümpelmann, M., Roost, D. V., & Elger, C. E. (1999). Real-time tracking of memory formation in the human rhinal cortex and hippocampus. *Science*, *285*(5433), 1582–1585. <https://doi.org/10.1126/science.285.5433.1582>
- Fiorillo, C. D. (2013). Two dimensions of value: Dopamine neurons represent reward but not aversiveness. *Science*, *341*(6145), 546–549. <https://doi.org/10.1126/science.1238699>
- Foa, E. B., Rothbaum, B. O., Riggs, D. S., & Murdock, T. B. (1991). Treatment of posttraumatic stress disorder in rape victims: A comparison between cognitive-behavioral procedures and counseling. *Journal of Consulting and Clinical Psychology*, *59*(5), 715–723. <https://doi.org/10.1037/0022-006X.59.5.715>
- Frey, U., & Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, *385*(6616), 533–536. <https://doi.org/10.1038/385533a0>
- Frey, U., & Morris, R. G. (1998). Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiation. *Trends in Neurosciences*, *21*(5), 181–188. [https://doi.org/10.1016/S0166-2236\(97\)01189-2](https://doi.org/10.1016/S0166-2236(97)01189-2)
- Friston, K. J. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. <https://doi.org/10.1038/nrn2787>
- Friston, K. J. (2018). Does predictive coding have a future? *Nature Neuroscience*, *21*(8), 1019–1021. <https://doi.org/10.1038/s41593-018-0200-7>
- Frith, C., & Dolan, R. (1996). The role of the prefrontal cortex in higher cognitive functions. *Cognitive Brain Research*, *5*(1), 175–181. [https://doi.org/10.1016/S0926-6410\(96\)00054-7](https://doi.org/10.1016/S0926-6410(96)00054-7)
- Fukuda, K., & Vogel, E. K. (2019). Visual short-term memory capacity predicts the bandwidth of visual long-term memory encoding. *Memory & Cognition*, *47*(8), 1481–1497. <https://doi.org/10.3758/s13421-019-00954-0>
- Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*(1), 101–128. <https://doi.org/10.1146/annurev-psych-122414-033625>
- Gläscher, J., Daw, N., Dayan, P., & O’Doherty, J. P. (2010). States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*, *66*(4), 585–595. <https://doi.org/10.1016/j.neuron.2010.04.016>
- 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. <https://doi.org/10.1016/j.jml.2016.11.001>
- Ham, T., Leff, A., de Boissezon, X., Joffe, A., & Sharp, D. J. (2013). Cognitive control and the salience network: An investigation of error processing and effective connectivity. *Journal of Neuroscience*, *33*(16), 7091–7098. <https://doi.org/10.1523/JNEUROSCI.4692-12.2013>
- Hammond, L. J. (1980). The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the Experimental Analysis of Behavior*, *34*(3), 297–304. <https://doi.org/10.1901/jeab.1980.34-297>

- Hauser, T. U., Eldar, E., Purg, N., Moutoussis, M., & Dolan, R. J. (2019). Distinct roles of dopamine and noradrenaline in incidental memory. *Journal of Neuroscience*, *39*(39), 7715–7721. <https://doi.org/10.1523/JNEUROSCI.0401-19.2019>
- Heilbron, M., & Chait, M. (2018). Great expectations: Is there evidence for predictive coding in auditory cortex? *Neuroscience*, *389*, 54–73. <https://doi.org/10.1016/j.neuroscience.2017.07.061>
- Hermans, D., Craske, M. G., Mineka, S., & Lovibond, P. F. (2006). Extinction in human fear conditioning. *Biological Psychiatry*, *60*(4), 361–368. <https://doi.org/10.1016/j.biopsych.2005.10.006>
- Hosoya, T., Baccus, S. A., & Meister, M. (2005). Dynamic predictive coding by the retina. *Nature*, *436*(7047), 71–77. <https://doi.org/10.1038/nature03689>
- Humphreys, G. W., Duncan, J., Treisman, A., & Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *353*(1373), 1245–1255. <https://doi.org/10.1098/rstb.1998.0280>
- Itti, L., & Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision Research*, *49*(10), 1295–1306. <https://doi.org/10.1016/j.visres.2008.09.007>
- Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nature Human Behaviour*, *3*(7), 719–732. <https://doi.org/10.1038/s41562-019-0597-3>
- Joëls, M., Fernandez, G., & Roozendaal, B. (2011). Stress and emotional memory: A matter of timing. *Trends in Cognitive Sciences*, *15*(6), 280–288. <https://doi.org/10.1016/j.tics.2011.04.004>
- Joëls, M., Sarabdjitsingh, R. A., & Karst, H. (2012). Unraveling the time domains of corticosteroid hormone influences on brain activity: Rapid, slow, and chronic modes. *Pharmacological Reviews*, *64*(4), 901–938. <https://doi.org/10.1124/pr.112.005892>
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*(1), 237–285. <https://doi.org/10.1613/jair.301>
- Kalbe, F., Bange, S., Lutz, A., & Schwabe, L. (2020). Expectancy violation drives memory boost for stressful events. *Psychological Science*, *31*(11), 1409–1421. <https://doi.org/10.1177/0956797620958650>
- Kalbe, F., & Schwabe, L. (2020a). Beyond arousal: Prediction error related to aversive events promotes episodic memory formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 234–246. <https://doi.org/10.1037/xlm0000728>
- Kalbe, F., & Schwabe, L. (2020b). On the search for a selective and retroactive strengthening of memory: Repeated failure to find category-specific behavioral tagging. *PsyArXiv*. <https://doi.org/10.31234/osf.io/ed3z8>
- Kalbe, F., & Schwabe, L. (2021). Prediction errors for aversive events shape long-term memory formation through a distinct neural mechanism. *bioRxiv*. <https://doi.org/10.1101/2021.03.19.436177>
- Kalbe, F., & Schwabe, L. (in press). On the search for a selective and retroactive strengthening of memory: Is there evidence for category-specific behavioral tagging? *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001075>

- Kim, G., Lewis-Peacock, J. A., Norman, K. A., & Turk-Browne, N. B. (2014). Pruning of memories by context-based prediction error. *Proceedings of the National Academy of Sciences*, *111*(24), 8997–9002. <https://doi.org/10.1073/pnas.1319438111>
- Kiverstein, J., Miller, M., & Rietveld, E. (2019). The feeling of grip: Novelty, error dynamics, and the predictive brain. *Synthese*, *196*(7), 2847–2869. <https://doi.org/10.1007/s11229-017-1583-9>
- Klein, S. B., Cosmides, L., Tooby, J., & Chance, S. (2002). Decisions and the evolution of memory: Multiple systems, multiple functions. *Psychological Review*, *109*(2), 306–329. <https://doi.org/10.1037/0033-295x.109.2.306>
- Klein, S. B., Robertson, T. E., & Delton, A. W. (2010). Facing the future: Memory as an evolved system for planning future acts. *Memory & Cognition*, *38*(1), 13–22. <https://doi.org/10.3758/MC.38.1.13>
- Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719. <https://doi.org/10.1016/j.tins.2004.10.007>
- Kuhl, B. A., Dudukovic, N. M., Kahn, I., & Wagner, A. D. (2007). Decreased demands on cognitive control reveal the neural processing benefits of forgetting. *Nature Neuroscience*, *10*(7), 908–914. <https://doi.org/10.1038/nn1918>
- LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, *7*(1), 54–64. <https://doi.org/10.1038/nrn1825>
- Lak, A., Stauffer, W. R., & Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proceedings of the National Academy of Sciences*, *111*(6), 2343–2348. <https://doi.org/10.1073/pnas.1321596111>
- Lench, H. C., Flores, S. A., & Bench, S. W. (2011). Discrete emotions predict changes in cognition, judgment, experience, behavior, and physiology: A meta-analysis of experimental emotion elicitations. *Psychological Bulletin*, *137*(5), 834–855. <https://doi.org/10.1037/a0024244>
- Lengyel, M., & Dayan, P. (2008). Hippocampal contributions to control: The third way. *Advances in Neural Information Process Systems*, *20*, 889–896.
- Lombardi, O., Holik, F., & Vanni, L. (2016). What is Shannon information? *Synthese*, *193*(7), 1983–2012. <https://doi.org/10.1007/s11229-015-0824-z>
- Lonsdorf, T. B., Menz, M. M., Andreatta, M., Fullana, M. A., Golkar, A., Haaker, J., Heitland, I., Hermann, A., Kuhn, M., Kruse, O., Meir Drexler, S., Meulders, A., Nees, F., Pittig, A., Richter, J., Römer, S., Shiban, Y., Schmitz, A., Straube, B., ... Merz, C. J. (2017). Don't fear 'fear conditioning': Methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neuroscience & Biobehavioral Reviews*, *77*, 247–285. <https://doi.org/10.1016/j.neubiorev.2017.02.026>
- Macmillan, N. A., & Creelman, C. D. (2004). *Detection Theory: A User's Guide* (2nd ed.). Psychology Press. <https://doi.org/10.4324/9781410611147>
- Maia, T. V. (2009). Reinforcement learning, conditioning, and the brain: Successes and challenges. *Cognitive, Affective, & Behavioral Neuroscience*, *9*(4), 343–364. <https://doi.org/10.3758/CABN.9.4.343>

- Martin, K. C., & Kosik, K. S. (2002). Synaptic tagging – who’s it? *Nature Reviews Neuroscience*, 3(10), 813–820. <https://doi.org/10.1038/nrn942>
- Mather, M., Clewett, D., Sakaki, M., & Harley, C. W. (2016). Norepinephrine ignites local hotspots of neuronal excitation: How arousal amplifies selectivity in perception and memory. *Behavioral and Brain Sciences*, 39, e200. <https://doi.org/10.1017/S0140525X15000667>
- Matsumoto, M., & Hikosaka, O. (2009). Two types of dopamine neuron distinctly convey positive and negative motivational signals. *Nature*, 459(7248), 837–841. <https://doi.org/10.1038/nature08028>
- Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition and Emotion*, 23(2), 209–237. <https://doi.org/10.1080/02699930802204677>
- McGaugh, J. L. (2000). Memory – A century of consolidation. *Science*, 287(5451), 248–251. <https://doi.org/10.1126/science.287.5451.248>
- McGaugh, J. L. (2018). Emotional arousal regulation of memory consolidation. *Current Opinion in Behavioral Sciences*, 19, 55–60. <https://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(2), 205–210. [https://doi.org/10.1016/S0959-4388\(02\)00306-9](https://doi.org/10.1016/S0959-4388(02)00306-9)
- McIntyre, C. K., Power, A. E., Roozendaal, B., & McGaugh, J. L. (2003). Role of the basolateral amygdala in memory consolidation. *Annals of the New York Academy of Sciences*, 985(1), 273–293. <https://doi.org/10.1111/j.1749-6632.2003.tb07088.x>
- Menon, V., & Uddin, L. Q. (2010). Saliency, switching, attention and control: A network model of insula function. *Brain Structure and Function*, 214(5), 655–667. <https://doi.org/10.1007/s00429-010-0262-0>
- Miller, R. R., Barnet, R. C., & Grahame, N. J. (1995). Assessment of the Rescorla-Wagner model. *Psychological Bulletin*, 117(3), 363–386. <https://doi.org/10.1037/0033-2909.117.3.363>
- Moncada, D., Ballarini, F., & Viola, H. (2015). Behavioral tagging: A translation of the synaptic tagging and capture hypothesis. *Neural Plasticity*, 2015, 650780. <https://doi.org/10.1155/2015/650780>
- Moncada, D., & Viola, H. (2007). Induction of long-term memory by exposure to novelty requires protein synthesis: Evidence for a behavioral tagging. *Journal of Neuroscience*, 27(28), 7476–7481. <https://doi.org/10.1523/JNEUROSCI.1083-07.2007>
- Nairne, J. S., & Pandeirada, J. N. S. (2008a). Adaptive memory: Is survival processing special? *Journal of Memory and Language*, 59(3), 377–385. <https://doi.org/10.1016/j.jml.2008.06.001>
- Nairne, J. S., & Pandeirada, J. N. S. (2008b). Adaptive memory: Remembering with a stone-age brain. *Current Directions in Psychological Science*, 17(4), 239–243. <https://doi.org/10.1111/j.1467-8721.2008.00582.x>
- Nater, U. M., Moor, C., Okere, U., Stallkamp, R., Martin, M., Ehlert, U., & Kliegel, M. (2007). Performance on a declarative memory task is better in high than low cortisol responders to psychosocial stress. *Psychoneuroendocrinology*, 32(6), 758–763. <https://doi.org/10.1016/j.psyneuen.2007.05.006>

- Neiss, R. (1988). Reconceptualizing arousal: Psychobiological states in motor performance. *Psychological Bulletin*, *103*(3), 345–366. <https://doi.org/10.1037/0033-2909.103.3.345>
- Nicoll, R. A., & Roche, K. W. (2013). Long-term potentiation: Peeling the onion. *Neuropharmacology*, *74*, 18–22. <https://doi.org/10.1016/j.neuropharm.2013.02.010>
- Niv, Y., & Schoenbaum, G. (2008). Dialogues on prediction errors. *Trends in Cognitive Sciences*, *12*, 265–272. <https://doi.org/10.1016/j.tics.2008.03.006>
- Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, *35*(21), 8145–8157. <https://doi.org/10.1016/10.1523/JNEUROSCI.2978-14.2015>
- O’Doherty, J. P., Buchanan, T. W., Seymour, B., & Dolan, R. J. (2006). Predictive neural coding of reward preference involves dissociable responses in human ventral midbrain and ventral striatum. *Neuron*, *49*(1), 157–166. <https://doi.org/10.1016/j.neuron.2005.11.014>
- O’Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*(2), 329–337. [https://doi.org/10.1016/S0896-6273\(03\)00169-7](https://doi.org/10.1016/S0896-6273(03)00169-7)
- Oyarzún, J. P., Packard, P. A., de Diego-Balaguer, R., & Fuentemilla, L. (2016). Motivated encoding selectively promotes memory for future inconsequential semantically-related events. *Neurobiology of Learning and Memory*, *133*, 1–6. <https://doi.org/10.1016/j.nlm.2016.05.005>
- Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*(2), 97–98. <https://doi.org/10.1038/nn802>
- Paré, D. (2003). Role of the basolateral amygdala in memory consolidation. *Progress in Neurobiology*, *70*(5), 409–420. [https://doi.org/10.1016/S0301-0082\(03\)00104-7](https://doi.org/10.1016/S0301-0082(03)00104-7)
- Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning & Memory*, *24*(1), 65–69. <https://doi.org/10.1101/lm.042978.116>
- Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, *139*(6), 1173–1203. <https://doi.org/10.1037/a0033044>
- Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*(7106), 1042–1045. <https://doi.org/10.1038/nature05051>
- Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(17), R764–R773. <https://doi.org/10.1016/j.cub.2013.05.041>
- Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. *Journal of Neuroscience*, *28*(11), 2745–2752. <https://doi.org/10.1523/JNEUROSCI.4286-07.2008>

- Pritzel, A., Uria, B., Srinivasan, S., Badia, A. P., Vinyals, O., Hassabis, D., Wierstra, D., & Blundell, C. (2017). Neural episodic control. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th International Conference on Machine Learning* (pp. 2827–2836). PMLR.
- Quent, J. A., Henson, R. N., & Greve, A. (2021). A predictive account of how novelty influences declarative memory. *Neurobiology of Learning and Memory*, *179*, 107382. <https://doi.org/10.1016/j.nlm.2021.107382>
- 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*(1), 79–87. <https://doi.org/10.1038/4580>
- Rauch, S., Eftekhari, A., & Ruzek, J. (2012). Review of exposure therapy: A gold standard for PTSD treatment. *Journal of Rehabilitation Research and Development*, *49*, 679–88. <https://doi.org/10.1682/JRRD.2011.08.0152>
- 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). Appleton-Century-Crofts.
- Richardson, M. P., Strange, B. A., & Dolan, R. J. (2004). Encoding of emotional memories depends on amygdala and hippocampus and their interactions. *Nature Neuroscience*, *7*(3), 278–285. <https://doi.org/10.1038/nn1190>
- Rogerson, T., Cai, D. J., Frank, A., Sano, Y., Shobe, J., Lopez-Aranda, M. F., & Silva, A. J. (2014). Synaptic tagging during memory allocation. *Nature Reviews Neuroscience*, *15*(3), 157–169. <https://doi.org/10.1038/nrn3667>
- Roozendaal, B. (2002). Stress and memory: Opposing effects of glucocorticoids on memory consolidation and memory retrieval. *Neurobiology of Learning and Memory*, *78*(3), 578–595. <https://doi.org/10.1006/nlme.2002.4080>
- Roozendaal, B., McEwen, B. S., & Chattarji, S. (2009). Stress, memory and the amygdala. *Nature Reviews Neuroscience*, *10*(6), 423–433. <https://doi.org/10.1038/nrn2651>
- Roozendaal, B., McReynolds, J. R., Van der Zee, E. A., Lee, S., McGaugh, J. L., & McIntyre, C. K. (2009). Glucocorticoid effects on memory consolidation depend on functional interactions between the medial prefrontal cortex and basolateral amygdala. *Journal of Neuroscience*, *29*(45), 14299–14308. <https://doi.org/10.1523/JNEUROSCI.3626-09.2009>
- Roozendaal, B., Okuda, S., Van der Zee, E. A., & McGaugh, J. L. (2006). Glucocorticoid enhancement of memory requires arousal-induced noradrenergic activation in the basolateral amygdala. *Proceedings of the National Academy of Sciences*, *103*(17), 6741–6746. <https://doi.org/10.1073/pnas.0601874103>
- Rouhani, N., & Niv, Y. (2021). Signed and unsigned reward prediction errors dynamically enhance learning and memory. *eLife*, *10*, e61077. <https://doi.org/10.7554/eLife.61077>
- 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*(9), 1430–1443. <https://doi.org/10.1037/xlm0000518>

- Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition*, 203, 104269. <https://doi.org/10.1016/j.cognition.2020.104269>
- Rugg, M. D., & Yonelinas, A. P. (2003). Human recognition memory: A cognitive neuroscience perspective. *Trends in Cognitive Sciences*, 7(7), 313–319. [https://doi.org/10.1016/S1364-6613\(03\)00131-1](https://doi.org/10.1016/S1364-6613(03)00131-1)
- Sartory, G., Cwik, J., Knuppertz, H., Schürholt, B., Lebens, M., Seitz, R., & Schulze, R. (2013). In search of the trauma memory: A meta-analysis of functional neuroimaging studies of symptom provocation in posttraumatic stress disorder (PTSD). *PLoS ONE*, 8. <https://doi.org/10.1371/journal.pone.0058150>
- Schultz, W. (2006). Behavioral theories and the neurophysiology of reward. *Annual Review of Psychology*, 57(1), 87–115. <https://doi.org/10.1146/annurev.psych.56.091103.070229>
- Schultz, W. (2016). Dopamine reward prediction error coding. *Dialogues in Clinical Neuroscience*, 18(1), 10. <https://doi.org/10.31887/DCNS.2016.18.1/wschultz>
- Schultz, W. (2019). Recent advances in understanding the role of phasic dopamine activity. *F1000Research*, 8, F1000 Faculty Rev–1680. <https://doi.org/10.12688/f1000research.19793.1>
- Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593–1599. <https://doi.org/10.1126/science.275.5306.1593>
- Schultz, W., Tremblay, L., & Hollerman, J. R. (1998). Reward prediction in primate basal ganglia and frontal cortex. *Neuropharmacology*, 37(4), 421–429. [https://doi.org/10.1016/S0028-3908\(98\)00071-9](https://doi.org/10.1016/S0028-3908(98)00071-9)
- Schwabe, L., Bohringer, A., Chatterjee, M., & Schachinger, H. (2008). Effects of pre-learning stress on memory for neutral, positive and negative words: Different roles of cortisol and autonomic arousal. *Neurobiology of Learning and Memory*, 90(1), 44–53. <https://doi.org/10.1016/j.nlm.2008.02.002>
- Schwabe, L., Joëls, M., Roozendaal, B., Wolf, O. T., & Oitzl, M. S. (2012). Stress effects on memory: An update and integration. *Neuroscience & Biobehavioral Reviews*, 36(7), 1740–1749. <https://doi.org/10.1016/j.neubiorev.2011.07.002>
- Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry*, 20(1), 11–21. <https://doi.org/10.1136/jnnp.20.1.11>
- Seo, H., & Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. *Journal of Neuroscience*, 27(31), 8366–8377. <https://doi.org/10.1523/JNEUROSCI.2369-07.2007>
- Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*, 27(3), 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences*, 14(10), 464–472. <https://doi.org/10.1016/j.tics.2010.08.002>
- Silvetti, M., Seurinck, R., & Verguts, T. (2011). Value and prediction error in medial frontal cortex: Integrating the single-unit and systems levels of analysis. *Frontiers in Human Neuroscience*, 5, 75. <https://doi.org/10.3389/fnhum.2011.00075>

- Smeets, T., Otgaar, H., Candel, I., & Wolf, O. T. (2008). True or false? Memory is differentially affected by stress-induced cortisol elevations and sympathetic activity at consolidation and retrieval. *Psychoneuroendocrinology*, *33*(10), 1378–1386. <https://doi.org/10.1016/j.psyneuen.2008.07.009>
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. <https://doi.org/10.1037/0096-3445.117.1.34>
- Spoormaker, V. I., Blechert, J., Goya-Maldonado, R., Sämman, P. G., Wilhelm, F. H., & Czisch, M. (2012). Additional support for the existence of skin conductance responses at unconditioned stimulus omission. *NeuroImage*, *63*(3), 1404–1407. <https://doi.org/10.1016/j.neuroimage.2012.08.050>
- Spratling, M. (2008). Predictive coding as a model of biased competition in visual attention. *Vision Research*, *48*(12), 1391–1408. <https://doi.org/10.1016/j.visres.2008.03.009>
- Squire, L. R. (2009). The legacy of patient H.M. for neuroscience. *Neuron*, *61*(1), 6–9. <https://doi.org/10.1016/j.neuron.2008.12.023>
- Squire, L. R., & Zola-Morgan, J. (1991). The cognitive neuroscience of human memory since H.M. *Annual Review of Neuroscience*, *14*, 297–324. <https://doi.org/10.1146/annurev-neuro-061010-113720>
- Sridharan, D., Levitin, D. J., & Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks. *Proceedings of the National Academy of Sciences*, *105*(34), 12569–12574. <https://doi.org/10.1073/pnas.0800005105>
- Standing, L. (1973). Learning 10000 pictures. *Quarterly Journal of Experimental Psychology*, *25*(2), 207–222. <https://doi.org/10.1080/14640747308400340>
- Strange, B. A., & Dolan, R. J. (2004). β -adrenergic modulation of emotional memory-evoked human amygdala and hippocampal responses. *Proceedings of the National Academy of Sciences*, *101*(31), 11454–11458. <https://doi.org/10.1073/pnas.0404282101>
- Strange, B. A., & Dolan, R. J. (2001). Adaptive anterior hippocampal responses to oddball stimuli. *Hippocampus*, *11*(6), 690–698. <https://doi.org/10.1002/hipo.1084>
- Strange, B. A., Duggins, A., Penny, W., Dolan, R. J., & Friston, K. J. (2005). Information theory, novelty and hippocampal responses: Unpredicted or unpredictable? *Neural Networks*, *18*(3), 225–230. <https://doi.org/10.1016/j.neunet.2004.12.004>
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44. <https://doi.org/10.1007/BF00115009>
- Sutton, R. S., Barto, A., & Williams, R. (1992). Reinforcement learning is direct adaptive optimal control. *IEEE Control Systems Magazine*, *12*(2), 19–22. <https://doi.org/10.1109/37.126844>
- Trapp, S., O'Doherty, J. P., & Schwabe, L. (2018). Stressful events as teaching signals for the brain. *Trends in Cognitive Sciences*, *22*(6), 475–478. <https://doi.org/10.1016/j.tics.2018.03.007>
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of Memory* (pp. 381–402). Academic Press.

- van Kesteren, M. T., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, 35(4), 211–219. <https://doi.org/10.1016/j.tins.2012.02.001>
- Vogel, S., Klüen, L. M., Fernández, G., & Schwabe, L. (2018). Stress affects the neural ensemble for integrating new information and prior knowledge. *NeuroImage*, 173, 176–187. <https://doi.org/10.1016/j.neuroimage.2018.02.038>
- Vogel, S., & Schwabe, L. (2016). Stress in the zoo: Tracking the impact of stress on memory formation over time. *Psychoneuroendocrinology*, 71, 64–72. <https://doi.org/10.1016/j.psyneuen.2016.04.027>
- Wessel, J. R., Danielmeier, C., Morton, J. B., & Ullsperger, M. (2012). Surprise and error: Common neuronal architecture for the processing of errors and novelty. *Journal of Neuroscience*, 32(22), 7528–7537. <https://doi.org/10.1523/JNEUROSCI.6352-11.2012>
- Whitehead, S. D., & Lin, L.-J. (1995). Reinforcement learning of non-Markov decision processes. *Artificial Intelligence*, 73(1), 271–306. [https://doi.org/10.1016/0004-3702\(94\)00012-P](https://doi.org/10.1016/0004-3702(94)00012-P)
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.
- Williams, L. M., Phillips, M. L., Brammer, M. J., Skerrett, D., Lagopoulos, J., Rennie, C., Bahramali, H., Olivieri, G., David, A. S., Peduto, A., & Gordon, E. (2001). Arousal dissociates amygdala and hippocampal fear responses: Evidence from simultaneous fMRI and skin conductance recording. *NeuroImage*, 14(5), 1070–1079. <https://doi.org/10.1006/nimg.2001.0904>
- Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, 114(1), 152–176. <https://doi.org/10.1037/0033-295X.114.1.152>
- Wolf, O. T. (2012). Immediate recall influences the effects of pre-encoding stress on emotional episodic long-term memory consolidation in healthy young men. *Stress*, 15(3), 272–280. <https://doi.org/10.3109/10253890.2011.622012>
- Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), 1341–1354. <https://doi.org/10.1037/0278-7393.20.6.1341>
- Yonelinas, A. P., & Parks, C. (2007). Receiver operating characteristics (ROCs) in recognition memory: A review. *Psychological Bulletin*, 133, 800–832. <https://doi.org/10.1037/0033-2909.133.5.800>

Appendix A: Study 1

A

Kalbe, F., & Schwabe, L. (in press). On the search for a selective and retroactive strengthening of memory: Is there evidence for category-specific behavioral tagging? *Journal of Experimental Psychology: General*. <https://doi.org/10.1037/xge0001075>

On the Search for a Selective and Retroactive Strengthening of Memory: Is There Evidence for Category-Specific Behavioral Tagging?

Felix Kalbe and Lars Schwabe

Department of Cognitive Psychology, Institute of Psychology, Universität Hamburg

Storing motivationally salient experiences preferentially in long-term memory is generally adaptive. Although such relevant experiences are often immediately obvious, a problem arises when the relevance of initially ambiguous events becomes evident sometime after encoding. Is there a mechanism that enables the retroactive enhancement of specific memories? Recent evidence suggests the existence of such a mechanism that selectively strengthens weak memories for neutral stimuli from one category when their respective category gains motivational significance later. Although such a selective retroactive memory enhancement has considerable implications for adaptive memory, evidence for this phenomenon is based on only few studies. Here, we report data from four attempts to replicate category-specific retroactive memory enhancements for neutral stimuli from a category that was later predictive of aversive electric shocks. Although our data showed enhanced memory for the arousing stimuli themselves as well as related subsequent stimuli, none of our experiments provided any evidence for category-specific retroactive memory enhancement when strictly replicating the analysis strategy from the original study. In an additional analysis focusing on high confidence memory only, one of four experiments indicated a significant retroactive memory effect but only in corrected recognition and not in d' based on signal detection theory. In an analysis pooled across all experiments, we found a small but significant retroactive memory effect again solely for high-confidence corrected recognition, although the corresponding Bayesian analysis indicated even substantial evidence for the null hypothesis. Overall, our data cast doubt on the reliability and generalizability of the proposed selective retroactive enhancement of initially weak memory.


Keywords: adaptive memory, behavioral tagging, episodic memory, reproducibility

Supplemental materials: <https://doi.org/10.1037/xge0001075.supp>

Our memories provide not only a window into the past but may also guide our future behavior. In particular, detailed memories of past experiences allow predicting future events as well as the potential

consequences of actions and can therefore serve as a basis for optimized choices in complex environments (Gershman & Daw, 2017; Murty et al., 2016). However, of the numerous experiences that we make every day, only few are of significant value for future decisions. According to the theory of adaptive memory, these motivationally significant experiences should be preferentially stored in episodic memory (Nairne et al., 2007; Nairne & Pandeirada, 2008; Shohamy & Adcock, 2010). Phylogenetically, such an adaptive memory might have been critical to survival by allowing the identification and subsequent avoidance of potentially threatening situations, thereby improving fitness (Nairne & Pandeirada, 2008). The preferential memory processing is relevant because limited memory resources during both encoding and retrieval should optimally be reserved for motivationally relevant experiences.

Such motivationally salient experiences are usually immediately obvious to an individual. Exciting or stressful experiences elicit physiological arousal during encoding, a well-known factor that promotes episodic memory formation (Cahill & McGaugh, 1998; LaBar & Cabeza, 2006; McGaugh, 2018; Schwabe et al., 2012; Vogel & Schwabe, 2016). However, other events appear initially neutral or mundane and their link to important consequences is only later revealed. Consider a bank customer entering her local branch as usual, when another presumed customer is leaving in a hurry. She barely

Lars Schwabe  <https://orcid.org/0000-0003-4429-4373>

Part of the data of Experiments 1 and 2 were analyzed to address a research question unrelated to the present article and has been reported in Kalbe, F., & Schwabe, L. (2020). Beyond arousal: Prediction error related to aversive events promotes episodic memory formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 234–246.

We gratefully acknowledge the assistance of Friederike Baier, Pia D'Agostino, Leandra Feldhusen, Hülya Keskin, Manuel Krohn, Vincent Kühn, Moana Lamm, Fabian Schacht, Felix Schiborn, Celine Schneller, Anne-Sophie Siegel, Elizabeth Sievert, Seher Teymuroglu, and Till Thelosen during data collection. We further thank Joseph Dunsmoor for providing the stimulus materials used in Experiments 2, 3, and 4 and for helpful advice on instructions and experimental procedure. The behavioral data analyzed in this series of experiments can be found at <https://osf.io/qpm3t/>.

Correspondence concerning this article should be addressed to Lars Schwabe, Department of Cognitive Psychology, Institute of Psychology, Universität Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany. Email: Lars.Schwabe@uni-hamburg.de

notices his face. Soon it is revealed to her that the person she saw leaving had just robbed the bank and she realizes that she will soon be asked by the police to give a detailed description of the alleged robber's unmasked face. Because the initial encounter with the alleged robber was not particularly remarkable and therefore not paralleled by significant arousal, his face was not encoded preferentially. Is there still a mechanism to make such initially weak memories last? Adaptive memory would call for a mechanism that temporarily and nonselectively stores recent experiences and transfers them to long-term memory when a motivationally significant event follows within a certain time window. Indeed, such a mechanism was first discovered at the synaptic level and inspired the *tag and capture hypothesis* (Frey & Morris, 1997, 1998; Martin & Kosik, 2002; Rogerson et al., 2014). According to this hypothesis, at least two distinguishable steps are necessary to achieve long-term potentiation for initially weak experiences, the dominant neurophysiological model of long-term memory (Bliss & Collingridge, 1993; Malenka & Nicoll, 1999). First, weak stimulation of a neuron creates a local transient tag at a synapse, which decays within hours and by itself is insufficient to create long-lasting memories. To produce long-term potentiation, additional plasticity-related proteins are required that result from a stronger stimulation of the neuron. These proteins bind to the synaptic tag set earlier (i.e., the capturing step) inducing long-term physiological changes in synaptic signaling. Critically, plasticity related proteins evoked through strong stimulation of the neuron can bind to synaptic tags set earlier through unrelated weak stimulation and therefore create lasting memories for events that would by themselves be too weak to produce long-term potentiation (Moncada & Viola, 2007; Redondo & Morris, 2011). This mechanism therefore provides a neurophysiological basis for the retroactive memory enhancement of events with an initially unclear motivational significance (Moncada et al., 2015).

Evidence that the synaptic tag and capture hypothesis can be translated to behavior has been found in both rodents (Almaguer-Melian et al., 2012; Ballarini et al., 2009; de Carvalho Myskiw et al., 2013; Moncada & Viola, 2007; Wang et al., 2010) and more recently in humans (Ballarini et al., 2013). In paradigms demonstrating a *behavioral tagging* mechanism, participants first superficially encode stimuli. This encoding session is then followed by either a significant event (e.g., an aversive or novel experience) or a nonsignificant control event. Subsequent memory tests typically show that the significant event—compared with a neutral control event—retroactively enhanced memory for the previously encoded stimuli. In these paradigms, retroactive memory enhancements are usually unspecific in the sense that an event enhances memory for any stimuli encoded within a certain time window before the significant event, even if these are not directly linked to the latter. In the case of the bank robbery, such unrelated details might include the color of the tie the bank clerk was wearing at the time of the robbery. From an adaptive memory perspective, promoting memory for such irrelevant details might be regarded as suboptimal when they lack any predictive value for the memory-promoting event.

A recent study suggests that there is—in addition to rather broad and unspecific behavioral tagging—a retroactive memory enhancement that is highly specific (Dunsmoor et al., 2015). This study combined an incidental encoding task with a fear learning procedure. In a preconditioning phase, participants first encoded neutral pictures of animals and tools and were asked to indicate to which of the two categories a picture

belonged. Following this weak encoding session, in a Pavlovian fear conditioning phase, additional, previously unseen pictures from the same two categories were presented. Pictures from one of the two categories (i.e., either animals or tools; CS^+) were followed by an aversive electric shock in two thirds of all trials, while pictures from the remaining category (CS^-) were never followed by a shock. Whether shocks followed pictures of animals or tools was counterbalanced across participants. A postconditioning phase with an identical procedure as the preconditioning phase but novel stimuli followed the fear-conditioning phase. To test participants' memory for stimuli from the three encoding phases, a surprise recognition test followed either immediately, 6 hr, or 24 hr later (manipulated between subjects). In this recognition test, participants saw all previously presented pictures of animals and tools together with the same number of previously unseen (new) pictures from both categories and classified each picture as either *old* or *new*. Results showed an enhanced recognition performance for CS^+ pictures encoded during fear-conditioning compared with CS^- pictures encoded in the same phase in all three delay groups. In the 24-hr delay group, this CS^+ memory carried over to pictures presented after the fear-conditioning phase, although these items were never paired with a shock themselves. Most importantly, however, the authors found category-specific retroactive memory enhancements in both the 6-hr and 24-hr delay groups, as indicated by better recognition of CS^+ pictures encoded before the fear-conditioning compared with CS^- pictures encoded in the same phase. This finding is particularly remarkable because participants had no information about shock contingencies being linked to one of the two categories when these pictures were encoded. When the recognition test followed immediately after the encoding, no category-specific retroactive memory enhancement was observable, suggesting the critical involvement of consolidation processes. Interestingly, there also was a negative linear relationship between the size of the retroactive memory effect and the temporal proximity of preconditioning items to the fear-conditioning procedure, suggesting that pictures from the CS^+ category that were encoded first (i.e., furthest from the following fear-conditioning) received the strongest memory enhancement. Furthermore, another group of participants encoded stimuli from the preconditioning phase more strongly through repeated presentation of each picture. These participants showed no signs of category-specific retroactive memory enhancement after 24 hr, indicating that only initially weak memories are susceptible to this effect, a finding that is congruent with the literature on synaptic tagging (Frey & Morris, 1997, 1998; Martin & Kosik, 2002; Rogerson et al., 2014).

Another study from the same group of authors showed that selective, category-specific retroactive memory enhancements cannot only be triggered through aversive events, but also through reward (Patil et al., 2017). Following a similar design as the study by Dunsmoor et al. (2015), the authors showed that memory for initially neutral pictures of animals and tools could be enhanced for the category that was later associated with high compared with low reward opportunities in a delayed matching-to-sample task. Notably, in this task, participants were rewarded for correct responses, whereas shocks were

independent of participants' actions in the study by Dunsmoor et al. (2015). In contrast to these findings, another study from an independent lab using a similar classical conditioning procedure as Dunsmoor et al. (2015) featuring monetary reward instead of aversive shocks obtained no evidence for category-specific retroactive memory enhancement (Oyarzún et al., 2016). To our knowledge, no other studies so far have investigated category-specific retroactive memory enhancement, neither in the aversive, nor in the appetitive domain.

The findings showing a selective, retroactive memory enhancement are exciting; they provide novel insights into how our memory works and may have considerable practical implications for clinical or legal settings. A selective behavioral tagging mechanism may also inspire new tools for boosting memory retrospectively. Given the far-reaching implications of selective, retroactive memory enhancements, we initially aimed to shed light on the cognitive mechanisms underlying this effect. However, what started as an attempt to unravel the fundamental mechanisms underlying selective behavioral tagging, turned out to be a search for the phenomenon itself. We present here evidence from four experiments aimed to replicate findings of category-specific retroactive memory enhancement through aversive electric shocks (Dunsmoor et al., 2015).

Experiment 1: Testing the Fear-Related Category-Specific Retroactive Memory Enhancement

Experiment 1 was designed to replicate findings of category-specific retroactive memory enhancement in the context of an aversive learning task (Dunsmoor et al., 2015). Because the experiment of Dunsmoor et al. (2015) showed that observed retroactive memory effects were most pronounced in a recognition test 24 hr after encoding, we used here a 24-hr interval between encoding and recognition test. Instead of the original stimulus set, we used pictures that were conceptually very similar to those used by Dunsmoor et al. (2015); that is, also pictures from the categories 'animals' and 'tools'. Procedural differences included the placing of the shock electrode on the lower leg (rather than on the wrist as in the original study) and employing a two-stage recognition test (rather than a single-stage as in the original study) that first asked participants to indicate whether an item was old or new, followed by their certainty with this decision. As further discussed below, we implemented a different CS-UCS timing compared with Dunsmoor et al. (2015). Finally, we did not control for stimulus typicality across encoding phases because this aspect was not mentioned in Dunsmoor et al. (2015). Instead, it was only revealed during later stages of the peer-review process for this article that Dunsmoor et al. (2015) controlled for stimulus typicality. This aspect is later explicitly addressed in Experiment 4.

Method

Participants

Forty-four healthy participants (30 women) between 19 and 33 years of age took part in this experiment ($M = 25.05$, $SD = 3.75$). This sample size was based on an a priori sample size calculation with G*Power 3 (Faul et al., 2007). Dunsmoor et al. (2015) reported retroactive memory improvements from a paired t -test for items conceptually related to the CS⁺ compared with items related

to the CS⁻ in the 24-hr retrieval group with weak encoding ($n = 30$) and obtained a t value of 2.48 with an effect size of $d_{av} = .41$. Based on this information, Cohen's d_z , another measure of effect size in within-subject designs used by G*Power, can be calculated using the following formula (Lakens, 2013):

$$\text{Cohen's } d_z = \frac{t}{\sqrt{n}}$$

Using the values reported by Dunsmoor et al. (2015) yielded the following estimate for Cohen's d_z :

$$\text{Cohen's } d_z = \frac{2.48}{\sqrt{30}} = 0.45$$

We treated this effect size as a point estimate for the category-specific retroactive memory effect in our power analysis. This indicated that, using a two-tailed paired t -test with $\alpha = .05$, at least 41 participants would be required to detect such an effect with 80% certainty. This target sample size also represents an approximately 40% increase compared with the 24-hr group in the original study ($n = 30$). Exclusion criteria for participation in this experiment comprised any current or past physical or mental illness, electric medical devices such as pacemakers, and pregnancy in women. Participants gave written informed consent prior to testing and received a monetary compensation of 20€ after completing the experiment. The ethics committee of the Faculty of Psychology and Human Movement Sciences of the Universität Hamburg approved the study protocol.

Materials

As in the original study, stimuli were 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. We acquired photographs from the Bank of Standardized Stimuli (Brodeur et al., 2010; Brodeur et al., 2014) and from publicly available Internet sources. All photographs were of neutral valence and selected to be unique exemplars of their respective category. For example, there were not two different photographs of dogs or two different photographs of hammers. From the total pool of 360 photographs, 180 (90 animals, 90 tools) were randomly selected per participant to serve as learning items, while the remaining 180 served as lures for the surprise recognition test on the second experimental day. The 180 learning items were then randomly allocated to the three different incidental encoding phases for the first experimental day, such that each phase featured 30 photographs of animals and 30 photographs of tools.

Procedure

The first experimental day featured an incidental encoding session with three phases: a preconditioning phase, a fear conditioning phase, and a postconditioning phase. Approximately 24 hr later, participants completed a surprise recognition test for photographs that had been presented in all three encoding phases on the previous day. Unlike in the study by Dunsmoor et al. (2015), we did not vary the interval between encoding and recognition test between subjects but kept it fixed at 24 hr, because the 24 hr group had previously shown the clearest evidence for both category-specific retroactive and prospective memory enhancement. Additionally, another study featuring a reward learning task also demonstrated category-specific

retroactive memory enhancement only after a 24-hr interval, but not in an immediate recognition test, suggesting a crucial role of a sufficiently long consolidation period before the recognition test (Patil et al., 2017).

Upon arrival on the first experimental day, participants gave written informed consent and received detailed written instructions about the following three learning phases. Importantly, they were not informed that the study investigated episodic memory, nor that a recognition test would follow on the second experimental day. In the *preconditioning phase*, participants saw 30 photographs of animals and 30 photographs of tools in a pseudorandomized order, such that no more than three photographs from the same category could appear in a row. Each stimulus was presented for 2.5 s, during which participants should indicate whether the photograph showed an animal or a tool by pressing the '1' or '2' button on the computer keyboard (see Figure 1). Each stimulus was followed by a black fixation cross on a white background for $6 \text{ s} \pm 2 \text{ s}$. The allocation of buttons to each of the two categories was counterbalanced across participants. The total duration of the preconditioning phase was approximately 8 min.

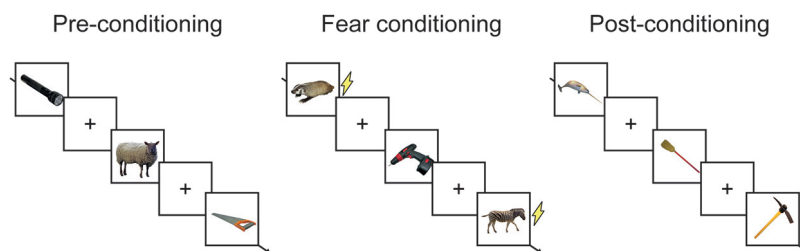
Before the *conditioning phase*, we attached electrodes on the distal phalanx of the second and third finger of the left hand to record skin conductance responses (SCRs). Skin conductance was measured using a MP-160 BIOPAC system (BIOPAC systems, Goleta, CA). An additional STM-200 module (BIOPAC systems, Goleta, CA) was connected to the MP-160 for electrical stimulation. The stimulation electrode was placed on the right lower leg, approximately 25 cm centrally above the heel. To determine the individual stimulation intensity, we used a standardized procedure consisting of twelve 200-ms single-pulse shocks with an initial intensity of 20 V. After each trial, participants rated the shock that they had just received as either painful or not painful in a forced-choice fashion. Whenever a shock was rated as not painful, its intensity was increased slightly in the following trial. Similarly, whenever participants rated a shock as painful, the intensity was decreased slightly. The goal was to select an intensity that participants perceived as

unpleasant, but not painful. In total, these steps following the preconditioning phase took approximately 10 min.

The following conditioning phase again consisted of 30 photographs of animals and 30 photographs of tools, none of which had been presented before. As in the preconditioning phase, stimuli were presented in a pseudorandomized order, so that no more than three photographs from the same category appeared in a row. Each photograph was presented centrally on the screen for 4.5 s, during which participants were instructed to make a binary prediction about the possible occurrence of a following shock using the '1' and '2' buttons on the keyboard, corresponding to *no shock* and *shock*, respectively. In 20 of the 60 trials in this phase, a 200-ms electric shock was presented immediately after the offset of the photograph. Note that in Dunsmoor et al. (2015), shocks coterminated with photograph presentation, leading to a 200ms relative offset of the shock in our replication attempt. This procedural difference was unintentional and addressed in a later experiment (Experiment 4).

Importantly, shock contingencies were linked to the item categories, such that one image category (e.g., tools) served as the CS^+ category, whereas the remaining category (e.g., animals) was never paired with a shock and thus served as the CS^- category. Whether photographs of animals or tools served as the CS^+ category was counterbalanced across participants. In CS^+ trials, the shock probability was two thirds, with a fixed number of 20 shocks occurring in the 30 CS^+ trials. In CS^- trials, on the other hand, none of the photographs was followed by a shock. Participants were not informed about category-shock contingencies but had to learn them by trial and error. To avoid that participants could misinterpret shocks as consequences of their actions, they were explicitly told that their choices had no effect on the probability that a shock would occur (Dunsmoor et al., 2015). Each trial was followed by a black fixation cross on a white background for $8 \pm 2 \text{ s}$, which enabled measuring the relatively slow SCRs elicited by electric shocks and allowed skin conductance levels to return to baseline before the next trial started. The total duration of the conditioning phase was approximately 12 min. After the

Figure 1
Procedure in Experiments 1–4



Note. In each phase, participants saw 60 unique photographs of animals and tools. During pre- and postconditioning, they were instructed to categorize each photograph as an animal or tool. During fear conditioning, photographs from one category (CS^+ ; animals in the example above) were followed by an electric shock in two-thirds of all trials, whereas photographs of the remaining category (CS^- ; tools in the example above) were never followed by a shock. Whether photographs of animals or tools served as the CS^+ category was counterbalanced across participants. For each photograph, participants were instructed to indicate whether they expected that a shock would follow. Note that in Experiment 3, the interval between preconditioning and Pavlovian fear conditioning was increased by 10 min, based on previous reports that this would lead to increased category-specific retroactive memory enhancement (Dunsmoor et al., 2015). Approximately 24h after encoding, participants completed a surprise recognition test in which they saw all previously presented photographs of animals and tools together with the same number of new photographs and indicated for each of them whether they thought it had been presented on the previous day. See the online article for the color version of this figure.

conditioning phase, we removed both the SCR- and the shock-electrodes. Participants then rated the shock intensity on a scale from 1 (*not unpleasant at all*) to 10 (*extremely unpleasant*).

The subsequent *postconditioning phase* consisted of 60 previously unseen photographs (30 animals and 30 tools) and otherwise followed an identical procedure as the preconditioning phase. Thus, the duration of the postconditioning phase was approximately 8 minutes again.

Participants returned for a *memory test* 22 hr to 26 hr after encoding on the first experimental day. They first completed a short questionnaire to assess whether they had already anticipated the following recognition test. To this end, after being informed about the following memory test, they rated how surprised they were about the upcoming memory test on a scale from 1 (*not surprised at all*) to 5 (*very surprised*). For later analyses, we inverted the scale of this measure so that larger values indicate less surprise as in Dunsmoor et al. (2015). Next, they received written instructions explaining details of the recognition test. In the recognition test, they were presented all 180 photographs from the three encoding phases of the previous day intermixed with an equal number of “new” photographs (i.e., photographs that had not been presented previously). Half of these lures were photographs of animals and half were photographs of tools. Stimuli were presented one by one centrally on a white background. For each of these photographs, participants first decided whether it was “old” or “new” in a forced-choice fashion. Then, participants had to indicate how confident they were that this decision was correct by pressing buttons corresponding to *very unsure* (German: *sehr unsicher*), *rather unsure* (*eher unsicher*), *rather sure* (*eher sicher*) and *very sure* (*sehr sicher*). If in any of the two stages no response was given within 5 s, the rest of the trial was skipped. Between trials, a black fixation cross on a white background was presented centrally for $1.5 \text{ s} \pm .5 \text{ s}$.

Data Analysis

Confirmatory statistical analyses were kept as close as possible to the analyses described in the original study by Dunsmoor et al. (2015). Specifically, these memory analyses were performed on corrected recognition scores to account for different response criteria between subjects. These were derived by subtracting the individual per image category false alarm rate from the per image category and per phase hit rate. Responses were collapsed across confidence, that is, only the forced-choice decision between “old” and “new” items was considered for memory performance. Besides *t*-tests on corrected recognition scores as reported by Dunsmoor et al. (2015), we also report *t*-tests on sensitivity scores (d') based on signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002). Before computing their *z* scores from the standard normal distribution, hit- and false-alarm-rates were restricted to the range of 1% to 99%. All *t*-tests were two-tailed.

Further, Bayes factors were calculated using the *ttestBF* R-function from the *BayesFactor* package to directly compare the adequacy of the null hypothesis H_0 that the true effect is equal to zero against the one-sided alternative hypothesis H_1 that the effect is greater than zero. We applied a Cauchy prior distribution with a default scale parameter of $r = .707$ (Morey et al., 2018; Rouder et al., 2009). The resulting BF_{10} metric indicates relative evidence for the H_1 versus the H_0 such that values greater than 1 favor the

alternative hypothesis H_1 and values smaller than 1 favor the null hypothesis H_0 . We interpret values greater than 3 as substantial evidence for the H_1 , while values smaller than 1/3 are interpreted as substantial evidence for the H_0 (Jarosz & Wiley, 2014).

As a manipulation check for successful fear conditioning, we analyzed skin conductance data obtained during the second encoding phase using both (a) a continuous decomposition analysis (CDA) and (b) a more classic through-to-peak (TTP) analysis, which was more similar to the SCR analysis in Dunsmoor et al. (2015), using Ledalab Version 3.4.9 (Benedek & Kaernbach, 2010). First, the skin conductance signal was downsampled to a resolution of 50 Hz and optimized using four sets of initial values. The minimum amplitude threshold was set to $.01 \mu\text{S}$. For each trial during the conditioning phase, we derived anticipatory SCRs as the average phasic driver within a response window of .5 s to 4.5 s after each stimulus onset to obtain CDA-estimates. Like Dunsmoor et al. (2015), we also obtained more classic through-to-peak results, expressed as the sum of significant SCR-amplitudes within the specified response window. Importantly, as shocks always appeared exactly 4.5 s after stimulus onset and therefore outside the response window, the resulting estimates could not have been biased by the UCS.

Results and Discussion

Successful Fear Conditioning

An analysis of skin conductance responses confirmed that our procedure successfully induced conditioned fear for items from the CS^+ category. During the conditioning phase, participants showed significantly higher anticipatory SCRs to CS^+ items compared with CS^- items (TTP: $t[43] = 4.20$, $p < .001$, $d_{av} = .51$; CDA: $t[43] = 4.79$, $p < .001$, $d_{av} = .52$; Figure 2).

Anticipation of the Recognition Test

On the second experimental day, participants were first informed about the following recognition test for photographs from the previous day and then rated how surprised they were by this task on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). Responses from six participants were missing. The average response in the remaining sample was 3.08 ($SD = .97$), showing that, on average, participants were moderately surprised. Four participants indicated that they were *not surprised at all*. Exclusion of these four participants had no effect on the pattern of results. Therefore, these participants were still included in the following analysis.

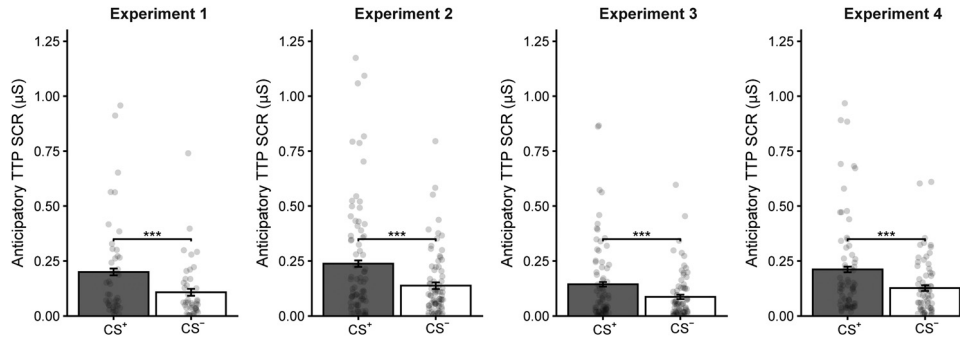
Overall Memory Performance

Overall, participants performed well in the recognition task (see Table 1), as reflected in a markedly higher average hit rate for items from all three encoding phases (i.e., the rate of correctly classifying previously seen photographs as old) of 69.6% ($SD = .11$) than the false alarm rate (i.e., the rate of incorrectly classifying previously unseen photographs as old) of 24.4% ($SD = .09$).

No Evidence for Category-Specific Retroactive Memory Enhancement

To address our main research question, we investigated how recognition performance for the photographs presented on the first experimental day was affected by the encoding phase (before, during or after the fear conditioning) and the conditioning category

Figure 2
Successful Fear Conditioning as Indicated by Average Anticipatory Skin Conductance Responses (SCRs) in Experiments 1–4 That Were Estimated Using a Through-to-Peak Analysis Similar to Dunsmoor et al. (2015)



Note. In all four experiments, participants showed significantly greater anticipatory skin-conductance responses during fear conditioning to CS⁺ items compared with CS⁻ items, confirming a successful fear induction. TTP = through-to-peak. Error bars represent ± 1 SEM.

*** $p < .001$.

(CS⁺ or CS⁻) an item belonged to through a repeated-measures ANOVA on corrected recognition scores. For the factor phase, Mauchly's test indicated that the sphericity assumption was violated, $W = .85$, $p = .030$. Hence, results for the factor phase are reported after applying a Greenhouse-Geisser correction. Overall, corrected recognition scores differed according to the phase an item was encoded in, $F(1.69, 72.73) = 17.1$, $p < .001$, $\eta^2_G = .05$. Whether an item belonged to the conditioned category, on the other hand, had no significant overall effect on recognition performance, although a trend was visible, $F(1, 43) = 3.92$, $p = .054$, $\eta^2_G = .01$. There was no significant interaction between the encoding phase and the conditioning category an item belonged to, $F(2, 86) = 1.65$, $p = .20$, $\eta^2_G = .004$. We further performed paired t -tests comparing the corrected recognition for items from the CS⁺ category versus items from the CS⁻ category separately per phase. These confirmed previous findings of an enhanced memory formation for CS⁺ items versus CS⁻ items in the conditioning phase, $t(43) = 2.31$, $p = .025$, $d_{av} = .35$ (Figure 3, upper left panel; Dunsmoor et al. 2015). At trend level, there was evidence that this memory benefit persisted for CS⁺ items over CS⁻ items in the postconditioning phase, even though these photographs were never directly paired with the UCS, $t(43) = 1.82$, $p = .076$, $d_{av} = .29$. Critically, for items that were encoded during the preconditioning phase, there was no evidence for a category-specific retroactive memory enhancement for CS⁺ items over CS⁻ items, $t(43) = .36$, $p = .72$, $d_{av} = .06$.

Finally, we also tested for preconditioning items the previously reported positive linear relationship between the temporal distance to the conditioning phase and the size of category-specific retroactive memory enhancements (Dunsmoor et al., 2015). To this end, CS⁺ and CS⁻ preconditioning items were each binned in tertiles corresponding to trials 0–10, 11–20, and 21–30 relative to the conditioning phase. A repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ as the dependent variable and the time bin as a within-subject factor showed no significant effect of time bins, $F(2, 86) = .20$, $p = .82$, $\eta^2_G = .002$. In contrast with previous reports (Dunsmoor et al., 2015), this finding indicates that the relative time of encoding of an item within the preconditioning phase had no effect on a putative category-specific retroactive memory enhancement.

Complementary Analyses

Although previous analyses showed no evidence for any category-specific retroactive memory enhancement, these relied on classic frequentist statistics and can therefore only indicate evidence *against*, but not *in support of* the null hypothesis. To this end, we reanalyzed previously reported classic paired t -tests with their Bayesian counterparts (see the Method section). For items encoded during the conditioning phase, these provided substantial support for the alternative hypothesis of a positive memory effect for CS⁺ compared with CS⁻ item from the same phase, $BF_{10} = 3.53$. Similarly, for items encoded after fear conditioning, results also favored the alternative hypothesis of a memory

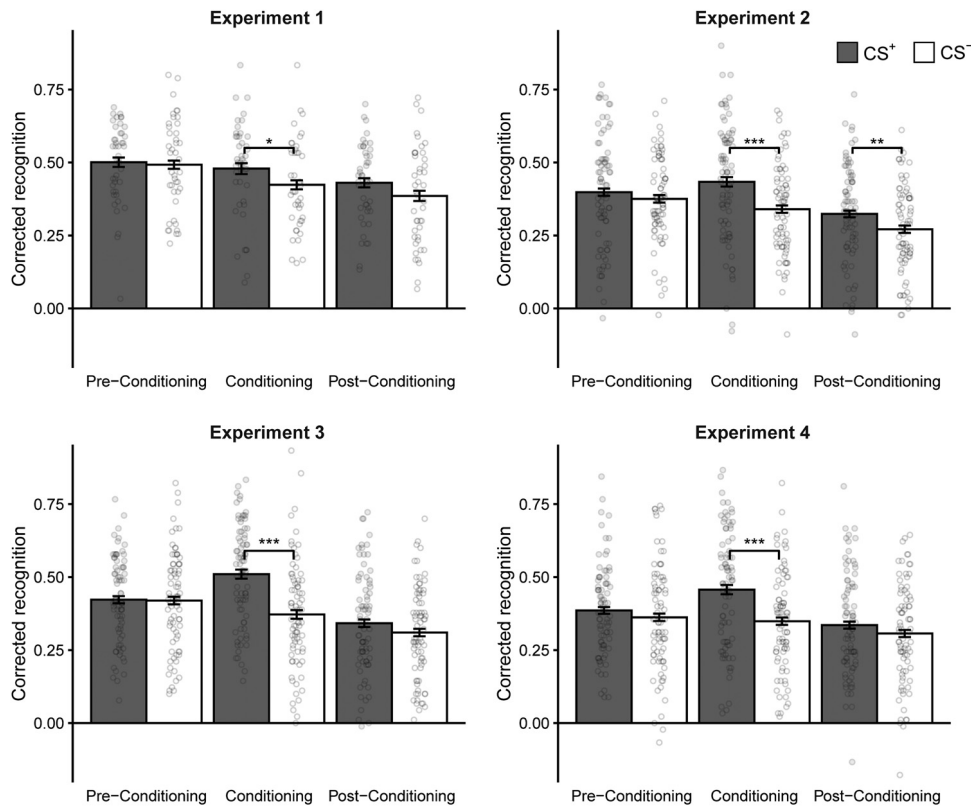
Table 1

Retrieval Memory Results in Experiment 1, Mean Proportion of Responses by Certainty

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.603	0.148	0.129	0.120	0.571	0.162	0.126	0.141
Conditioning	0.580	0.148	0.134	0.138	0.511	0.154	0.154	0.181
Postconditioning	0.521	0.158	0.152	0.170	0.482	0.145	0.168	0.205
New	0.121	0.126	0.236	0.518	0.121	0.119	0.227	0.533

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

Figure 3
Recognition Performance Expressed as Hit Rate Minus False Alarm Rate in Experiments 1–4 by Encoding Phase and Conditioning Category



Note. In all four experiments, recognition was improved for items from the CS⁺ category that were encoded during Pavlovian fear conditioning. Only in Experiment 2 was this effect significantly carried over to items encoded after the end of the fear conditioning. Most importantly, none of the four experiments provided any evidence for category-specific retroactive memory enhancement. Error bars represent ± 1 SEM.

* $p < .05$. ** $p < .01$. *** $p < .001$.

advantage for CS⁺ items relative to CS⁻ items, although evidence was only anecdotal, $BF_{10} = 1.41$. Most importantly, for preconditioning items, Bayesian analysis further provided substantial support for the null hypothesis rejecting any category-specific retroactive memory enhancement, $BF_{10} = .22$.

Dunsmoor et al. (2015) performed all memory analyses on corrected recognition scores, defined as hit rates minus false alarm rates. Here, we repeated their main analyses using sensitivity scores (d') based on signal detection theory, another common measure of recognition performance in the memory literature (Macmillan & Creelman, 2005; Wickens, 2002). These parallel analyses showed no significant differences between CS⁺ and CS⁻ items in any of the three encoding phases, all t s < 1.60 , all p s $> .12$. In the online supplemental materials, we further present results of parallel analyses using generalized linear mixed-effect models, showing an identical pattern of results as in the analysis based on memory sensitivity.

To identify possible factors hindering us from replicating the category-specific retroactive memory effect, we performed additional analyses beyond merely replicating the analysis strategy reported by Dunsmoor et al. (2015). Notably, participants were slightly less surprised by the recognition test than in the original study. However, there was no significant correlation between

memory test anticipation and recognition performance, Spearman's $r_s = .22$, $p = .18$. Further, overall memory performance per participant did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS⁺ minus mean SCRs to CS⁻), TTP: Spearman's $r_s = -.01$, $p = .93$, CDA: Spearman's $r_s = -.01$, $p = .95$. Only in very few trials of the recognition test, participants failed to respond quickly enough rejecting the notion that this might have biased our results. The mean number of missed trials per participant was .39 ($SD = .75$) of 360.

Experiment 2: Testing the Fear-Related Category-Specific Retroactive Memory Enhancement With Increased Statistical Power and the Original Stimulus Set

Despite successful fear conditioning and a replication of the procedure and analysis strategy from Dunsmoor et al. (2015), we found no evidence for category-specific retroactive memory enhancements in Experiment 1. It should be noted that, although conceptually very similar, Experiment 1 did not use the original stimulus set. Additionally, there were subtle differences in the procedure. For example, in our recognition test, "old" versus "new" decisions and memory confidence were tested separately for each

item, while this was a single step in the original study (Dunsmoor et al., 2015). Although there is no theoretical justification how these small deviations from the original study should prevent the detection of the proposed category-specific retroactive memory effect, we aimed to investigate whether we could replicate the findings when using the original stimulus set and sticking closer to the original procedure. Therefore, we contacted the lead author of the original study and asked him to provide us the stimulus materials, including all experimental instructions, which were used in the original study (Dunsmoor et al., 2015). We received these materials and used them for a direct replication of the original study by Dunsmoor et al. (2015). Additionally, we substantially increased the sample size compared with Experiment 1 to minimize the chance of null findings attributable to insufficient statistical power. Experiment 2 used the same variation in the CS-UCS timing as Experiment 1. Further, we did not control for stimulus typicality in this experiment as this aspect was not mentioned in Dunsmoor et al. (2015) and brought forward to us at a later stage.

Method

Participants

Eighty-four healthy participants (60 women) between 19 and 35 years of age took part in Experiment 2 ($M = 25.23$, $SD = 4.08$). Four participants had to be excluded from the analysis, either because they did not return for the memory test on the second experimental day or because of technical and experimenter errors. Because these exclusions, there was a slight imbalance regarding the between-subjects factor conditioned image category, such that for 41 participants tools served as CS⁺ category, whereas animals served as CS⁺ category for only 39 participants. We examined whether this slight imbalance affected our results by randomly excluding two participants with tools as the CS⁺ category from our analysis (10 permutations). Because the pattern of results remained unchanged, the following analyses were performed for the full sample of 80 participants. Again, the target sample size was determined using an a priori power analysis in G*Power 3 (Faul et al., 2007) with the aim to considerably increase the statistical power compared with Experiment 1 and thereby minimize the chance of null findings due to an insufficient sample size. As in Experiment 1, we assumed $d_z = .45$ as a point-estimate for the previously reported category-specific retroactive memory effect (Dunsmoor et al., 2015). A two-tailed paired t -test with $\alpha = .05$ required at least 82 participants to achieve a statistical power of .98. Exclusion criteria were identical to Experiment 1. None of the participants from Experiment 1 participated in this experiment. As before, participants received a monetary compensation of 20€. The study protocol was approved by the ethics committee of the Faculty of Psychology and Human Movement Science at the Universität Hamburg.

Materials

For this experiment, we used the stimulus set from the study by Dunsmoor et al. (2015), consisting of 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. As in Experiment 1, photographs were of neutral valence and each tool and animal represented a unique exemplar of its respective category. The stimulus set was randomly divided

into learning items and lures per participant and learning items were allocated to the three encoding phases in the same manner as in Experiment 1.

Procedure

The procedure in this experiment was largely identical to Experiment 1, except for some minor changes to achieve consistency with the original study (Dunsmoor et al., 2015). More precisely, we changed the location of the stimulation electrode from the right lower leg to the right wrist. Because this area tends to be more sensitive to electric stimulation, we also reduced the initial intensity in the procedure for determining the pain threshold from 20V in Experiment 1 to 10V in this experiment. Furthermore, we replaced the two-step forced-choice decision in the surprise recognition test on the second experimental day with a task assessing both “old” versus “new” decisions and certainty in a single step as reported in the original study (Dunsmoor et al., 2015). More precisely, for each stimulus in the recognition test, participants performed only a single button press with the four possible options that the currently presented item was either *definitely old* (German: *sicher alt*), *maybe old* (*eher alt*), *maybe new* (*eher neu*), or *definitely new* (*sicher neu*) by pressing the “1”, “2”, “3”, or “4” button on the keyboard, respectively. Additionally, the 5-s time limit per response that was used in Experiment 1 was removed. In sum, Experiment 2 used the same experimental procedure and stimuli as the study by Dunsmoor et al. (2015) but a significantly larger sample size.

Data Analysis

The statistical analysis was identical to Experiment 1 and the statistical analysis of Dunsmoor et al. (2015), complemented by analyses based on signal detection theory parameters and a Bayesian analysis.

Results and Discussion

Successful Fear Conditioning

An analysis of skin conductance data indicated that our procedure successfully induced conditioned fear for CS⁺ items. Specifically, during Pavlovian conditioning participants showed greater anticipatory SCRs to items from the CS⁺ category compared with items from the CS⁻ category (TTP: $t[79] = 4.75$, $p < .001$, $d_{av} = .48$; CDA: $t[79] = 4.32$, $p < .001$, $d_{av} = .35$; Figure 2).

Anticipation of the Recognition Test

As in Experiment 1, participants rated how surprised they were by the recognition test on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). On average, they indicated that they were moderately surprised ($M = 3.11$, $SD = 1.12$). Nine participants chose the *not surprised at all* option. Because excluding these participants did not affect the pattern of results, they were still included in the following analyses.

Overall Memory Performance

Participants performed overall very well in the surprise recognition test (see Table 2) with an average hit rate for items from all

Table 2*Retrieval Memory Results in Experiment 2, Mean Proportion of Responses by Certainty*

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.421	0.224	0.216	0.138	0.388	0.245	0.245	0.123
Conditioning	0.454	0.227	0.205	0.114	0.331	0.265	0.267	0.137
Postconditioning	0.352	0.219	0.258	0.171	0.298	0.230	0.307	0.165
New	0.078	0.169	0.368	0.386	0.083	0.174	0.381	0.362

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

three encoding phases of 60.9% ($SD = .14$) and an average false alarm rate of 25.2% ($SD = .10$).

No Evidence for Category-Specific Retroactive Memory Enhancement

We performed a repeated-measures ANOVA on corrected recognition scores to identify factors affecting memory formation over the task. As in Experiment 1, the recognition performance generally differed between phases, $F(2, 158) = 26.8, p < .001, \eta^2_G = .06$. The recognition performance was also generally different between CS⁺ and CS⁻ items, $F(1, 79) = 17.0, p < .001, \eta^2_G = .03$. Finally, there was a significant interaction between the encoding phase and the item category, indicating that the effect of membership of an item to the CS⁺ versus CS⁻ category differed between the encoding phases, $F(2, 158) = 6.66, p = .002, \eta^2_G = .007$. To further qualify these results, we performed paired t -tests comparing corrected recognition scores for items belonging to the CS⁺ versus CS⁻ category separately per phase. For photographs presented during Pavlovian conditioning, we obtained an enhanced memory for CS⁺ items compared with CS⁻ items, $t(79) = 4.89, p < .001, d_{av} = .53$ (Figure 3, upper right panel). This memory benefit for items belonging to the CS⁺ category carried over to the postconditioning phase, even though shock leads were removed beforehand, as indicated by improved corrected recognition scores, $t(79) = 3.32, p = .001, d_{av} = .33$. Most importantly, despite the high statistical power in this replication study, corrected recognition scores provided no evidence for a retroactive memory enhancing effect for items from the CS⁺ category over items from the CS⁻ category presented before the Pavlovian conditioning phase, $t(79) = 1.28, p = .20, d_{av} = .14$. As in Experiment 1, we also tested for preconditioning items the previously reported linear relationship between their temporal distance and the size of the category-specific retroactive memory effect using the same procedure as in Experiment 1. Contrary to this hypothesis, a repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ items as the dependent variable and the time bin as a within-subject factor showed no significant effect of time bins, $F(2, 158) = 1.08, p = .34, \eta^2_G = .007$. This finding shows that the relative time of encoding of an item within the preconditioning phase had no effect on the proposed category-specific retroactive memory enhancement.

Complementary Analyses

As in Experiment 1, we performed additional Bayesian paired t -tests to quantify relative evidence for the null versus the alternative hypotheses regarding effects of fear conditioning on memory

formation in the different encoding phases. These confirmed previous findings by showing substantial evidence for the alternative hypothesis of an enhanced memory for CS⁺ versus CS⁻ items that were encoded during Pavlovian conditioning, $BF_{10} = 6223$. Similarly, a Bayesian analysis indicating substantial support for the hypothesis of a memory advantage for CS⁺ over CS⁻ items that were encoded after fear conditioning, $BF_{10} = 36.38$. Most critically, for the comparison of CS⁺ and CS⁻ items encoded before fear conditioning, results from the Bayesian analysis spoke against category-specific retroactive memory enhancement, although unlike in Experiment 1, evidence for the null hypothesis was only anecdotal, $BF_{10} = .48$.

Although Dunsmoor et al. (2015) based their critical analyses on corrected recognition scores, we also aimed to replicate their findings using memory sensitivity scores (d'). As expected, results were very similar to those based on corrected recognition scores. Specifically, we found improved memory for CS⁺ items encoded during fear conditioning compared with CS⁻ items from the same phase, $t(79) = 4.31, p < .001, d_{av} = .49$. This improved memory sensitivity for CS⁺ items also carried over to the postconditioning phase, $t(79) = 2.58, p = .012, d_{av} = .27$. As for corrected recognition scores, there was no evidence for category-specific retroactive memory enhancement in memory sensitivity scores (d'), $t(79) = 1.28, p = .20, d_{av} = .14$. Again, analyses based on generalized linear mixed-effect models showing the same pattern of results are included in the online supplemental materials.

To identify possible factors contributing to the lack of category-specific retroactive memory enhancements in this experiment, we again performed additional analysis beyond those reported by Dunsmoor et al. (2015). Although participants were slightly less surprised by the recognition test than in Dunsmoor et al. (2015), there again was no significant correlation between the anticipation of the memory test and recognition performance, Spearman's $r_s = .15, p = .19$. Further, individual memory performance did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS⁺ minus mean SCRs to CS⁻), TTP: Spearman's $r_s = -.08, p = .50$; CDA: Spearman's $r_s = -.01, p = .95$.

Experiment 3: Testing the Effect of an Increased Interval Between Preconditioning and Fear-Conditioning on Category-Specific Retroactive Memory Enhancement

Thus far, we were unable to find any evidence for category-specific retroactive memory enhancements for weakly encoded stimuli belonging to a category that was later associated with the occurrence of shocks in a fear conditioning paradigm. Experiment

2 showed the absence of the category-specific retroactive memory effect could not be attributed to the stimulus set, nor to small deviations in the procedure. Additionally, as Experiment 2 had high statistical power, it is highly unlikely that the absence of the category-specific retroactive memory effect was due to an insufficient sample size.

This third replication attempt was designed to investigate one aspect that moderated the size of the category-specific retroactive memory effect in the original study, namely the interval between the encoding during the preconditioning phase and the subsequent fear-conditioning (Dunsmoor et al., 2015). It had been shown that items from the preconditioning phase that were presented the longest before fear-conditioning showed the strongest category-specific retroactive memory enhancement, whereas this effect seemed to diminish the closer items were encoded relative to the conditioning phase (Dunsmoor et al., 2015). This finding was in line with previous work on nonspecific behavioral tagging in rodents, which suggests that there might be a minimal interval between the weak encoding (setting the tag) and associated arousing event to enable retroactive memory enhancement (de Carvalho Myskiw et al., 2013; Moncada et al., 2015). However, it is important to note that, in these experiments investigating unspecific behavioral tagging, the interval between initial weak encoding and the subsequent memory promoting event was relatively long, typically more than one hour. Dunsmoor et al. (2015), on the other hand, observed positive effects of the temporal distance of a preconditioning item to conditioning procedure at a much shorter time scale, that is, only minutes. Here, we built the encoding-conditioning interval on the finding from Dunsmoor et al. (2015) to maximize the chances of detecting a category-specific retroactive memory enhancement. Therefore, in this third experiment, we increased the interval between the preconditioning and the fear conditioning phase by 10 minutes to investigate whether this change could produce the category-specific retroactive memory enhancement that was not detectable in Experiments 1 and 2. Apart from this single aspect, we retained both the procedure and the high statistical power from Experiment 2. Therefore, the 200-ms variation in the CS-UCS timing compared with Dunsmoor et al. (2015) was also retained in this experiment. Again, we did not control for stimulus typicality in this experiment as this aspect was not mentioned in Dunsmoor et al. (2015) and brought forward to us only at a later stage.

Method

Participants

Eighty-four healthy volunteers (59 women) between 18 and 33 years of age participated in this experiment ($M = 25.11$, $SD = 3.57$). Six participants had to be excluded from the analysis, either because they did not return for the memory test on the second experimental day or because of technical and experimenter errors. As in Experiment 2, these exclusions led to a slight imbalance regarding the between-subjects factor conditioned image category, such that for 40 participants tools served as CS⁺ category, whereas animals served as CS⁺ category for only 38 participants. Again, we examined whether this imbalance affected our results by randomly excluding two participants with tools as the CS⁺ category from our analysis (10 permutations). Because the pattern of results

remained unchanged, the following analyses were performed for the full sample of 78 participants. The target sample size was calculated using an a priori power analysis with identical parameters as in Experiment 2. Exclusion criteria were identical as in Experiments 1 and 2. None of the participants had previously participated in Experiment 1 nor in Experiment 2. Again, participants received a monetary compensation of 20€. The study protocol was approved by the ethics committee of the Faculty of Human Movement Science at the Universität Hamburg.

Materials

We used the same stimulus set as in Experiment 2, corresponding to the material used by Dunsmoor et al. (2015) and consisting of 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. As in Experiment 1 and Experiment 2, per participant, half of the stimuli from each category were randomly selected as learning items, whereas the remaining half served as lures. The learning items were allocated to each of the three encoding phases in the same manner as in Experiment 1 and Experiment 2. Furthermore, the assignment of photographs of tools and animals as CS⁺ and CS⁻, respectively, was counterbalanced across participants.

Procedure

The only difference compared with the procedure in Experiment 2 was the extension of the interval between the preconditioning phase and the subsequent fear conditioning phase. This change was based on the finding that the category-specific retroactive memory effect was positively correlated with the temporal distance between the encoding of an item and the following fear conditioning procedure in the original study (Dunsmoor et al., 2015) as well as evidence from studies in rodents (de Carvalho Myskiw et al., 2013; Moncada et al., 2015). In this experiment, when participants finished the preconditioning phase—unlike in Experiment 1 and 2—we did not immediately attach the electrodes. Instead, participants were first presented the following series of questionnaires: The State-Trait Anxiety Inventory (Spielberger, 1983), a multidimensional mood questionnaire (Steyer et al., 1997), a chronic stress questionnaire (Schulz et al., 2004), the Beck Depression Inventory (Beck et al., 1996), the Social Interaction Anxiety Scale (Mattick & Clarke, 1998), and the Positive and Negative Affect Schedule (Watson et al., 1988). After participants had worked on these questionnaires for exactly 10 minutes, they were interrupted and told that the remaining questions could be finished at a later stage. In fact, questionnaires were only added to keep participants occupied during the prolonged interval before the fear-conditioning. For this reason, we also chose a greater number of questionnaires than could usually be completed within 10 minutes, so that no participant would finish them earlier. Afterward, the experiment continued in the same manner as described for Experiment 2, by first attaching electrodes and determining the pain threshold (taking an additional approximately 10 min), before the start of the fear-conditioning phase, followed by the postconditioning phase and the 24-hr-delayed recognition test.

Data Analysis

The statistical analysis was identical to Experiments 1 and 2.

Results and Discussion

Successful Fear Conditioning

As in both previous experiments, an analysis of skin conductance data confirmed that our procedure was successful in inducing conditioned fear for CS⁺ items. SCR data for one additional participant were missing due to experimenter error. For the remaining sample of 77 participants, during Pavlovian fear conditioning anticipatory SCRs to items from the CS⁺ category were significantly higher compared with items from the CS⁻ category (TTP: $t[76] = 3.97, p < .001, d_{av} = .39$; CDA: $t[76] = 4.10, p < .001, d_{av} = .30$; Figure 2).

Anticipation of the Recognition Test

As in the previous experiments, participants rated how surprised they were by the recognition test on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). Data from one participant were missing as a result of experimenter error. On average, the remaining 77 participants indicated moderate levels of surprise ($M = 2.95, SD = .97$). Three participants chose the *not surprised at all* option. Because excluding these participants did not affect the pattern of results, they were still included in the following analyses.

Overall Memory Performance

As in Experiments 1 and 2, participants performed overall very well in the surprise recognition test (see Table 3) with an average hit rate for items from all three encoding phases of 62.7% ($SD = .16$) and an average false alarm rate of 23.1% ($SD = .10$).

No Evidence for Category-Specific Retroactive Memory Enhancement

To analyze factors affecting memory performance for the different phases of the task, we ran a repeated-measures ANOVA on corrected recognition scores. As in the two previous experiments, corrected recognition scores generally differed between phases, $F(2, 154) = 35.72, p < .001, \eta^2_G = .08$. Corrected recognition scores were also generally different between items from the CS⁺ and CS⁻ categories, $F(1, 77) = 17.8, p < .001, \eta^2_G = .03$. Finally, this effect of item category membership differed between the encoding phases, as indicated by a significant interaction between the encoding phase and the item category, $F(2, 154) = 22.12, p < .001, \eta^2_G = .03$. To further qualify these results, we performed paired t -tests to compare corrected recognition scores for items belonging to the CS⁺ versus CS⁻ category separately per

encoding phase. As in Experiments 1 and 2, these showed an enhanced memory performance for CS⁺ items encoded during Pavlovian conditioning compared with CS⁻ items encoded in the same phase, $t(77) = 6.60, p < .001, d_{av} = .77$ (Figure 3, lower left panel). As in Experiment 1, there also was a trend toward improved recognition memory for CS⁺ items encoded after Pavlovian conditioning compared with CS⁻ items encoded in the same phase, although unlike in Experiment 2, this trend was not statistically significant, $t(77) = 1.90, p = .061, d_{av} = .19$. Above all, despite the increase in the interval between encoding and Pavlovian conditioning, we obtained no evidence for a retroactive enhancement of memory for CS⁺ items compared with CS⁻ items encoded before Pavlovian conditioning in corrected recognition scores, $t(77) = .17, p = .86, d_{av} = .02$. Notably, even at descriptive level, the memory difference between CS⁺ and CS⁻ items encoded before fear conditioning was negligible. We again tested the possibility of a previously suggested linear trend between preconditioning items' temporal distance to the conditioning phase and the size of retroactive memory enhancement. As in Experiment 1 and 2, a repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ items as the dependent variable and the time bin as a within-subject factor showed no significant effect of time bins, $F(2, 154) = .17, p = .85, \eta^2_G = .001$. This indicates that the relative time of encoding of an item within the preconditioning phase had no effect on putative category-specific retroactive memory enhancement.

Complementary Analyses

As for both previous experiments, to quantify relative evidence for the null versus the alternative hypothesis of memory enhancements through fear learning in each of the three encoding phases, we conducted complementary Bayesian paired t -test. As before, these indicated substantial evidence for enhanced memory formation of CS⁺ relative to CS⁻ items that were encoding during fear conditioning, $BF_{10} = 4727037$. A corresponding Bayesian analysis for the postconditioning phase also favored the alternative hypothesis of enhanced memory for CS⁺ items, although evidence was only anecdotal, $BF_{10} = 1.34$. As in both previous experiments, the Bayesian analysis favored the null hypothesis rejecting the notion of category-specific retroactive memory enhancements and as in Experiment 1, evidence for the null hypothesis was substantial, $BF_{10} = .14$.

Whereas previous analyses focused on corrected recognition scores to closely replicate Dunsmoor et al. (2015), we also performed parallel analyses on memory sensitivity (d'). These yielded

Table 3

Retrieval Memory Results in Experiment 3, Mean Proportion of Responses by Certainty

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.407	0.242	0.225	0.126	0.420	0.235	0.221	0.125
Conditioning	0.481	0.256	0.166	0.097	0.352	0.255	0.261	0.132
Postconditioning	0.341	0.227	0.256	0.175	0.310	0.235	0.272	0.183
New	0.070	0.156	0.362	0.412	0.066	0.169	0.372	0.393

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

the same pattern of results as analyses based on corrected recognition. Specifically, the analysis based on d' confirmed previous findings of enhanced memory for CS⁺ items encoded during fear conditioning, $t(77) = 5.95, p < .001, d_{av} = .69$. For items encoded after conditioning, a similar, but nonsignificant trend was obtained, $t(77) = 1.72, p = .089, d_{av} = .20$. Above all, memory sensitivity scores (d') indicated no evidence for any category-specific retroactive memory enhancement for CS⁺ items from the preconditioning phase, $t(77) = .45, p = .65, d_{av} = .05$. Again, analyses based on generalized linear mixed-effect models showing the same pattern of results are presented in the online supplemental materials.

As in all previous experiments, there was no significant correlation between the anticipation of the memory test and recognition performance, Spearman's $r_s = -.04, p = .74$. Again, individual memory performance did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS⁺ minus mean SCRs to CS⁻), TTP: Spearman's $r_s = -.04, p = .76$; CDA: Spearman's $r_s = .11, p = .36$.

Experiment 4: Replicating Category-Specific Retroactive Memory Enhancements After Adopting Original UCS Timings and Balanced Stimulus Typicality Across Phases

The three previous experiments aimed to replicate findings of category-specific retroactive memory enhancements for stimuli from a category that was later associated with shock occurrences in a fear conditioning procedure. Compared with Experiment 1, Experiments 2 and 3 adopted additional details from the original procedure, namely the original stimulus set and the same format for the recognition tests. Based on comments from authors of the original study and reviewers, two additional deviations from Dunsmoor et al. (2015) were identified that applied to Experiments 1 to 3. First, in the original study, shocks coterminated with the stimulus presentation during fear-conditioning, whereas in Experiments 1–3 shock onsets were administered exactly at the point of stimulus offsets. Although this only leads to a 200-ms relative difference between studies (i.e., one shock length), it implies that stimuli were still present when shocks occurred in Dunsmoor et al. (2015), whereas in our Experiments 1 to 3 shocks followed immediately after stimulus offset. We address this issue here by using exactly the same shock timings that were used in Dunsmoor et al. (2015).

Second, Dunsmoor et al. (2015) controlled typicality and superordinate categories of stimuli, such that these were balanced across each of the three encoding phases and the recognition test. Unfortunately, they did not report on this in their study and we only learned about this aspect through the peer review process for this article. This contrasts with our procedure in Experiments 1–3, in which the set of stimuli was randomly distributed to each encoding phase. Therefore, our allocation of stimuli was unique per participant. We aimed to investigate whether this procedural difference might explain the lack of category-specific retroactive memory enhancements in our previous experiments. This fourth replication attempt had been preregistered and prereviewed before the beginning of data collection. The preregistration can be found at <https://osf.io/9hzmk>.

Method

Participants

Eighty-four healthy men and women between 18 and 34 years of age participated in this experiment ($M = 25.17, SD = 4.26$). Data from 13 participants had to be excluded because of an error in an early version of the experimental software that would in some trials incorrectly administer shocks to CS⁻ items. Because these exclusions might negatively affect the statistical power, we decided to recruit replacements for these 13 participants. One additional participant had to be excluded due to technical problems on the first experimental day. Therefore, the final sample included in the memory analysis consisted of 83 participants.

As in Experiment 2, there was a slight imbalance regarding the between-subjects factor conditioned image category, such that for 42 participants tools served as CS⁺ category, whereas animals served as CS⁺ category for 41 participants. Again, we examined whether this imbalance affected our results by randomly excluding two participants with tools as the CS⁺ category from our analysis (10 permutations). Because the pattern of results remained unchanged, the following analyses were performed for the full sample of 83 participants. The target sample size was calculated using an a priori power analysis with identical parameters as in Experiment 2 and 3. Exclusion criteria were identical as in Experiment 1, 2, and 3. None of the participants had previously participated in any of the other Experiments. Participants received a monetary compensation of 30€. The study protocol was approved by the ethics committee of the Faculty of Psychology and Human Movement Science at the Universität Hamburg.

Materials

We used the same stimulus set as in Experiments 2 and 3, corresponding to the material used by Dunsmoor et al. (2015) and consisting of 180 color photographs of animals and 180 color photographs of tools isolated on white backgrounds. Unlike in Experiments 1–3, stimuli were not randomly allocated as learning items or distractors. Instead, we received the fixed stimulus allocation table that was used in Dunsmoor et al. (2015; Joseph E. Dunsmoor, personal communication, August 13, 2018) which was intended to match each of the encoding phases in terms of stimulus typicality and superordinate categories. In an online pilot-study, we recruited an additional independent sample of 41 participants (31 women, 10 men; aged 19–42 years; $M = 26.55, SD = 6.08$) who rated the typicality of all 360 stimuli. In random succession, they saw all 360 photographs (180 animals and 180 tools) and rated how typical each photograph was for its respective category on a scale from 1 (*very untypical*) to 10 (*very typical*). Ratings were self-paced (i.e., there was no time limit per photograph).

Results showed that simply adopting the allocation table from Dunsmoor et al. (2015) would lead to significant differences in typicality between encoding phases. After swapping four photographs of tools and two photographs of animals between sets, we obtained even typicality per category across sets (Figure 7 in the online supplemental materials). This procedure ensured that we had comparable typicality ratings per category across sets on the one hand, while sticking as closely as possible to the stimulus allocation used in Dunsmoor et al. (2015). The resulting stimulus sets consisted of three encoding sets with 30

photographs of animals and 30 photographs of tools each and a fourth set consisting of 90 photographs of animals and 90 photographs of tools that were used as lures in the recognition test. For each participant, the allocation of encoding sets to encoding phases was randomized. Further, as in all previous experiments, the assignment of photographs of tools and animals as CS⁺ and CS⁻, respectively, was counterbalanced across participants.

Procedure

The procedure in this experiment was identical to Experiment 2, except that we changed the timing of the shock (i.e., the UCS) during fear conditioning to be identical to Dunsmoor et al. (2015). During fear conditioning, a 200-ms shock occurred (under the same contingencies as in the previous two experiments), presented 4.3 s after stimulus onset and thus coterminated with the stimulus.

Data Analysis

The statistical analysis was identical to Experiment 1 and 3, and the statistical analysis of Dunsmoor et al. (2015), complemented by additional exploratory and a Bayesian analysis.

Results and Discussion

Successful Fear Conditioning

We again performed an analysis of skin conductance data to confirm the success of our fear-conditioning procedure. SCR data for twelve participants were not usable because of equipment misconfiguration. For the remaining sample of 71 participants, the TTP analysis (i.e., a more traditional approach of analyzing SCR data also utilized by Dunsmoor et al., 2015) indicated successful Pavlovian fear conditioning as expressed in increased anticipatory SCRs to CS⁺ items compared with CS⁻ items, $t(70) = 4.59, p < .001, d_{av} = .48$; for the CDA there was no significant effect, $t(70) = 1.24, p = .22, d_{av} = .08$ (see Figure 2).

Anticipation of the Recognition test

Again, participants rated how surprised they were by the recognition test on a scale ranging from 1 (*very surprised*) to 5 (*not surprised at all*). On average, they indicated that they were moderately and slightly more surprised than in Experiments 1–3 ($M = 2.82, SD = 1.14$). Eight participants reported being *not surprised at all*. Because excluding these participants did not affect the pattern of results, they were still included in the following analyses.

Overall Memory Performance

As in all three previous experiments, participants performed overall well in the recognition test (see Table 4). The average hit rate for items from all three encoding phases was 62.8% ($SD = .14$), with an average false alarm rate of 26.2% ($SD = .10$).

No Evidence for Category-Specific Retroactive Memory Enhancement

As in all three previous experiments, we ran a repeated-measures ANOVA on corrected recognition scores to analyze factors affecting memory performance for the different phases of the task. For the factor phase, Mauchly's test indicated that the sphericity assumption was violated, $W = .89, p = .008$. Hence, results for the factor phase are reported after applying a Greenhouse-Geisser correction. Corrected recognition scores generally differed between phases, $F(1.80, 147.30) = 18.19, p < .001, \eta^2_G = .036$. They were also generally different between items from the CS⁺ and CS⁻ categories, $F(1, 82) = 17.61, p < .001, \eta^2_G = .022$. Finally, this effect of item category membership differed between the encoding phases, as indicated by a significant interaction between the encoding phase and the item category, $F(2, 164) = 10.19, p < .001, \eta^2_G = .012$. These results were further qualified by paired *t*-tests comparing corrected recognition scores for items belonging to the CS⁺ versus CS⁻ category separately per encoding phase. As in all three previous experiments, these showed an enhanced memory performance for CS⁺ items encoded during Pavlovian conditioning compared with CS⁻ items encoded in the same phase, $t(82) = 5.75, p < .001, d_{av} = .58$ (Figure 3, lower right panel). As in Experiments 1 and 3, there also was a (nonsignificant) trend toward improved recognition memory for CS⁺ items encoded after Pavlovian conditioning compared with CS⁻ items encoded in the same phase, $t(82) = 1.71, p = .091, d_{av} = .14$. Most importantly, even after additionally adopting the exact UCS-CS timings from Dunsmoor et al. (2015) and controlling for stimulus typicality across phases, we obtained no evidence for a category-specific retroactive enhancement of memory for CS⁺ items compared with CS⁻ items encoded before Pavlovian conditioning in corrected recognition scores, $t(82) = 1.37, p = .18, d_{av} = .14$. We again tested the possibility of a previously suggested linear trend between pre-conditioning items' temporal distance to the conditioning phase and the size of retroactive memory enhancement. As in all three previous experiments, a repeated-measures ANOVA with the corrected recognition advantage for CS⁺ items compared with CS⁻ items as the dependent variable and the time bin as a within-subject

Table 4

Retrieval Memory Results in Experiment 4, Mean Proportion of Responses by Certainty

Measure	CS ⁺				CS ⁻			
	DO	MO	MN	DN	DO	MO	MN	DN
Preconditioning	0.412	0.243	0.226	0.120	0.375	0.243	0.243	0.139
Conditioning	0.497	0.229	0.186	0.088	0.348	0.257	0.251	0.144
Postconditioning	0.364	0.240	0.250	0.146	0.332	0.231	0.275	0.162
New	0.090	0.179	0.393	0.339	0.084	0.171	0.377	0.367

Note. DO = definitely old; MO = maybe old; MN = maybe new; DN = definitely new.

factor showed no significant effect of time bins, $F(2, 164) = .68, p = .51, \eta^2_G = .004$. Thus, we could not find any effect of the relative time of encoding of an item within the preconditioning phase on the size of the putative category-specific retroactive memory enhancement.

Complementary Analyses

We conducted complementary Bayesian paired t -tests to quantify relative evidence for the null versus the alternative hypothesis of memory enhancements through fear learning in each of the three encoding phases. As before, these indicated substantial evidence for enhanced memory formation of CS^+ relative to CS^- items that were encoded during fear conditioning, $BF_{10} = 176943$. A corresponding Bayesian analysis for the postconditioning phase slightly favored the null hypothesis of no memory advantage for CS^+ items, although evidence was only anecdotal, $BF_{10} = .93$. As in the three previous experiments, the Bayesian analysis favored the null hypothesis rejecting the notion of category-specific retroactive memory enhancements, although evidence for the null hypothesis was only anecdotal, $BF_{10} = .54$.

In addition to analyses focusing on corrected recognition scores to closely replicate Dunsmoor et al. (2015), we also performed parallel analysis on memory sensitivity (d'). These yielded the same pattern of results as analyses based on corrected recognition. Specifically, analysis based on d' confirmed previous findings of enhanced memory for CS^+ items encoded during fear conditioning, $t(82) = 5.37, p < .001, d_{av} = .57$. For items encoded after conditioning, there were no significant differences in d' between CS^+ and CS^- items, $t(82) = 1.36, p = .18, d_{av} = .14$. Most importantly, memory sensitivity scores (d') indicated no evidence for any category-specific retroactive memory enhancement for CS^+ items from the preconditioning phase, $t(82) = 1.12, p = .27, d_{av} = .11$. Analyses based on generalized linear mixed-effect models again showed the same pattern of results and are included in the online supplemental materials.

As in all three previous experiments, we found no significant correlation between the anticipation of the memory test and recognition performance, Spearman's $r_s = .13, p = .25$. Again, individual memory performance did not correlate with induced arousal during fear conditioning (measured through mean SCRs to CS^+ minus mean SCRs to CS^-), TTP: Spearman's $r_s = .03, p = .78$; CDA: Spearman's $r_s = .01, p = .91$.

Analyses Focusing on High Confidence Hits

Although Dunsmoor et al. (2015) collapsed responses from the surprise recognition test across confidence, we also reanalyzed our data by focusing only high confidence hits using the same paired t -tests on corrected recognition scores as reported in the article, complemented by their Bayesian counterparts to quantify the relative evidence for the null versus the alternative hypothesis of category-specific retroactive memory enhancement. For this analysis, we used a definition of high confidence hits that treated any *rather old* responses like *new* responses (Dunsmoor et al., 2012; Keller & Dunsmoor, 2020). Therefore, only *definitely old* responses could result in either a hit or a false alarm, whereas *rather old* responses were always scored as either misses or correct rejections depending on the actual status of the item. Note that focusing on high confidence hits therefore implies a different scoring of existing

responses, while no trials were omitted from the recognition analysis.

For Experiment 1, these analyses focusing on high confidence hits showed no evidence for category-specific retroactive memory enhancement on corrected recognition scores, $t(43) = 1.05, p = .30, d_{av} = .18, BF_{10} = .46$, nor on memory sensitivity (d'), $t(43) = .58, p = .56, d_{av} = .11, BF_{10} = .27$. In contrast to previously reported results after collapsing across memory confidence, for Experiment 2 an analysis focusing on high confidence hits showed the proposed category-specific retroactive enhancement on corrected recognition scores, $t(79) = 2.31, p = .024, d_{av} = .22, BF_{10} = 2.95$, but not on memory sensitivity (d'), $t(79) = 1.30, p = .20, d_{av} = .14, BF_{10} = .50$. For Experiment 3, results were again consistent with those obtained from the analysis of recognition collapsed over confidence and showed no evidence for category-specific retroactive memory enhancement in neither corrected recognition, $t(77) = .98, p = .33, d_{av} = .10, BF_{10} = .07$, nor in memory sensitivity (d'), $t(77) = .67, p = .50, d_{av} = .08, BF_{10} = .08$. Likewise, analyses on high confidence memory for Experiment 4 provided again neither evidence for the category-specific retroactive memory effect on corrected recognition, $t(82) = 1.72, p = .088, d_{av} = .19, BF_{10} = .95$, nor on memory sensitivity (d'), $t(82) = .20, p = .84, d_{av} = .08, BF_{10} = .14$.

Response Bias Analysis

A possible explanation for the inconsistent findings between corrected recognition and d' regarding high-confidence memory in Experiment 2 (and at trend level in Experiment 4) could be that findings appearing to show category-specific retroactive memory enhancement in corrected recognition for high confidence hits instead reflect a response bias toward more liberal *old* responses for items from the CS^+ category without any actual difference in memory sensitivity between items from the CS^+ versus CS^- category that were encoded during preconditioning (Dougal & Rotello, 2007; Rotello et al., 2008). We investigated this possibility by calculating response bias scores c based on signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002) and comparing them for items from the CS^+ versus CS^- category separately for each experiment and encoding phase. Detailed results from this analysis are provided in the online supplemental materials (Tables 1–4). In short, we found that participants overall showed a bias to classify items from the CS^+ category (over item from the CS^- category) as *old* when these were encoded during fear-conditioning. For the critical influence of response biases on findings of category-specific retroactive memory enhancement, Experiments 2 and 4 were the most interesting, because these were the only two experiments in which an analysis of high confidence corrected recognition provided some evidence for this effect (although only at trend level in Experiment 4). In Experiment 2, participants descriptively, but nonsignificantly, showed a slightly increased response bias in the high-confidence hit rate toward items from the CS^+ category that were encoded before fear-conditioning, $t(79) = 1.25, p = .21, d_{av} = .14$. For Experiment 4, this effect was significant, indicating that participants more liberally classified preconditioning items from the CS^+ category (compared with items from the CS^- category) as 'old', regardless of their actual status, $t(82) = 2.06, p = .042, d_{av} = .22$. For Experiments 1 and 3, there was no

significant difference in response bias for high confidence memory of items from the preconditioning phase (both p s > .36).

Pooled Analysis Across All Experiments

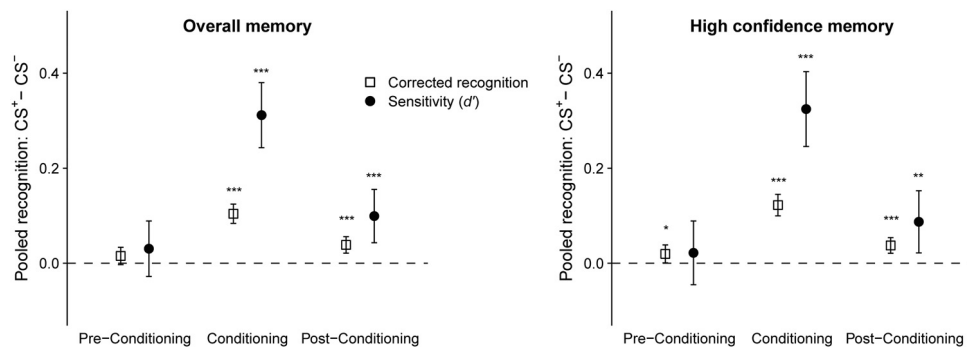
Experiments 1–4 were designed to replicate the previously reported finding of a category-specific retroactive memory enhancement through subsequent electric shocks (Dunsmoor et al., 2015). Although we varied certain aspects regarding stimuli and the procedure between these experiments, the procedure of Experiments 1–4 was conceptually very similar. To summarize findings from the four studies in a combined statistical model, we pooled data across experiments and fit separate linear mixed-effects models for both corrected recognition and memory sensitivity (d') using the *R* library *lme4* (Bates et al., 2015) for each of the three conditioning phases. In each of these three models, the conditioned category an item belonged to (binary coding: 0 for CS^- ; 1 for CS^+) was treated as a fixed effect (Hedges & Vevea, 1998) to explain corrected recognition scores. Additionally, random intercepts were fitted both per subject and per experiment, to account for differences in memory performance between participants and procedural differences between experiments, respectively. To further qualify the results reported in the previous section, we also ran separate models for memory collapsed over confidence levels and high confidence memory only, respectively.

Even after pooling data collapsed across confidence (parallel to Dunsmoor et al., 2015) from Experiments 1–4 with a total of 285 unique participants, there was no evidence for category-specific retroactive memory enhancement, as indicated by neither a significant effect of conditioning category membership of preconditioning items on corrected recognition scores, $\beta = .015$, 95% CI $[-.003, .033]$, $t(284) = 1.66$, $p = .097$, $BF_{10} = .17$, nor a significant effect of

conditioning category membership of preconditioning items on d' , $\beta = .03$, 95% CI $[-.028, .090]$, $t(284) = 1.03$, $p = .30$, $BF_{10} = .12$ (see Figure 4). Bayes factors obtained in both cases indicated substantial evidence for the null hypothesis speaking against category-specific retroactive memory enhancement. Although there was no evidence for a selective retroactive memory effect, the pooled analysis revealed that CS^+ photographs encoded during fear-conditioning were significantly better recognized than CS^- photographs, as reflected in both corrected recognition, $\beta = .10$, 95% CI $[.084, .124]$, $t(284) = 10.07$, $p < .001$, $BF_{10} = 544211770$, and in d' , $\beta = .31$, 95% CI $[.24, .38]$, $t(284) = 8.93$, $p < .001$, $BF_{10} = 13621099$. Finally, the pooled analysis collapsed over confidence confirmed results from Experiment 2 that this memory enhancement for CS^+ photographs relative to CS^- photographs carried over to the postconditioning phase, even though these items were never directly paired with the UCS. This was reflected in both corrected recognition, $\beta = .039$, 95% CI $[.021, .056]$, $t(284) = 4.37$, $p < .001$, $BF_{10} = 3.75$, and in d' , $\beta = .099$, 95% CI $[.043, .155]$, $t(284) = 3.47$, $p < .001$, $BF_{10} = 1.30$.

Next, we fitted the same pooled models for the exploratory analyses on high confidence recognition memory. This pooled analysis showed a significant category-specific retroactive memory enhancement in corrected recognition scores, although a Bayesian analysis of this model indicated substantial evidence for the null hypothesis, thus speaking against a category-specific retroactive memory enhancement, $\beta = .20$, 95% CI $[.0008, .039]$, $t(284) = 2.05$, $p = .042$, $BF_{10} = .22$. The same analysis on d' indicated no significant category-specific retroactive memory enhancement and substantial evidence for the null hypothesis, $\beta = .02$, 95% CI $[-.045, .089]$, $t(284) = .64$, $p = .52$, $BF_{10} = .10$. As for the analysis collapsed across confidence levels, the analysis focusing on high confidence recognition memory revealed clear evidence that CS^+ photographs encoded during fear-conditioning were significantly better recognized than CS^- photographs in both corrected

Figure 4
Results of a Pooled Analysis Across Data From Experiments 1–4 Using Linear-Mixed Effect Models



Note. The left panel shows the advantage in CS^+ over CS^- recognition performance after collapsing across memory confidence, whereas the right panel shows parallel results for high confidence memory only. In both analyses, items from the CS^+ category encoded during both Pavlovian fear conditioning and after fear conditioning were recognized significantly better compared with items from the CS^- category encoded within their respective phase, as reflected in both corrected recognition scores and d' from signal detection theory. For items encoded before the conditioning phase, there was a retroactive enhancement for items from the CS^+ vs. CS^- category when only high confidence memory was analyzed, but not when overall memory performance was analyzed. Moreover, this effect for high confidence memory was only present for corrected recognition, but not for d' . Error bars represent 95% confidence intervals.

* $p < .05$. ** $p < .01$. *** $p < .001$.

recognition scores, $\beta = .12$, 95% CI [.10, .14], $t(284) = 10.60$, $p < .001$, $BF_{10} = 1.26 \times 10^{11}$; as well as d' , $\beta = .32$, 95% CI [.25, .40], $t(284) = 8.09$, $p < .001$, $BF_{10} = 2938815$. Finally, the pooled analysis for high confidence recognition memory confirmed that this memory advantage also carried over to the postconditioning phase, as reflected in significantly increased CS^+ over CS^- scores in both corrected recognition, $\beta = .037$, 95% CI [.021, .054], $t(284) = 4.47$, $p < .001$, $BF_{10} = 3.62$, as well as in d' , $\beta = .087$, 95% CI [.022, .152], $t(284) = 2.62$, $p = .009$, $BF_{10} = .49$.

General Discussion

Adaptive episodic memory has been theorized to preferentially store motivationally significant experiences that can be useful to guide future behavior (Nairne et al., 2007; Nairne & Pandeirada, 2008; Shohamy & Adcock, 2010). How can such an adaptive prioritization be achieved for stimuli that appear neutral during encoding, but are subsequently revealed to relate to important consequences? Recently, a possible mechanism has been suggested that retroactively and selectively promotes memory for initially neutral items when their respective category is later predictive of either aversive or appetitive events (Dunsmoor et al., 2015; Patil et al., 2017). These findings have challenged existing models of episodic memory formation by demonstrating for the first time that postencoding processes can selectively enhance memory for a specific group of stimuli, but not others, based on the (categorical) relatedness of the stimuli to the emotional event. In this framework, memories can exist in a weak, transient form ('tagged') that relies on a subsequent event ('capture') to store them permanently. This tag and capture framework had previously been developed at the level of individual neurons and was referred to as synaptic tagging (Frey & Morris, 1997, 1998). Studies in rodents and more recently in humans have successfully translated this framework to the behavioral level (Ballarini et al., 2009, 2013; de Carvalho Myskiw et al., 2013; Moncada et al., 2015). Importantly, these studies investigated general, nonselective retroactive effects on memory, irrespective of a semantical link between tagged stimuli and the subsequent memory-promoting event. Only recently, it has been reported that retroactive enhancements may selectively promote memory for one category of stimuli that has been linked to a subsequent arousing event, while leaving irrelevant stimuli unaffected (Dunsmoor et al., 2015; Patil et al., 2017). In addition to this category-specific retroactive (backward) effect, selective category-specific memory enhancements were also observed when appetitive or aversive stimuli were present during encoding (online) and for items from the relevant category that were encoded after these salient stimuli were present (i.e., a forward effect). Together, these findings of highly selective backward and forward memory effects are in contrast to more traditional models of memory formation, which have focused on effects that are driven through the allocation of attention during online encoding (Mulligan, 1998; Uncapher & Rugg, 2005) and general offline effects of physiological arousal that enhance consolidation in a nonselective fashion (McGaugh, 2015), irrespective of the semantic or conceptual relatedness of stimuli. In particular, the finding of a category-specific retroactive memory enhancement is incompatible with previous attentional models, as unlike online and forward enhancements, this backward effect cannot be explained by increased attention to stimuli from the category that had been

linked with salient outcomes, since this associative link was only established after the encoding of these items. Therefore, this highly selective retroactive memory enhancement is at the heart of this new framework.

In a series of four experiments, we aimed to replicate findings of the first published study showing category-specific retroactive memory enhancement for initially neutral stimuli through a following Pavlovian fear-conditioning procedure that linked aversive electric shocks to only one category of stimuli (Dunsmoor et al., 2015). Based on recent reports (Dunsmoor et al., 2015; Patil et al., 2017), we expected that memory for the initially neutral items would retroactively be enhanced when these are later revealed to belong to a relevant category. In sharp contrast to our hypotheses, analyses of overall recognition memory performance (as in Dunsmoor et al., 2015) failed to produce any evidence for a category-specific retroactive memory enhancement through aversive learning in all four experiments. Parallel Bayesian analyses provided substantial evidence for the null hypothesis speaking against a category-specific retroactive memory effect in Experiments 1 and 3 and anecdotal evidence for the null hypothesis in Experiments 2 and 4.

In a pooled analysis across all four experiments, we observed a similar pattern of results: When recognition memory was collapsed over confidence, evidence for category-specific retroactive memory enhancement was found neither in corrected recognition scores, nor in memory sensitivity (d'). In both cases, Bayes factors indicated substantial evidence for the null hypothesis. Only when additional analyses focused on high confidence memory and corrected recognition was there some evidence for the predicted category-specific retroactive memory effect, which was however only significant in one out of four experiments and was not paralleled by a significant improvement in memory sensitivity (d'), nor was it supported by a Bayesian analysis.

How can the inconsistencies between the previous reports of category-specific retroactive memory enhancements and the current findings be explained? Although close replications can be challenging (Stroebe & Strack, 2014) and seemingly small deviations from the original procedure can dramatically affect the replicability of a finding (Noah et al., 2018; Wagenmakers et al., 2016), Experiments 1, 2, and 4 were designed to match the procedure of the previous studies regarding various aspects such as timing, instructions, and stimuli, while substantially increasing the sample size. We focused only on the group of participants in which there was a 24-hr interval between encoding and recognition test, as these participants had shown the most robust evidence for category-specific retroactive memory enhancement (Dunsmoor et al., 2015). Other groups featured in the original study, such as an immediate retrieval or a strong encoding 24-hr retrieval group had not shown evidence for category-specific retroactive memory enhancement. Importantly, these group differences only become meaningful once the existence of the phenomenon is demonstrated in the first place. Achieved statistical powers were generally greater than 95% (except for Experiment 1, which used a sample size comparable to previous reports suggesting a selective behavioral tagging effect). Thus, a lack of statistical power is very unlikely.

Two further aspects that could have potentially affected the replicability of category-specific retroactive memory enhancements in Experiments 1–3 were (a) deviations in the relative timing of the

CS to the UCS compared with Dunsmoor et al. (2015) and (b) the random allocation of stimuli to each learning phase instead of controlling for typicality across their respective categories. In Dunsmoor et al. (2015), each 200-ms shock coterminated with the end of the stimulus presentation, whereas in our Experiments 1–3, the 200-ms shocks started with the end of the stimulus presentation. This resulting a 200-ms deviation in CS-UCS timing compared with Dunsmoor et al. (2015) was unintentional. Potentially, this issue could be relevant as the differential timing between CS and UCS can be used to differentiate between trace and delay conditioning, which involve different processes (Kochli et al., 2015; McLaughlin et al., 2002; Weike et al., 2007). However, trace conditioning would only be present if an additional pause were implemented between stimulus offset and the following UCS. Because this was not the case in Experiments 1–3, our procedure may still be considered a delay conditioning procedure like the one used in Dunsmoor et al. (2015). Moreover, we obtained significantly higher anticipatory SCRs for CS⁺ compared with CS⁻ items, indicating that our fear conditioning manipulation was successful. Furthermore, we could replicate the memory benefit for CS⁺ items that were presented during fear conditioning and partly the prospective memory effect for items that were presented after fear-conditioning, indicating that, despite the deviation in CS-UCS timing, the UCS was still able to modulate memory in Experiments 1 to 3. To our knowledge, there is no theoretical justification why only the retroactive effect, but not the online, nor the prospective effect should be affected by this difference in timing. Finally, after explicitly addressing the issue of CS-UCS timing in Experiment 4, we obtained a similar pattern of results as in Experiments 1–3 that most prominently did not show any signs of category-specific retroactive memory enhancement.

Regarding stimulus typicality, we unfortunately only learned during the peer review process that Dunsmoor et al. (2015) kept stimulus typicality constant across learning phases as this aspect was not mentioned at all in their original article. Even after balancing stimulus typicality across encoding phases in Experiment 4, we still found no evidence for any category-specific retroactive memory enhancement.

Another factor that has been suggested to moderate the extent of category-specific memory enhancement is the interval between the encoding of initially neutral stimuli and the following significant (either aversive or appetitive) event. Specifically, Dunsmoor et al. (2015) reported a linear trend between the distance of learning items to the following significant event and the strength of category-specific retroactive memory enhancement. This linear trend is in line with previous work on nonspecific behavioral tagging in animals suggesting that a minimal interval between initial learning and the following event is necessary for such effects to unfold (de Carvalho Myskiw et al., 2013; Moncada et al., 2015). We specifically addressed this issue in Experiment 3 by extending the interval between preconditioning and subsequent Pavlovian conditioning. It is important to note that, for practical reasons, this interval had to be at least approximately 10 min even in Experiments 1, 2, and 4. This time was needed to attach electrodes and adjust shock intensities and should correspond with Dunsmoor et al. (2015). For Experiment 3, we effectively doubled this interval to 20 min, which did not lead to the expected increase of the putative category-specific retroactive memory effect. Furthermore, none of Experiments 1, 2, 3, or 4 provided any evidence for the previously reported linear trend between

the temporal distance of an item of the preconditioning phase to the conditioning procedure and the size of category-specific retroactive memory enhancement.

Retroactive memory effects have been further theorized to only strengthen initially weak memories, but to have no additional benefit for already strongly encoded stimuli (Dunsmoor et al., 2015; Moncada & Viola, 2007; Wang et al., 2010). Accordingly, it could be speculated that our sample of participants included better learners, which might have prevented category-specific retroactive memory enhancement due to strong initial encoding. However, recognition performance in the present experiments was, with the exception of Experiment 1, comparable with previous studies reporting selective retroactive memory enhancements (Dunsmoor et al., 2015; Patil et al., 2017). Because Experiment 1 featured both a slightly different set of stimuli (although from the same categories) as well as a different format for the recognition test, this might explain the slightly increased overall preconditioning performance in this experiment compared with Dunsmoor et al. (2015). Both aspects were addressed in Experiments 2, 3, and 4 such that these featured the same set of stimuli and the same recognition test procedure. In these three experiments, we obtained similar memory performances during preconditioning as in Dunsmoor et al. (2015): For instance, CS⁻ preconditioning items were correctly classified as definitely old in 42.6% of all cases for the 24-hr retrieval group in Dunsmoor et al. (2015) and the corresponding performance ranged from 37.5% to 42.0% in our Experiments 2–4 (Tables 2–4). This renders overly strong memories as explanation for the absence of a selective retroactive memory effects rather unlikely. Additionally, despite participants indicating that they were overall slightly less surprised by the recognition test compared with Dunsmoor et al. (2015) and it cannot be completely ruled out that such differences may have influenced our results, although this would clearly question the robustness of the suggested category-specific tagging effect, there is no clear theoretical rationale why such a subtle difference should abolish the tagging effects. None of our experiments revealed any correlation between levels of surprise and memory performance.

It might be argued that emotion has a higher impact on memory for items recognized with high confidence (Kim & Cabeza, 2009; Phelps & Sharot, 2008). We therefore ran additional analysis that focused on high confidence memory only. In one of the four experiments (Experiment 2), this exploratory recognition analysis based on corrected recognition scores and focusing on high confidence hits showed a significant category-specific retroactive memory effect, although a parallel Bayesian analysis indicated that evidence was nonsubstantial. For Experiment 4, there was a nonsignificant trend in the same direction ($p = .088$). Interestingly, this retroactive memory enhancement for high confidence hits in Experiment 2 and respective trend in Experiment 4 were only detectable in corrected recognition scores, but not in memory sensitivity (d') from signal detection theory. Further, in the remaining two experiments, there was no evidence for a category-specific retroactive effect for high confidence memory and a Bayesian analysis on high confidence corrected recognition contrarily favored the null hypothesis. A pooled analysis across all four experiments that focused on high confidence corrected recognition showed a small but significant category-specific retroactive memory effect. A parallel Bayesian analysis, however, showed even substantial evidence for the null hypothesis rejecting the notion of category-

specific retroactive memory enhancement. The same pooled analysis for memory sensitivity (d') was much clearer: Neither for recognition scores collapsed across confidence, nor for those focusing on high confidence hits was there any evidence for the category-specific retroactive memory effect, with a parallel Bayesian analysis indicating substantial evidence for the null hypothesis in both cases. Although it is to be acknowledged that the two experiments in which we obtained an effect or a similar trend for a retroactive effect in high confidence memory might be considered the closest replication attempts to Dunsmoor et al. (2015), even in these experiments the evidence was not robust across memory parameters.

In the face of the findings of a significant category-specific retroactive memory effect for high confidence corrected recognition scores in Experiment 2 and the pooled analysis, it must also be noted we run multiple analyses (overall memory analysis, high confidence memory analysis, linear mixed models) across multiple parameters (corrected recognition score and d') and multiple experiments. This wide array of tests comes with a significantly increased risk of false positives (i.e., an inflated alpha-error rate). Only in one of the four experiments, there was a significant result and only in corrected recognition scores, but not in d' from signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002). As we aimed for the maximum sensitivity regarding possible effects, we did not correct for the relatively high number of statistical tests. If any correction for multiple testing was performed, none of the effects or trends for high confidence memory would be even close to statistical significance. Therefore, additional caution against interpreting the findings on high confidence memory as a successful replication of category-specific retroactive memory enhancement is warranted.

The observed discrepancy in results between analyses based on d' versus corrected recognition scores is interesting because both methods of estimating discrimination performance rely on different models of recognition memory (Snodgrass & Corwin, 1988). Whereas d' is rooted in signal detection theory (Macmillan & Creelman, 2005; Wickens, 2002) and assumes curvilinear receiver operating characteristics (ROCs), corrected recognition scores as calculated by Dunsmoor et al. (2015) stem from the two-high-threshold model of recognition (Bröder et al., 2013; Snodgrass & Corwin, 1988) and assume linear ROCs. The issue of selecting the correct model is particularly important as we also found that participants showed for items from the CS^- category a more conservative response bias c (from signal detection theory) than for items from the CS^+ category for high confidence responses at least in Experiment 4. Ideally, this response bias should not influence memory discrimination scores, as it does not reflect true memory but rather a response tendency. Indeed, when that the assumptions of signal detection theory are correct, memory sensitivity (d') and response bias (c) are theoretically independent from each other (Snodgrass & Corwin, 1988). Likewise, if the model underlying corrected recognition scores is correct (i.e., the two-high-threshold model), these scores should equally be independent of the response bias. Although there has been some debate regarding the question which of these two approaches is generally more appropriate in the memory context, most empirical findings favor the use of signal detection theory (and therefore d') over the two-high-threshold model (associated with corrected recognition) when analyzing recognition performance (Dube & Rotello, 2012; Pazzaglia et al., 2013; Slotnick & Dodson, 2005). Future research on

the category-specific retroactive memory effect should optimally report results from both measures, consider theoretical implications if such an effect was detectable in only one measure but not the other, and consider possible response biases.

It should be noted that, although none of our four experiments provided consistent evidence for the existence of category-specific retroactive ('backward') memory enhancement, there was some evidence for the selective online and forward memory enhancements. In line with category-specific online effects, in all four experiments we consistently found a memory advantage for items from the CS^+ category that were presented during Pavlovian fear conditioning compared with items from the CS^- category encoded in the same learning phase. This finding corroborates previous studies showing enhanced memory for stimuli linked to arousing events (Dunsmoor et al., 2015; Dunsmoor & Kroes, 2019; Salehi et al., 2010; Vogel & Schwabe, 2016). In the context of adaptive memory, such a mechanism enables the preferential storage of stimuli that are associated with threat which may facilitate coping to similar situations in the future (Nairne et al., 2007; Nairne & Pandeirada, 2008). It is important to note that this memory enhancement for CS^+ items in Experiments 1 to 4 was evaluated by comparing them with CS^- items from the same category. Therefore, an alternative interpretation of these findings could be that CS^+ items encoded during fear conditioning did not experience a memory promotion per se, but instead that memory for CS^- items was diminished through fear conditioning. Modifying the task to test these two options is beyond the scope of our replication attempt.

Beyond selective backward and online memory enhancements, the proposed tag-and-capture framework predicts category-specific memory enhancement in a forward, prospective direction. Our results provided indeed evidence for a selective influence of emotionally arousing events on the encoding of subsequent related events. More specifically, the enhanced memory for stimuli paired with aversive shocks seemed to extend to subsequent stimuli belonging to the same category as the CS^+ . Although there was clear evidence for such a selective forward enhancement in the pooled analysis across all four experiments, it is to be noted that this effect was only significant in Experiment 2 and at trend level in Experiments 1, 3, and 4, suggesting a small to moderate effect.

Together, our results suggest a selective memory enhancement for aversive, threat-related stimuli, both online, while a threat is present (e.g., during the fear conditioning procedure), and in a forward direction for threat-related stimuli that are encoded after the threat (e.g., in the postconditioning phase). Both, the online and forward effects may be related to changes in stimulus saliency. During encoding stimuli predictive of motivationally relevant events will be more salient. Likewise, the previously learned association between stimuli and aversive events may increase the saliency of subsequently encoded stimuli that are conceptually linked to the threat-related stimuli. Such increases in saliency may help stimuli to directly exceed the threshold for long-term memory storage. The resulting selectivity in episodic memory has considerable impact on the architecture of our autobiographical memory and, although being generally adaptive, may propel dysfunctional memory in a variety of psychiatric disorders, such as anxiety disorders (Airaksinen et al., 2005; Coles et al., 2007; de Quervain et al., 2017), posttraumatic stress

disorder (Brown et al., 2014; Isaac et al., 2006), or depression (Airaksinen et al., 2007; Lemogne et al., 2006; McDermott & Ebmeier, 2009). In contrast to the online and forward memory enhancements, selective backward enhancements would require the retroactive enhancement of initially weakly encoded (tagged) stimuli to overcome the threshold for long-term storage. Most importantly, however, we obtained only very limited evidence for a selective retroactive (backward) memory enhancement. This raises the question how the brain adapts when certain stimuli only gain relevance after their initial encoding. One solution in line with previous studies is by nonselectively enhancing memory for events preceding an aversive (e.g., stressful) event, regardless of their relation to the relevant event (Cahill et al., 2003; Smeets et al., 2008). In fact, we cannot exclude that such a general, unspecific memory enhancement took place in our experiments. For example, memory for items from the preconditioning phase might have been promoted unspecifically through the following fear-conditioning procedure. Even in Dunsmoor et al. (2015), such a general effect might have played a role in addition to category-specific enhancements for CS⁺ items. However, because this task was specifically designed to investigate category-specific, rather than general retroactive memory enhancement through the within-subject comparison of CS⁺ and CS⁻ items, this question is beyond the scope of this replication attempt.

Another solution that has not been considered by the literature so far could be that in these cases, an even more specific retroactive enhancement takes place, which does not apply to a relatively wide array of stimuli of the same category but only strengthens the memory trace of a single stimulus. Relating back to the example of the bank customer's encounter with the bank robber, importance lies on the memory for only the specific face of the robber and not for other faces seen shortly before (e.g., that of all men). Therefore, adaptive memory would call for a memory promotion of only the specific face and not other faces from the same abstract category. Whether such a mechanism exists, however, is currently unknown and needs to be tested in future research.

In summary, the present series of experiments searched for category-specific, selective retroactive memory enhancement of initially neutral stimuli as suggested by two recent studies (Dunsmoor et al., 2015; Patil et al., 2017). Our data yielded only very limited evidence for a category-specific retroactive memory enhancement in line with Dunsmoor et al. (2015). We acknowledge that although we aimed to stick as closely as possible to the experimental procedure reported by Dunsmoor et al. (2015), subtle differences between studies (e.g., related to the specific sample) can hardly be ruled out. The fact that we did not obtain any evidence for a category-specific retroactive memory enhancement when strictly replicating the reported analysis across four separate experiments, with three of them being highly powered, suggests that this effect is not reliable. At least, the present data clearly question the generalizability of the suggested category-specific retroactive memory enhancement. Still, arousing events might promote episodic memory for recently encountered stimuli in a general, nonselective fashion, as previous evidence suggests (Christianson et al., 1991; McGaugh, 2018; McGaugh & Roozendaal, 2002). These findings of nonselective memory enhancement are in line with previous applications of the synaptic tag-and-

capture mechanism to the behavioral level, which demonstrated memory enhancement for weakly encoded stimuli through following arousing events even in absence of a semantical link between these two (Ballarini et al., 2009, 2013; de Carvalho Myskiw et al., 2013). From a theoretical point of view, this nonselective memory promotion might be regarded as a "safe" alternative to a category-specific retroactive memory promotion, since it does not require a model of events and their putative consequences, which is at risk to be incorrect and might therefore miss important predictors of significant outcomes. On the other hand, such non-specific memory promotion is not only inefficient as invalid cues are subjected to the same memory promotion as valid cues but might also contribute to psychopathology associated with errant memory functions such as posttraumatic stress disorder (Brown et al., 2014; Pitman, 1989). Elucidating how our memory balances the need for efficiency on the one hand and the need for an enhanced storage of experiences that preceded a significant event on the other hand remains a challenge for future research.

Context Paragraph

Our lab focusses on how emotion and stress can bias memory formation. Thus, we were intrigued by recent reports (Dunsmoor et al., 2015; Patil et al., 2017) suggesting a highly specific behavioral tagging mechanism according to which an emotionally arousing event could retroactively enhance memory selectively for preceding events that were conceptually relevant to the emotional event. The proposed mechanism would be highly adaptive in that it would enable our memory to retroactively enhance selectively the storage of material that turned out to be important later on. We aimed to elucidate the mechanisms underlying this selective, retroactive memory enhancement. However, when we tried to replicate the effect in the first place, we originally did not find evidence for a selective retroactive memory enhancement. Only in specific exploratory analyses proposed during the peer-review process, we obtained some limited evidence for the effect. Given the tremendous implications of the suggested retroactive and selective memory enhancement for understanding memory in general and for disorders such as PTSD, we believe that it is important to bring the findings of this series of experiments to the attention of our colleagues. Our hope is that these findings will inspire new theories and experimental paradigms to address the fundamental issue of how our memory can preferentially store events that are relevant for a subsequent emotional episode.

References

- Airaksinen, E., Larsson, M., & Forsell, Y. (2005). Neuropsychological functions in anxiety disorders in population-based samples: Evidence of episodic memory dysfunction. *Journal of Psychiatric Research, 39*(2), 207–214. <https://doi.org/10.1016/j.jpsychires.2004.06.001>
- Airaksinen, E., Wahlin, Å., Forsell, Y., & Larsson, M. (2007). Low episodic memory performance as a premorbid marker of depression: Evidence from a 3-year follow-up. *Acta Psychiatrica Scandinavica, 115*(6), 458–465. <https://doi.org/10.1111/j.1600-0447.2006.00932.x>
- Almaguer-Melian, W., Bergado-Rosado, J., Pavon-Fuentes, N., Alberti-Amador, E., Merceron-Martinez, D., & Frey, J. U. (2012). Novelty exposure overcomes foot shock-induced spatial-memory impairment by processes of synaptic-tagging in rats. *Proceedings of the National*

- Academy of Sciences of the United States of America*, 109(3), 953–958. <https://doi.org/10.1073/pnas.1114198109>
- Ballarini, F., Martínez, M. C., Díaz Perez, M., Moncada, D., & Viola, H. (2013). Memory in elementary school children is improved by an unrelated novel experience. *PLoS ONE*, 8(6), e66875. <https://doi.org/10.1371/journal.pone.0066875>
- Ballarini, F., Moncada, D., Martinez, M. C., Alen, N., & Viola, H. (2009). Behavioral tagging is a general mechanism of long-term memory formation. *Proceedings of the National Academy of Sciences of the United States of America*, 106(34), 14599–14604. <https://doi.org/10.1073/pnas.0907078106>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Beck Depression Inventory II*. The Psychological Corporation.
- Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, 190(1), 80–91. <https://doi.org/10.1016/j.jneumeth.2010.04.028>
- Bliss, T. V. P., & Collingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature*, 361(6407), 31–39. <https://doi.org/10.1038/361031a0>
- Bröder, A., Kellen, D., Schütz, J., & Rohrmeier, C. (2013). Validating a two-high-threshold measurement model for confidence rating data in recognition. *Memory*, 21(8), 916–944. <https://doi.org/10.1080/09658211.2013.767348>
- 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(5), e10773. <https://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(9), e106953. <https://doi.org/10.1371/journal.pone.0106953>
- Brown, A. D., Addis, D. R., Romano, T. A., Marmar, C. R., Bryant, R. A., Hirst, W., & Schacter, D. L. (2014). Episodic and semantic components of autobiographical memories and imagined future events in post-traumatic stress disorder. *Memory*, 22(6), 595–604. <https://doi.org/10.1080/09658211.2013.807842>
- Cahill, L., Gorski, L., & Le, K. (2003). Enhanced human memory consolidation with post-learning stress: Interaction with the degree of arousal at encoding. *Learning & Memory*, 10(4), 270–274. <https://doi.org/10.1101/lm.62403>
- Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences*, 21(7), 294–299. [https://doi.org/10.1016/S0166-2236\(97\)01214-9](https://doi.org/10.1016/S0166-2236(97)01214-9)
- 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(4), 693–701. <https://doi.org/10.1037/0278-7393.17.4.693>
- Coles, M. E., Turk, C. L., & Heimberg, R. G. (2007). Memory bias for threat in generalized anxiety disorder: The potential importance of stimulus relevance. *Cognitive Behaviour Therapy*, 36(2), 65–73. <https://doi.org/10.1080/16506070601070459>
- de Carvalho Myskiw, J., Benetti, F., & Izquierdo, I. (2013). Behavioral tagging of extinction learning. *Proceedings of the National Academy of Sciences of the United States of America*, 110(3), 1071–1076. <https://doi.org/10.1073/pnas.1220875110>
- de Quervain, D., Schwabe, L., & Roozendaal, B. (2017). Stress, glucocorticoids and memory: Implications for treating fear-related disorders. *Nature Reviews Neuroscience*, 18(1), 7–19. <https://doi.org/10.1038/nrn.2016.155>
- Dougal, S., & Rotello, C. M. (2007). Remembering” emotional words is based on response bias, not recollection. *Psychonomic Bulletin & Review*, 14(3), 423–429. <https://doi.org/10.3758/BF03194083>
- Dube, C., & Rotello, C. M. (2012). Binary ROCs in perception and recognition memory are curved. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 38(1), 130–151. <https://doi.org/10.1037/a0024957>
- Dunsmoor, J. E., & Kroes, M. C. (2019). Episodic memory and Pavlovian conditioning: Ships passing in the night. *Current Opinion in Behavioral Sciences*, 26, 32–39. <https://doi.org/10.1016/j.cobeha.2018.09.019>
- Dunsmoor, J. E., Martin, A., & LaBar, K. S. (2012). Role of conceptual knowledge in learning and retention of conditioned fear. *Biological Psychology*, 89(2), 300–305. <https://doi.org/10.1016/j.biopsycho.2011.11.002>
- Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, 520(7547), 345–348. <https://doi.org/10.1038/nature14106>
- 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(2), 175–191. <https://doi.org/10.3758/BF03193146>
- Frey, U., & Morris, R. G. M. (1997). Synaptic tagging and long-term potentiation. *Nature*, 385(6616), 533–536. <https://doi.org/10.1038/385533a0>
- Frey, U., & Morris, R. G. M. (1998). Synaptic tagging: Implications for late maintenance of hippocampal long-term potentiation. *Trends in Neurosciences*, 21(5), 181–188. [https://doi.org/10.1016/S0166-2236\(97\)01189-2](https://doi.org/10.1016/S0166-2236(97)01189-2)
- Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, 68(1), 101–128. <https://doi.org/10.1146/annurev-psych-122414-033625>
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486–504. <https://doi.org/10.1037/1082-989X.3.4.486>
- Isaac, C. L., Cushway, D., & Jones, G. V. (2006). Is posttraumatic stress disorder associated with specific deficits in episodic memory? *Clinical Psychology Review*, 26(8), 939–955. <https://doi.org/10.1016/j.cpr.2005.12.004>
- Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, 7(1), 2. <https://doi.org/10.7771/1932-6246.1167>
- Keller, N. E., & Dunsmoor, J. E. (2020). The effects of aversive-to-appetitive counterconditioning on implicit and explicit fear memory. *Learning & Memory*, 27(1), 1212–1219. <https://doi.org/10.1101/lm.050740.119>
- Kim, H., & Cabeza, R. (2009). Common and specific brain regions in high- versus low-confidence recognition memory. *Brain Research*, 1282, 103–113. <https://doi.org/10.1016/j.brainres.2009.05.080>
- Kochli, D. E., Thompson, E. C., Fricke, E. A., Postle, A. F., & Quinn, J. J. (2015). The amygdala is critical for trace, delay, and contextual fear conditioning. *Learning & Memory*, 22(2), 92–100. <https://doi.org/10.1101/lm.034918.114>
- LaBar, K. S., & Cabeza, R. (2006). Cognitive neuroscience of emotional memory. *Nature Reviews Neuroscience*, 7(1), 54–64. <https://doi.org/10.1038/nrn1825>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Lemogne, C., Piolino, P., Friszer, S., Claret, A., Girault, N., Jouvent, R., Allilaire, J. F., & Fossati, P. (2006). Episodic autobiographical memory in depression: Specificity, autoegetic consciousness, and self-perspective.

- Consciousness and Cognition*, 15(2), 258–268. <https://doi.org/10.1016/j.concog.2005.07.005>
- Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Erlbaum.
- Malenka, R. C., & Nicoll, R. A. (1999). Long-term potentiation—A decade of progress? *Science*, 285(5435), 1870–1874. <https://doi.org/10.1126/science.285.5435.1870>
- Martin, K. C., & Kosik, K. S. (2002). Synaptic tagging—Who's it? *Nature Reviews Neuroscience*, 3(10), 813–820. <https://doi.org/10.1038/nrn942>
- Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour Research and Therapy*, 36(4), 455–470. [https://doi.org/10.1016/S0005-7967\(97\)10031-6](https://doi.org/10.1016/S0005-7967(97)10031-6)
- McDermott, L. M., & Ebmeier, K. P. (2009). A meta-analysis of depression severity and cognitive function. *Journal of Affective Disorders*, 119(1-3), 1–8. <https://doi.org/10.1016/j.jad.2009.04.022>
- McGaugh, J. L. (2015). Consolidating memories. *Annual Review of Psychology*, 66(1), 1–24. <https://doi.org/10.1146/annurev-psych-010814-014954>
- McGaugh, J. L. (2018). Emotional arousal regulation of memory consolidation. *Current Opinion in Behavioral Sciences*, 19, 55–60. <https://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(2), 205–210. [https://doi.org/10.1016/S0959-4388\(02\)00306-9](https://doi.org/10.1016/S0959-4388(02)00306-9)
- McLaughlin, J., Skaggs, H., Churchwell, J., & Powell, D. A. (2002). Medial prefrontal cortex and Pavlovian conditioning: Trace versus delay conditioning. *Behavioral Neuroscience*, 116(1), 37–47. <https://doi.org/10.1037/0735-7044.116.1.37>
- Moncada, D., Ballarini, F., & Viola, H. (2015). Behavioral tagging: A translation of the synaptic tagging and capture hypothesis. *Neural Plasticity*, 2015, 650780. <https://doi.org/10.1155/2015/650780>
- Moncada, D., & Viola, H. (2007). Induction of long-term memory by exposure to novelty requires protein synthesis: Evidence for a behavioral tagging. *The Journal of Neuroscience*, 27(28), 7476–7481. <https://doi.org/10.1523/JNEUROSCI.1083-07.2007>
- Morey, R. D., Rouder, J. N., Jamil, T., Urbanek, S., Forner, K., & Ly, A. (2018). Package “BayesFactor” 0.9.12-4.2. <https://cran.r-project.org/web/packages/BayesFactor/index.html>
- Mulligan, N. W. (1998). The role of attention during encoding in implicit and explicit memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24(1), 27–47. <https://doi.org/10.1037/0278-7393.24.1.27>
- Murty, V. P., FeldmanHall, O., Hunter, L. E., Phelps, E. A., & Davachi, L. (2016). Episodic memories predict adaptive value-based decision-making. *Journal of Experimental Psychology: General*, 145(5), 548–558. <https://doi.org/10.1037/xge0000158>
- Nairne, J. S., & Pandeirada, J. N. S. (2008). Adaptive memory: Remembering with a stone-age brain. *Current Directions in Psychological Science*, 17(4), 239–243. <https://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(2), 263–273. <https://doi.org/10.1037/0278-7393.33.2.263>
- Noah, T., Schul, Y., & Mayo, R. (2018). When both the original study and its failed replication are correct: Feeling observed eliminates the facial-feedback effect. *Journal of Personality and Social Psychology*, 114(5), 657–664. <https://doi.org/10.1037/pspa0000121>
- Oyarzún, J. P., Packard, P. A., de Diego-Balaguer, R., & Fuentemilla, L. (2016). Motivated encoding selectively promotes memory for future inconsequential semantically-related events. *Neurobiology of Learning and Memory*, 133, 1–6. <https://doi.org/10.1016/j.nlm.2016.05.005>
- Patil, A., Murty, V. P., Dunsmoor, J. E., Phelps, E. A., & Davachi, L. (2017). Reward retroactively enhances memory consolidation for related items. *Learning & Memory*, 24(1), 6565–6569. <https://doi.org/10.1101/lm.042978.116>
- Pazzaglia, A. M., Dube, C., & Rotello, C. M. (2013). A critical comparison of discrete-state and continuous models of recognition memory: Implications for recognition and beyond. *Psychological Bulletin*, 139(6), 1173–1203. <https://doi.org/10.1037/a0033044>
- Phelps, E. A., & Sharot, T. (2008). How (and why) emotion enhances the subjective sense of recollection. *Current Directions in Psychological Science*, 17(2), 147–152. <https://doi.org/10.1111/j.1467-8721.2008.00565.x>
- Pitman, R. K. (1989). Post-traumatic stress disorder, hormones, and memory. *Biological Psychiatry*, 26(3), 221–223. [https://doi.org/10.1016/0006-3223\(89\)90033-4](https://doi.org/10.1016/0006-3223(89)90033-4)
- Redondo, R. L., & Morris, R. G. M. (2011). Making memories last: The synaptic tagging and capture hypothesis. *Nature Reviews Neuroscience*, 12(1), 17–30. <https://doi.org/10.1038/nrn2963>
- Rogerson, T., Cai, D. J., Frank, A., Sano, Y., Shobe, J., Lopez-Aranda, M. F., & Silva, A. J. (2014). Synaptic tagging during memory allocation. *Nature Reviews Neuroscience*, 15(3), 157–169. <https://doi.org/10.1038/nrn3667>
- Rotello, C. M., Masson, M. E. J., & Verde, M. F. (2008). Type I error rates and power analyses for single-point sensitivity measures. *Perception & Psychophysics*, 70(2), 389–401. <https://doi.org/10.3758/pp.70.2.389>
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t*-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16(2), 225–237. <https://doi.org/10.3758/PBR.16.2.225>
- Salehi, B., Cordero, M. I., & Sandi, C. (2010). Learning under stress: The inverted-U-shape function revisited. *Learning & Memory*, 17(10), 522–530. <https://doi.org/10.1101/lm.1914110>
- Schulz, P., Schlotz, W., & Becker, P. (2004). *Trierer Inventar zum chronischen Stress: TICS* [Trier Inventory for Chronic Stress (TICS)]. Hogrefe.
- 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(7), 1740–1749. <https://doi.org/10.1016/j.neubiorev.2011.07.002>
- Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences*, 14(10), 464–472. <https://doi.org/10.1016/j.tics.2010.08.002>
- Slotnick, S. D., & Dodson, C. S. (2005). Support for a continuous (single-process) model of recognition memory and source memory. *Memory & Cognition*, 33(1), 151–170. <https://doi.org/10.3758/BF03195305>
- Smeets, T., Otgaar, H., Candel, I., & Wolf, O. T. (2008). True or false? Memory is differentially affected by stress-induced cortisol elevations and sympathetic activity at consolidation and retrieval. *Psychoneuroendocrinology*, 33(10), 1378–1386. <https://doi.org/10.1016/j.psyneuen.2008.07.009>
- Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, 117(1), 34–50. <https://doi.org/10.1037/0096-3445.117.1.34>
- Spielberger, C. D. (1983). *Manual for the State-Trait Anxiety Inventory (Form Y)*. Consulting Psychologists Press.
- Steyer, R., Schwenkmezger, P., Notz, P., & Eid, M. (1997). *Der Mehrdimensionale Befindlichkeitsfragebogen MDBF* [Multidimensional mood questionnaire]. Hogrefe.
- Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science*, 9(1), 59–71. <https://doi.org/10.1177/1745691613514450>
- Uncapher, M. R., & Rugg, M. D. (2005). Effects of divided attention on fMRI correlates of memory encoding. *Journal of Cognitive*

- Neuroscience*, 17(12), 1923–1935. <https://doi.org/10.1162/089892905775008616>
- Vogel, S., & Schwabe, L. (2016). Stress in the zoo: Tracking the impact of stress on memory formation over time. *Psychoneuroendocrinology*, 71, 64–72. <https://doi.org/10.1016/j.psyneuen.2016.04.027>
- Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Albohn, D. N., Allard, E. S., Beek, T., Benning, S. D., Blouin-Hudon, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A., Connell, L., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, 11(6), 917–928. <https://doi.org/10.1177/1745691616674458>
- Wang, S.-H., Redondo, R. L., & Morris, R. G. M. (2010). Relevance of synaptic tagging and capture to the persistence of long-term potentiation and everyday spatial memory. *Proceedings of the National Academy of Sciences of the United States of America*, 107(45), 19537–19542. <https://doi.org/10.1073/pnas.1008638107>
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS Scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070. <https://doi.org/10.1037/0022-3514.54.6.1063>
- Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.
- Weike, A. I., Schupp, H. T., and Hamm, A. O. (2007). Fear acquisition requires awareness in trace but not delay conditioning. *Psychophysiology*, 44, 170–180. <https://doi.org/10.1111/j.1469-8986.2006.00469.x>

Received December 18, 2019

Revision received March 5, 2021

Accepted March 8, 2021 ■

SUPPLEMENTARY MATERIAL

**On the search for a selective and retroactive strengthening of memory: is there evidence
for category-specific behavioral tagging?**

Felix Kalbe and Lars Schwabe

Universität Hamburg

Experiment 1

Mixed-effect analysis

As an alternative analysis of the effects of fear conditioning on recognition performance across encoding phases, we estimated generalized linear mixed-effect models (GLMMs) with a logit-link function using the *lme4* R package (Bates et al., 2015). The dependent variable was participants' binary classification of an item as either old (coded 1) or new (coded 0) in each trial of the recognition task, with collapsed responses across confidence. As fixed effects, we added category membership of an item (coded 0 for CS⁻ items and coded 1 for CS⁺ items), the phase the item was encountered in, as well as their interaction. As the encoding phase was categorical, we used dummy coding with 'new' items (i.e., lure items that only occurred in the recognition test) serving as the reference category. Additionally, we estimated random intercepts for each participant. Specifying a random intercept for each item led to singular fit in this experiment. Therefore, this specific random effect was omitted for this experiment.

Results showed a non-significant trend towards a positive interaction between conditioning category and the fear-conditioning phase, $\beta = 0.41$, 95% CI [-0.006, 0.835], $z = 1.93$, $p = .053$. In other words, participants tended to be more likely to correctly classify previously seen items from the fear-conditioning phase as old when these were from the CS⁺ category. There was no evidence for an interaction between the conditioning category and the post-conditioning phase, rejecting the notion of a proactive memory effect, $\beta = -0.001$, 95% CI [-0.41, 0.40], $z = 0.006$, $p = .99$. Most importantly, there was no evidence for any category-specific retroactive memory enhancement as indicated by the lack of an interaction between the conditioning category and the pre-conditioning phase, $\beta = 0.07$, 95% CI [-0.35, 0.49], $z = 0.31$, $p = .75$.

Response bias

Supplementary Table 1

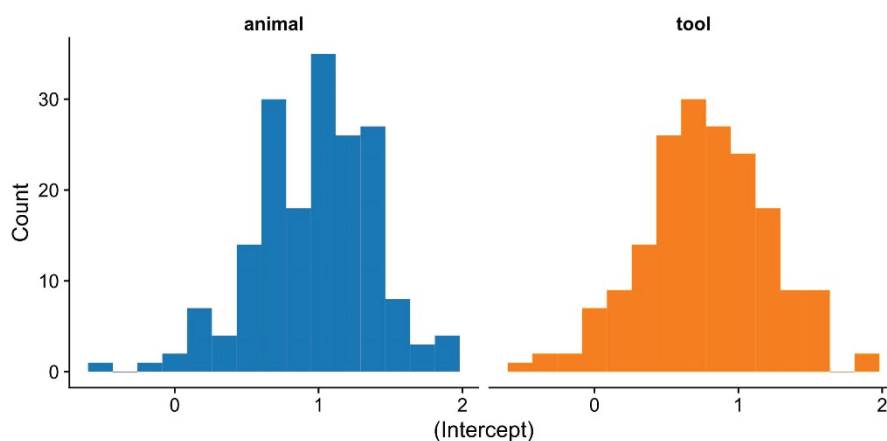
Comparison of the response bias c between conditioning categories in Experiment 1

	Collapsed across confidence				Only high confidence			
	Pre-Conditioning	Conditioning	Post-Conditioning		Pre-Conditioning	Conditioning	Post-Conditioning	
$M (SD); CS^+$	-0.005 (0.302)	0.025 (0.304)	0.109 (0.295)		0.491 (0.435)	0.510 (0.419)	0.607 (0.448)	
$M (SD); CS^-$	0.053 (0.345)	0.151 (0.400)	0.217 (0.360)		0.547 (0.357)	0.632 (0.439)	0.675 (0.383)	
d_{av}	-0.177	-0.358	-0.329		-0.142	-0.283	-0.164	
$t(43)$	-0.969	-1.944	-1.801		-0.932	-1.645	-1.052	
p	.34	.058	.079		.36	.11	.30	

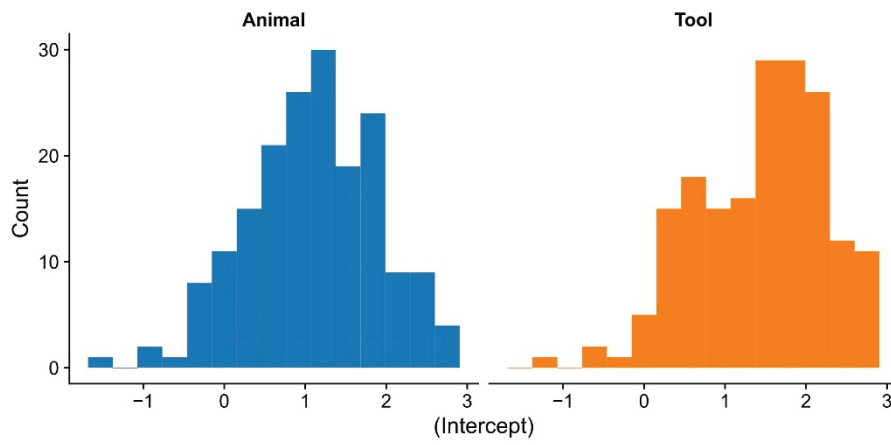
Note: Smaller values of c indicate a more liberal response bias, while larger values of c indicate a more conservative response bias. d_{av} , t , and p refer to the comparison of the response bias c between items from the CS^+ vs. CS^- category. Negative values of d_{av} indicate a bias towards more conservative responses for items from the CS^- category, while positive values of d_{av} indicate a bias towards more conservative responses for items from the CS^+ category.

Item analysis

To examine whether certain items were remembered at particularly high rates or associated with increased false alarm probabilities, we estimated separate generalized linear mixed-effect models (GLMMs) for ‘new’ and ‘old’ items with a logit-link function using the *lme4* R package (Bates et al., 2015). These models only included an intercept as a fixed effect to explain whether an item would be correctly classified as either old or new (coded 0 for incorrect and 1 for correct). Additionally, we estimated random intercepts for each item presented in the recognition test. Supplementary Figure 1 shows the distribution of per-item intercepts and photograph category membership (i.e., animals vs. tools) for previously seen items, while Supplementary Figure 2 shows the same distribution for lure items.



Supplementary Figure 1. Distribution of per-item intercepts in Experiment 1 reflecting the probability of correctly classifying previously seen photographs as ‘old’ in the recognition test.



Supplementary Figure 2. Distribution of per-item intercepts in Experiment 1 reflecting the probability of correctly classifying previously unseen photographs as ‘new’ in the recognition test.

Experiment 2

Mixed-effect analysis

As for Experiment 1, we ran an alternative analysis of the effects of fear conditioning on recognition performance across encoding phases by estimating generalized linear mixed-effect models (GLMMs). We used the same model as in Experiment 1, but additionally estimated a random intercept for each specific item of the recognition test.

Results replicated those reported in the main text by showing a memory advantage for CS⁺ items encoded during fear-conditioning (i.e., a significant interaction between conditioning category and the fear-conditioning phase), $\beta = 0.41$, 95% *CI* [0.26, 0.56], $z = 5.41$, $p < .001$. Likewise, results showed that this effect also extended to items that were encoded after fear-conditioning (i.e., a significant interaction between the conditioning category and the post-conditioning phase), $\beta = 0.21$, 95% *CI* [0.06, 0.36], $z = 2.83$, $p = .005$. Again, this analysis showed no evidence for category-specific retroactive memory enhancement (i.e., no significant interaction between the conditioning category and the pre-conditioning phase), $\beta = 0.11$, 95% *CI* [-0.04, 0.26], $z = 1.43$, $p < .16$.

Response Bias

Supplementary Table 2

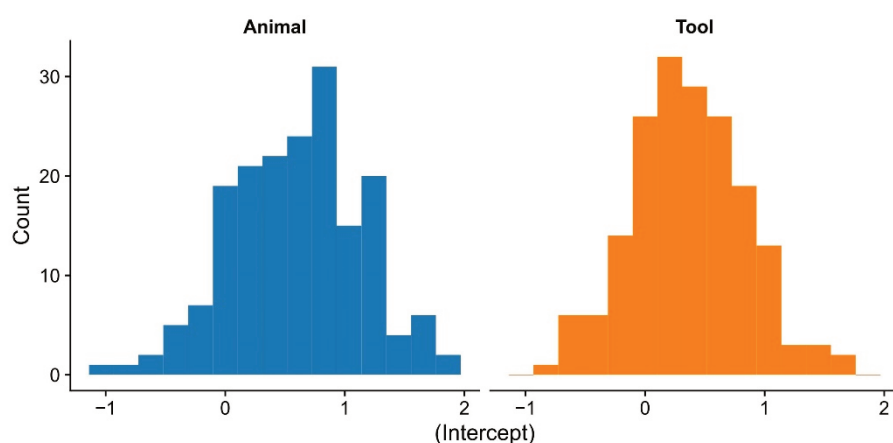
Comparison of the response bias c between conditioning categories in Experiment 2

	Collapsed across confidence				Only high confidence			
	Pre-Conditioning	Conditioning	Post-Conditioning		Pre-Conditioning	Conditioning	Post-Conditioning	
$M (SD)$; CS ⁺	0.153 (0.346)	0.095 (0.367)	0.275 (0.268)		0.901 (0.403)	0.857 (0.459)	0.996 (0.372)	
$M (SD)$; CS ⁻	0.171 (0.412)	0.213 (0.441)	0.320 (0.396)		0.964 (0.499)	1.059 (0.535)	1.103 (0.481)	
d_{av}	-0.047	-0.294	-0.137		-0.140	-0.408	-0.249	
$t(79)$	-0.335	-1.908	-0.875		-1.251	-3.556	-2.061	
p	.74	.060	.38		.21	< .001	.043	

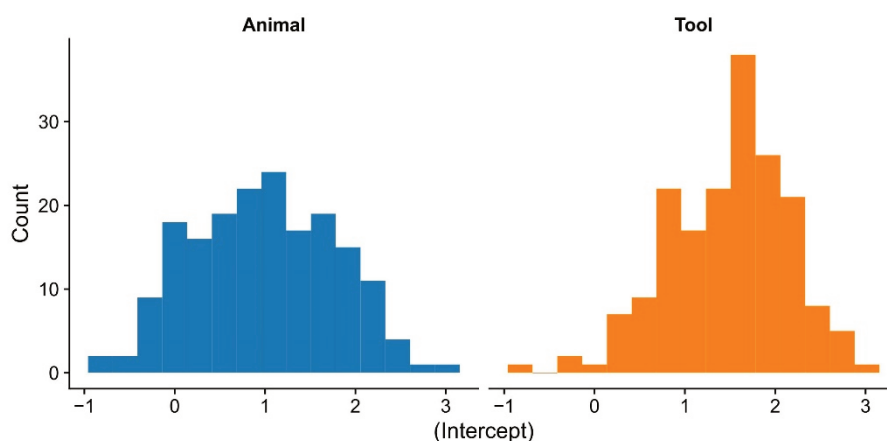
Note: Smaller values of c indicate a more liberal response bias, while larger values of c indicate a more conservative response bias. d_{av} , t , and p refer to the comparison of the response bias c between items from the CS⁺ vs. CS⁻ category. Negative values of d_{av} indicate a bias towards more conservative responses for items from the CS⁻ category, while positive values of d_{av} indicate a bias towards more conservative responses for items from the CS⁺ category.

Item analysis

We ran parallel GLMMs as in Experiment 1 to examine whether certain items were remembered at particularly high rates or associated with increased false alarm probabilities. Supplementary Figure 3 shows the distribution of per-item intercepts and photograph category membership (i.e., animals vs. tools) for previously seen items. Supplementary Figure 4 shows the same distribution for ‘new’ items.



Supplementary Figure 3. Distribution of per-item intercepts in Experiment 2 reflecting the probability of correctly classifying previously seen photographs as ‘old’ in the recognition test.



Supplementary Figure 4. Distribution of per-item intercepts in Experiment 2 reflecting the probability of correctly classifying previously unseen photographs as ‘new’ in the recognition test.

Experiment 3

Mixed-effect analysis

As an alternative analysis of the effects of fear conditioning on recognition performance across encoding phases, we again applied the same GLMM as in Experiment 2.

Again, results replicated the memory advantage for CS⁺ items encoded during fear conditioning (i.e., a significant interaction between conditioning category and the fear-conditioning phase), $\beta = 0.70$, 95% *CI* [0.54, 0.85], $z = 8.62$, $p < .001$. There was also a (non-significant) trend towards improved memory formation for CS⁺ items that were encoded after fear-conditioning (i.e., an interaction between the conditioning category and the post-conditioning phase), $\beta = 0.13$, 95% *CI* [-0.02, 0.28], $z = 1.70$, $p = .089$. As in both previous experiments, we found no evidence for any category-specific retroactive memory enhancement (i.e., no significant interaction between the conditioning category and the pre-conditioning phase), $\beta = -0.02$, 95% *CI* [-0.18, 0.13], $z = 0.27$, $p = .79$.

Response Bias

Supplementary Table 3

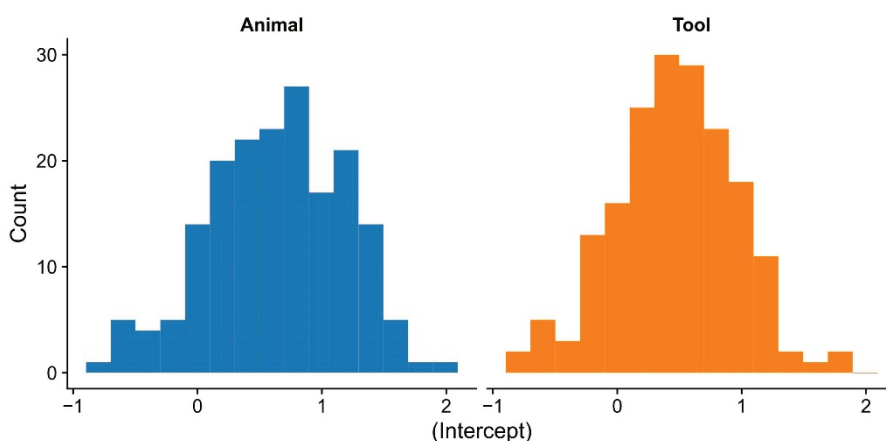
Comparison of the response bias c between conditioning categories in Experiment 3

	Collapsed across confidence				Only high confidence			
	Pre-Conditioning	Conditioning	Post-Conditioning		Pre-Conditioning	Conditioning	Post-Conditioning	
$M (SD)$; CS ⁺	0.198 (0.381)	0.048 (0.389)	0.306 (0.408)		0.948 (0.443)	0.851 (0.440)	1.045 (0.419)	
$M (SD)$; CS ⁻	0.170 (0.453)	0.241 (0.404)	0.339 (0.435)		0.940 (0.422)	1.049 (0.444)	1.114 (0.423)	
d_{av}	0.068	-0.486	-0.078		0.018	-0.449	-0.163	
$t(77)$	0.534	-4.002	-0.679		0.170	-4.046	-1.477	
p	.59	< .001	.50		.87	< .001	.14	

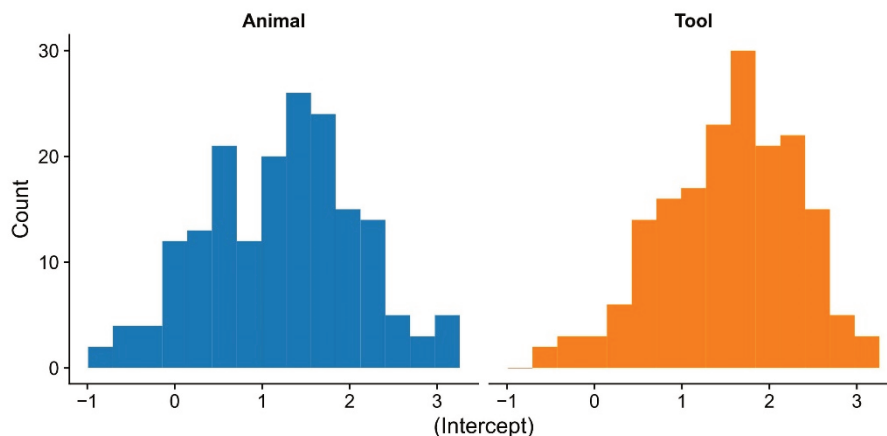
Note: Smaller values of c indicate a more liberal response bias, while larger values of c indicate a more conservative response bias. d_{av} , t , and p refer to the comparison of the response bias c between items from the CS⁺ vs. CS⁻ category. Negative values of d_{av} indicate a bias towards more conservative responses for items from the CS⁻ category, while positive values of d_{av} indicate a bias towards more conservative responses for items from the CS⁺ category.

Item analysis

We ran parallel GLMMs as in Experiments 1 and 2 to examine whether certain items were remembered at particularly high rates or associated with increased false alarm probabilities. Supplementary Figure 5 shows the distribution of per-item intercepts and photograph category membership (i.e., animals vs. tools) for previously seen items. Supplementary Figure 6 shows the same distribution for ‘new’ items.



Supplementary Figure 5. Distribution of per-item intercepts in Experiment 3 reflecting the probability of correctly classifying previously seen photographs as ‘old’ in the recognition test.

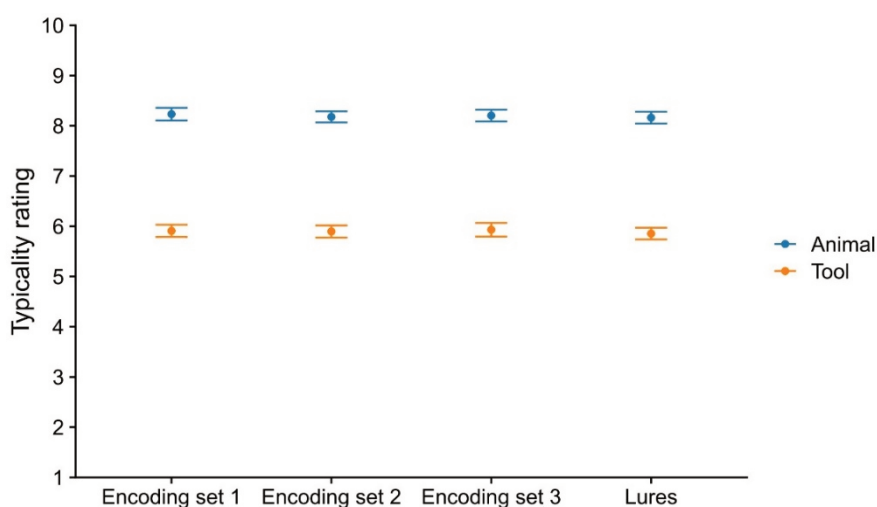


Supplementary Figure 6. Distribution of per-item intercepts in Experiment 3 reflecting the probability of correctly classifying previously unseen photographs as ‘new’ in the recognition test.

Experiment 4

Typicality ratings and stimulus allocation

Prior to data collection for Experiment 4, we obtained typicality ratings for each of the 360 photographs used by Dunsmoor et al. (2015). As simply adopting their allocation of photographs to encoding phases would have led to unequal typicality across phases, we performed a total of 6 swaps to ensure that typicality would be comparable across phases (see main text for details). We tested the resulting stimulus sets for any systematic differences in typicality using a repeated measures ANOVA with object category (i.e., animals vs. tools) and encoding phase as within-subject factors (Supplementary Figure 7). Mauchly's Test indicated that the sphericity assumption was violated for both the factor encoding phase ($W = 0.75, p = .049$) as well as for the interaction between encoding phase and object category ($W = 0.51, p < .001$). Therefore, a Greenhouse-Geisser correction was applied for these effects. Results showed that photographs of animals were generally rated as more typical than photographs of tools, $F(1, 40) = 118.78, p < .001, \eta^2_G = .38$. Critically, there was neither a significant main effect of encoding phase, $F(2.61, 104.51) = 0.90, p = .43, \eta^2_G = .0003$, nor a significant interaction between encoding phase and object category, $F(2.15, 85.98) = 0.12, p = .90, \eta^2_G = .00005$.



Supplementary figure 7. Mean typicality ratings of photographs used in Experiment 4 by encoding phase and object category. Error bars show ± 1 standard error of the mean. An

independent sample of 41 participants rated how typical each photograph was for its respective category (i.e., as an animal or tool) on a scale from 1 ('very untypical') to 10 ('very typical'). Note that in Experiment 4, it was randomized across participants which of the three encoding sets would be allocated to which of the three encoding phases (pre-conditioning, conditioning, and post-conditioning).

Mixed-effect analysis

We again applied the same GLMM as in Experiments 2 and 3 as an alternative analysis of the effects of fear conditioning on recognition performance across encoding phases.

In line with previous results, we found a memory advantage for CS⁺ items that were encoded during fear-conditioning (i.e., a significant interaction between conditioning category and the fear-conditioning phase), $\beta = 0.53$, 95% *CI* [0.38, 0.68], $z = 7.01$, $p < .001$. As in Experiment 3, we again found a non-significant trend towards improved memory for CS⁺ items that were encoded in the post-conditioning phase (i.e., an interaction between conditioning category and the post-conditioning phase), $\beta = 0.13$, 95% *CI* [-0.01, 0.27], $z = 1.77$, $p = .076$. Most importantly and in line with all three previous experiments, we found no evidence for category-specific retroactive memory enhancement (i.e., no interaction between conditioning category and the pre-conditioning phase), $\beta = 0.09$, 95% *CI* [-0.05, 0.24], $z = 1.22$, $p = .22$.

Response Bias

Supplementary Table 4

Comparison of the response bias c between conditioning categories in Experiment 4

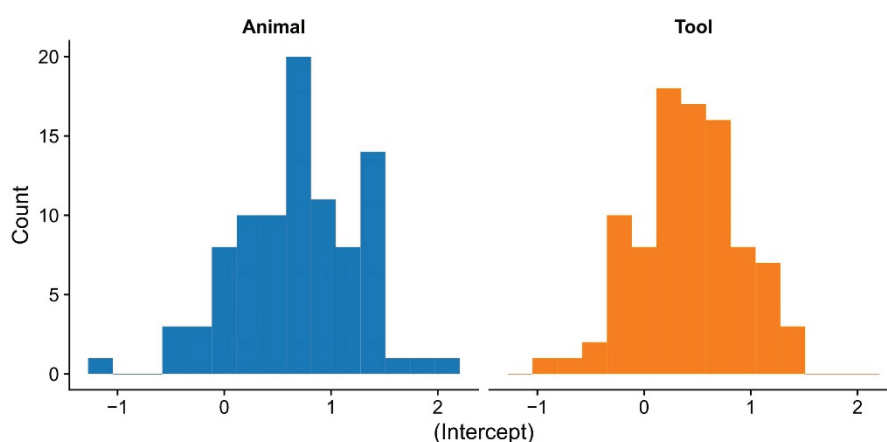
	Collapsed across confidence				Only high confidence			
	Pre-Conditioning	Conditioning	Post-Conditioning	Pre-Conditioning	Conditioning	Post-Conditioning	Pre-Conditioning	Post-Conditioning
$M (SD); CS^+$	0.114 (0.360)	-0.013 (0.400)	0.185 (0.343)	0.869 (0.436)	0.755 (0.472)	0.944 (0.435)		
$M (SD); CS^-$	0.196 (0.352)	0.212 (0.390)	0.275 (0.352)	0.966 (0.430)	1.008 (0.458)	1.033 (0.453)		
d_{av}	-0.233	-0.567	-0.258	-0.223	-0.544	-0.201		
$t(82)$	-1.667	-3.940	-1.936	-2.065	-5.042	-2.028		
p	.099	< .001	.056	.042	< .001	.046		

Note: Smaller values of c indicate a more liberal response bias, while larger values of c indicate a more conservative response bias. d_{av} , t , and p refer to the comparison of the response bias c between items from the CS^+ vs. CS^- category. Negative values of d_{av} indicate a bias towards more conservative responses for items from the CS^- category, while positive values of d_{av} indicate a bias towards more conservative responses for items from the CS^+ category.

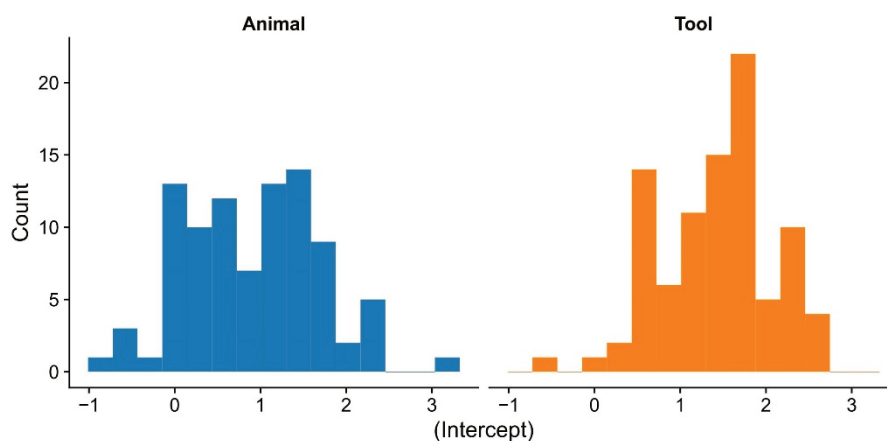
Item analysis

We ran parallel GLMMs as in the three previous experiments to examine whether certain items were remembered at particularly high rates or associated with increased false alarm probabilities. Supplemental Figure 8 shows the distribution of per-item intercepts and photograph category membership (i.e., animals vs. tools) for previously seen items.

Supplemental Figure 9 shows the same distribution for ‘new’ items.



Supplementary Figure 8. Distribution of per-item intercepts in Experiment 4 reflecting the probability of correctly classifying previously seen photographs as ‘old’ in the recognition test.



Supplementary Figure 9. Distribution of per-item intercepts in Experiment 4 reflecting the probability of correctly classifying previously unseen photographs as ‘new’ in the recognition test.

References

- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48.
<https://doi.org/10.18637/jss.v067.i01>

Appendix B: Study 2

B

Kalbe, F., & Schwabe, L. (2020). Beyond arousal: Prediction error related to aversive events promotes episodic memory formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 234-246. <https://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).

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 ex-

pectations 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 down-sampled to a resolution of 50 Hz and optimized using four sets of

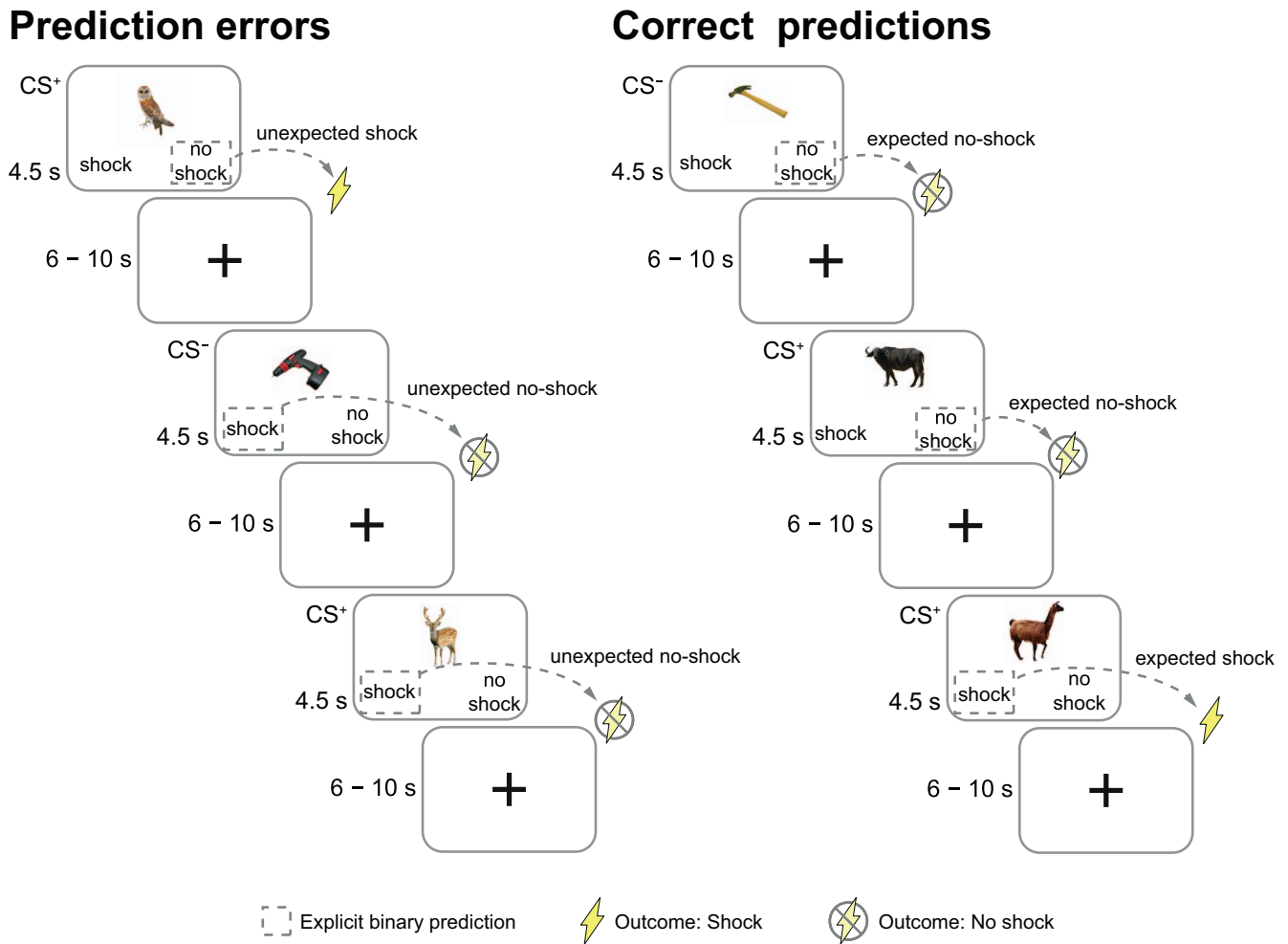


Figure 1. Task design. In each of the 60 trials of a combined incidental learning and fear conditioning task, participants saw a picture of an animal or a tool and predicted whether they would receive an aversive electric shock or not. One of the two stimulus categories (animals or tools) was randomly 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 \times 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 ($M = 0.44 \mu S, SD = 0.21 \mu S$) versus CS^- items ($M = 0.42 \mu S, SD = 0.24 \mu S$), 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

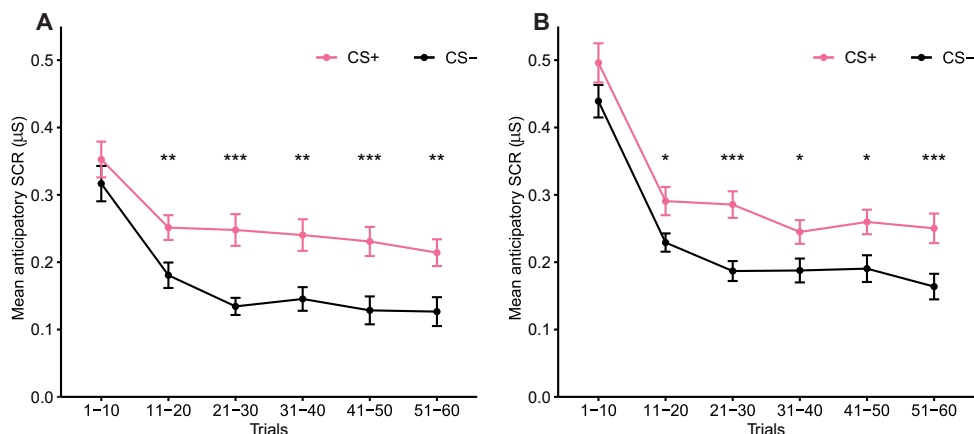


Figure 2. Average anticipatory skin conductance responses by block and condition. Apart from the first 10 trials, anticipatory skin conductance responses were always significantly higher for items from the conditional stimulus (CS^+) category compared with items from the CS^- category in both Experiment 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⁺), $\phi = .38$, $p < .001$. 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-

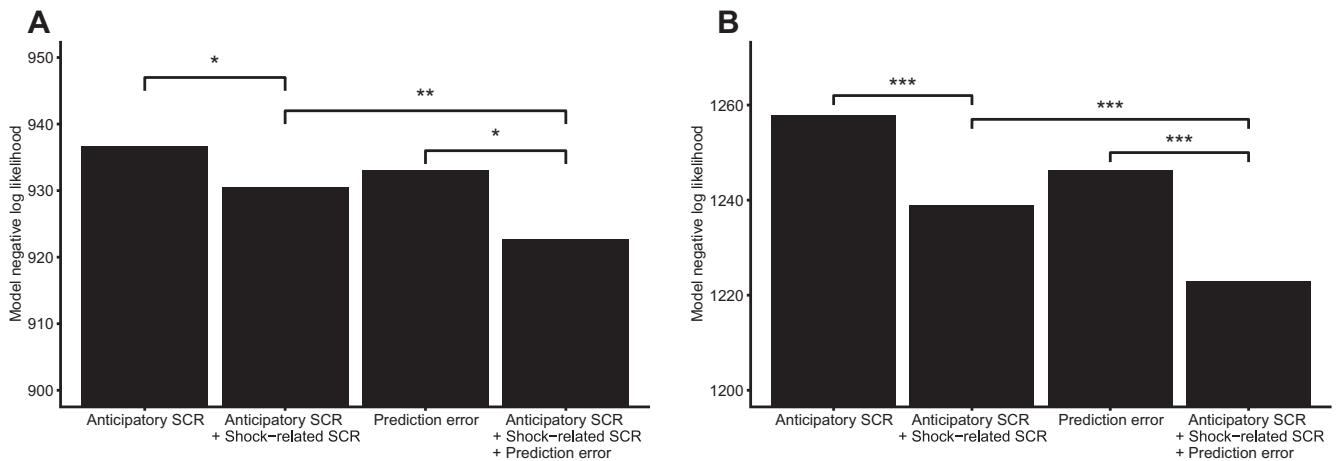


Figure 3. Generalized linear mixed-model (GLMM) fit indices for models with different sets of independent variables to predict the binary recognition of an item in the incidental learning paradigm in Experiment 1 (A) and Experiment 2 (B). Smaller values indicate a better model fit. Notably, the best-fitting 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 $\alpha = .05$. 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, $F(1, 79) = 19.10$, $p < .001$, $\eta_G^2 = .018$, as well as the block, $F(5, 395) = 40.10$, $p < .001$, $\eta_G^2 = .105$. There was no significant interaction between condition and block, $F(5, 395) = 0.62$, $p = .68$, $\eta_G^2 = .0001$. At descriptive level, however, we found that within the first five trials, there was almost no difference in mean anticipatory SCRs between CS⁺ items ($M = 0.60 \mu S$, $SD = 0.33 \mu S$) compared with CS⁻ items ($M = 0.61 \mu S$, $SD = 0.28 \mu S$), 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, $\phi = .44, p < .001$. 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 (Spoomaker 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, Kryptos, 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](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](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](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.biopsycho.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](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](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/nrn.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](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](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 ■

Appendix C: Study 3

C

Kalbe, F., & Schwabe, L. (2021). Prediction errors for aversive events shape long-term memory formation through a distinct neural mechanism. *bioRxiv*. <https://doi.org/10.1101/2021.03.19.436177>

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

Running title: How aversive PEs shape long-term memory

Felix Kalbe¹ and Lars Schwabe¹

¹Institute of Psychology, Universität Hamburg, Von-Melle-Park 5, 20254 Hamburg, Germany

Corresponding author: L.S. (lars.schwabe@uni-hamburg.de)

ABSTRACT

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 distinct neural mechanism that might serve to set these memories apart from those with expected outcomes.

Keywords: arousal, associative learning, medial temporal lobe, salience network, schema network

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 (Glimcher, 2011; Niv, 2009). 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 (Gershman & Daw, 2017; Shohamy & Adcock, 2010), PEs received little attention in episodic memory research (for early exceptions, see Henson & 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 (Ergo et al., 2020; Greve et al., 2017; Jang et al., 2019; Kalbe & Schwabe, 2020; Rouhani et al., 2018). A fundamental question concerns how PEs boost long-term memory formation.

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 & 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, Hermans et al., 2014; McGaugh & Roozendaal, 2002; Richardson et al., 2004; Strange & Dolan, 2004). 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 critically 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 arousal (Kalbe & 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 (Metereau & Dreher, 2013; Summerfield & Egnér, 2009) may be mediated by the salience network (Fouragnan et al., 2018; Ham et al., 2013), 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 & 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 in each trial to the actual trial outcome indicated whether participants experienced a negative PE (unexpected shock omission), a positive PE (unexpected shock), or no PE at all (Delgado et al., 2008; McHugh et al., 2014; Schultz, 1998). Memory was probed in a recognition test 24 hours after encoding. To unravel the neural mechanisms underlying PE-related enhancements of episodic memory, we used behavioral modelling, 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 \pm SD=24.97 \pm 4.65 years) participated in this experiment. Eleven participants had to be excluded from the analysis due to excessive head motion in the scanner (>5mm 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 & 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 & 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 .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 (Figure 1A). About 24 hours 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, USA) 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, USA). Another pair of electrodes was placed on the right side of the right lower leg, approximately 20cm 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, USA) connected to a BIOPAC STM100C stimulator (BIOPAC systems, Goleta, CA, USA). After participants were placed in the scanner, they first completed an unrelated task that included stimuli that were critically 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 4.5s (Figure 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 of an MRI-compatible response box (Current Designs Inc., Philadelphia, USA). 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 200ms shock to the right lower leg followed immediately after image offset. Between trials, there was a jittered white fixation cross presented for 5s to

8s. This relatively long inter-trial interval allowed us to observe the slowly emerging SCR in response 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 1/3 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. Between blocks, participants had the opportunity to ask the experimenter to slightly reduce the shock intensity in cases when shocks had become painful.

After an interval of 22h to 26h, 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). For each image, participants had a maximum of 6s 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 to 2s.

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 2mm depth, slice orientation parallel to the AC-PC line, phase-encoding in AP direction, repetition time (TR) of 2000ms, echo time (TE) of 30ms, 60-degree flip angle, 224mm × 224mm field of view (FOV), 2mm isotropic resolution, EPI factor of 112, echo distance of 0.58ms. 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-weighted scan was acquired with 256 slices, coronal orientation, repetition time (TR) of 2300ms, echo time (TE) of 2.12ms, a 240mm x 240mm field of view (FOV), and a 0.8mm × 0.8mm × 0.9mm 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:

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

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:

$$sPE(t) = O(t) - P(t)$$

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:

$$PU(t) = 1 - \left| |sPE(t)| - 0.5 \right| \times 2$$

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.

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 & 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 4.5s after stimulus onset). Additionally, we defined outcome-related SCRs to occur 0.5s after the outcome of the current trial was revealed (i.e., whether a shock would occur or not) until 2.9s 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 1000Hz to 50Hz and optimized using four sets of initial values. The minimum amplitude threshold was set to 0.01 μ S 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 & 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.

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 8mm full-width half-maximum Gaussian kernel. For multivoxel pattern analysis, the unsmoothed normalized images were used.

Univariate fMRI analyses

Based on results from behavioral modelling, we identified (1) quadratic prediction uncertainty, (2) quadratic signed PEs, and (3) 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 modelled 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 vs. 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 1/128Hz 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 $p_{FWE} < .05$.

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 vs. 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 & 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 tests using Bonferroni correction. 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).

Multivoxel pattern analyses (MVPA)

To test which of the regions identified in the univariate analysis contained pattern information about (1) the extent of PEs and (2) the probability that an encoded item would later be recognized, we performed a multivoxel pattern analysis (MVPA; Kriegeskorte, 2011). As PEs are continuous, decoding them from neural data constitutes a regression problem, while the later recognition of an item is binary, and its decoding therefore constitutes a classification problem. Hence, these problems required slightly different machine learning algorithms, although both were selected from the class of support vector machines and the general data preparation and model fitting procedure was very similar for both problems.

The MVPA was performed on t-maps that were generated in SPM12 from the unsmoothed, normalized functional data from each participant. Separate generalized linear models were estimated to extract several trial-specific t-maps of each of the following points in time relative to each outcome: -4, -2, 0, 2, 4. Extracting multiple activation maps per trial in this way allowed us to address specifically the question when exactly relevant pattern information was present. Note that stimulus onsets were always 4.5s before outcome onsets and were therefore represented by the offset -4. Also note that the temporal distance of 2s between offsets corresponds with the TR of the EPI sequence. In each GLM, each single onset of the outcome event, offset by the currently estimated point in time, was entered as its own regressor. Therefore, for each trial and offset, we generated unique beta-maps, which were then transformed to t-maps to normalize them (Misaki et al., 2010). The GLM used the same parameters as in the univariate analyses, namely, concatenation of experimental blocks, a high-pass filter at 1/128Hz, and an AR(1) process to adjust for temporal autocorrelation.

T-maps representing individual trials and offsets per participant were then further processed in Python 3 using the Nilearn module (Abraham et al., 2014). Whole-brain t-maps were masked with ROIs identified in the univariate analysis. Specifically, in the case of larger regions (e.g., insula), we created new masks by identifying the peak voxel per region from the second level analysis of univariate results reported earlier and including voxels within a 6mm-radius of each peak voxel. For the bilateral hippocampus, we used existing anatomical masks (Harvard-Oxford structural atlas; Desikan et al., 2006).

To prepare extracted data from each ROI for use with common machine learning algorithms, the 4-dimensional t-maps (three spatial and one temporal dimension) were reshaped to a samples-by-features matrix (number of trials \times number of voxels in ROI). Further, data were z-standardized using the StandardScaler implementation in scikit-learn (Pedregosa et al., 2011). To predict PEs from neural data, we trained a support vector regression (SVR) with a linear kernel as implemented in scikit-learn with the regularization

hyperparameter C fixed at 1 and the negative mean squared error as the performance metric. Similarly, to predict the binary recognition of an item, we trained a linear support vector classifier using the LinearSVC implementation in scikit-learn with the regularization parameter C fixed at 1 and the area under the receiver operating characteristic (ROC) curve as the performance metric to account for imbalanced classes due to uneven numbers of hits and misses for each participant. For both decoding tasks, we used leave-one-block-out cross-validation to evaluate decoding performance, such that three blocks were always used for training and the remaining block was used for validation. Performance metrics from all four possible training-validation combinations were averaged to compute the mean performance.

To establish a baseline performance at chance level that can be used to compare each fitted model against, for each “true” performance score, we also performed the exact same preprocessing and training procedure using 100 separate random permutations of the true labels as a permutation test (Nichols & Holmes, 2002). Therefore, the above-chance-performance of a predictive model could be conceptualized here as the distance between the performance achieved with true labels and the mean performance in the permutation test.

Large-scale network-connectivity analyses

We performed analyses of functional connectivity in the CONN toolbox (Whitfield-Gabrieli & 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 ($|sPE| < 0.5$) vs. 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 vs. outcome onsets, shocked vs. unshocked, and low vs. high PEs. This resulted in a total of 8 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 to 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: (i) dorsal anterior cingulate cortex, bilateral anterior insula, bilateral rostral prefrontal cortex and bilateral supramarginal gyrus forming the *salience network* (Menon, 2011); (ii) medial prefrontal cortex, bilateral angular gyrus and precuneus forming the schema network (van Kesteren et al., 2012; Vogel et al., 2018a); and (iii) 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, $F(2,98)=3.62$, $p=.030$, $\eta_G=.011$ (Figure 1B). Post-hoc paired t-tests revealed that participants showed increased anticipatory SCRs to both CS^{a+} pictures ($t(49)=2.38$, $p_{corr}=.042$ (Bonferroni-corrected), $d_{av}=0.27$) and CS^{b+} pictures compared with CS^- pictures ($t(49)=2.10$, $p=.041$, $p_{corr}=.082$ (Bonferroni-corrected), $d_{av}=0.20$). Explicit shock ratings further showed that participants learned to associate picture categories with their respective shock probabilities over the course of the task (Figure 1C). Participants had a significantly higher shock expectancy for CS^{a+} than for CS^{b+} ($t(49)=11.53$, $p_{corr}<.001$ (Bonferroni-corrected), $d_{av}=2.67$) and for CS^{b+} than for CS^- ($t(49)=10.81$, $p_{corr}<.001$ (Bonferroni-corrected), $d_{av}=1.87$).

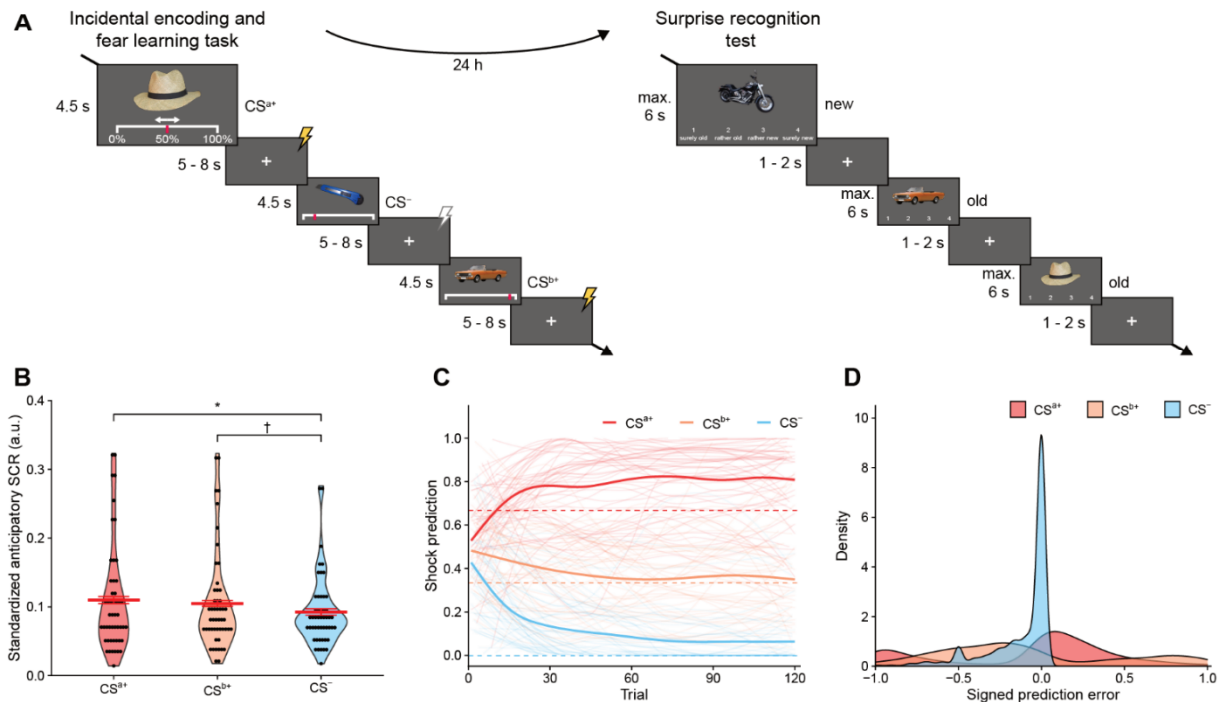


Figure 1. Experimental task and performance parameters

(A) Participants completed a combined incidental encoding and fear learning task and a surprise recognition test for its contents about 24h 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 distributed relatively symmetrical for CS^{a+} and CS^{b+} pictures around zero. PEs for CS⁻ pictures were mostly zero, reflecting that participants learned that items from this category were never paired with a shock.

† $p < .05$, * $p_{\text{corr}} < .05$ (Bonferroni-corrected).

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 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 varied almost symmetrically around 0 (Figure 1D), allowing similarly reliable conclusions about effects of both negative PEs and positive PEs. 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 hours 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'), $M_{\text{hitrate}}=60.9\%$ ($SD=0.149$), $M_{\text{FArate}}=21.1\%$ ($SD=0.098$). Participants were significantly more certain with their responses for hits ($M=0.59$; $SD=0.18$) than for false alarms ($M=0.26$, $SD=0.20$), $t(49)=15.92$, $p<.001$, $d_{\text{av}}=1.70$.

A repeated-measures ANOVA showed that hit rates differed significantly between CS categories, $F(2,98)=7.29$, $p=.001$, $\eta^2=0.05$. For false alarm rates, on the other hand, there was no such difference between CS categories, $F(2,98)=0.25$, $p=.77$, $\eta^2=.003$, 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 ($t(49)=4.15$, $p_{\text{corr}}<.001$ (Bonferroni-corrected), $d_{\text{av}}=0.54$), which was associated with a shock probability of 33%, and the CS^- category ($t(49)=2.64$, $p_{\text{corr}}=.022$ (Bonferroni-corrected), $d_{\text{av}}=0.40$; Figure 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 ($F(2,98)=3.70$, $p=.028$, $\eta^2=.03$), with

post-hoc t-tests confirming an increased memory sensitivity for CS^{a+} items compared with both CS^{b+} items ($t(49)=2.28$, $p=.027$, $p_{\text{corr}}=.053$ (Bonferroni-corrected), $d_{\text{av}}=0.38$) and CS⁻ items ($t(49)=2.42$, $p_{\text{corr}}=.038$ (Bonferroni-corrected), $d_{\text{av}}=0.35$).

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 ($F(1,49)=1.12$, $p=.294$, $\eta^2=.022$). Further, a 2x2 repeated-measures ANOVA confirmed increased hit rates for CS^{a+} over CS^{b+} items even after controlling for shocks ($F(1,49)=19.47$, $p<.001$, $\eta^2=.08$). Notably, this ANOVA even showed a tendency towards decreased hit rates for shocked items ($F(1,49)=3.76$, $p=.058$, $\eta^2=.006$), with no significant interaction ($F(1,49)<0.001$, $p=.997$, $\eta^2<.0001$). These findings indicate that differences between CS categories in the number of presented shocks cannot explain the differential memory performance and that other factors drive the boost in memory.

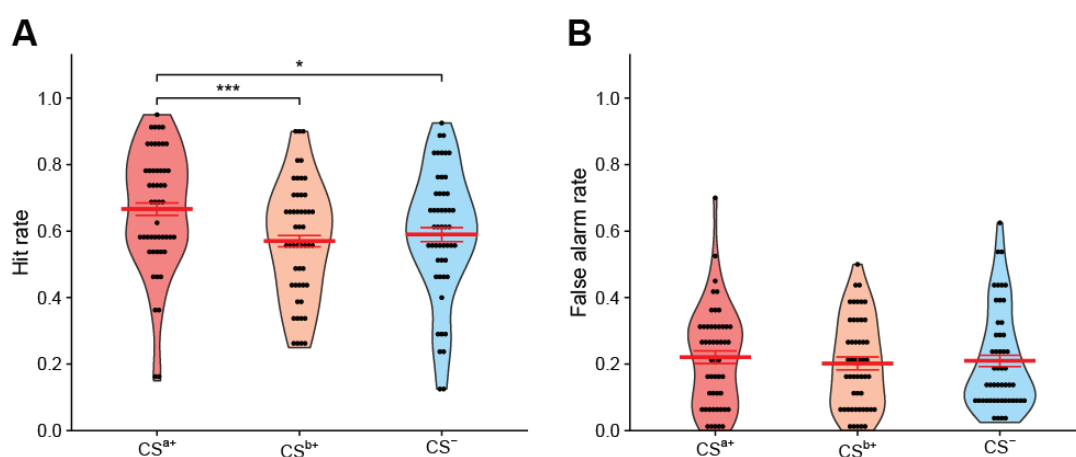


Figure 2. General recognition performance by CS category.

(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 (Supplemental Figure 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.

* $p_{\text{corr}} < .05$ (Bonferroni-corrected), *** $p_{\text{corr}} < .001$ (Bonferroni-corrected)

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.

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 assuming that effects of negative vs. positive PEs would 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). 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], $z=2.68$, $p=.007$. Even after adding the binary occurrences of shocks to the model, this effect remained significant ($\beta=0.58$, 95%-CI [0.29, 0.87], $z=3.30$, $p<.001$), rejecting the notion that aversive shocks alone drive this effect. Further, 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], $z=3.01$, $p=.002$, 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 (Figure 3A). For negative PEs, we again observed a memory enhancement with greater PE magnitude, $\beta=0.49$, 95%-CI [0.15, 0.82], $z=2.85$, $p=.004$. 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], $z=3.67$, $p<.001$. Random β s 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, $r=-.395$, $t(48)=2.98$, $p=.005$ (Figure 3B).

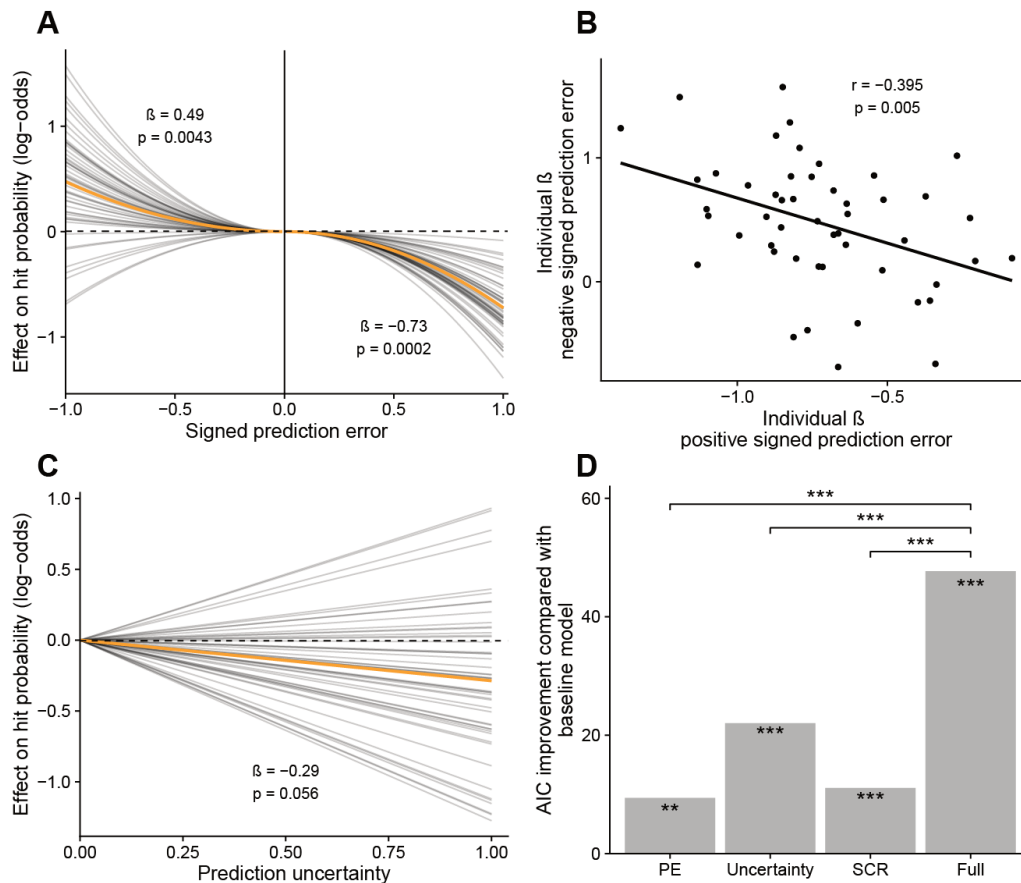


Figure 3. Behavioral model of long-term memory formation reveals modulating influences of prediction errors and prediction uncertainty

(A) 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.

(B) Effects of quadratic negative and positive PEs were negatively correlated at the level of participants.

(C) Prediction uncertainty was, independently of outcomes, associated with decreased hit probabilities. As in (A), the orange line indicates the estimated fixed effect, while thin black lines show participant-specific effect estimates.

(D) Model comparisons showed that a full model combining PEs, uncertainty, and anticipatory and outcome-related SCRs explained memory formation better than any model containing only one of these measures. This was confirmed by both likelihood-ratio tests as well as the lowest (i.e., best) AIC value for this model. Notably, any model containing only a single predictor (i.e., either PEs, uncertainty, or SCRs) also performed significantly better than our baseline model comprised only of any a random intercept per participant (comparisons indicated by markings within each bar).

** $p < .01$, *** $p < .001$.

An alternative explanation for the memory-modulating effects of PEs could be that the mere expectation of an aversive shock drives the effect on long-term memory. If this were the case, then contrasting the explicit shock predictions with observed outcomes should not explain memory formation beyond the influence of mere predictions. To test this alternative account, we first fitted a mixed-effects logistic regression model with participants' explicit shock predictions (ranging from 0 to 1) to explain the binary recognition of an item. Indeed, greater expectancy of a shock was associated with an increased probability that the item would subsequently be recognized, $\beta=0.40$, 95%-CI [0.12, 0.67], $z=2.83$, $p=.005$. After adding the subsequent binary occurrence of shocks, this positive effect of predictions remained ($\beta=0.44$, 95%-CI [0.16, 0.73], $z=3.08$, $p=.002$), while shocks themselves were not found to affect the probability that an item would be recognized ($\beta=-0.04$, 95%-CI [-0.20, 0.13], $z=0.46$, $p=.65$). To test whether PEs could explain memory formation beyond main effects of predictions and outcomes, we added them to the model using the inverted S-shaped transformation reported above. In contrast to the previous models, this model no longer indicated any memory-modulating effects of mere predictions, $\beta=-0.67$, 95%-CI [-1.53, 0.20], $z=1.50$, $p=.13$. However, aversive shocks were now associated with increased hit probabilities, $\beta=0.83$, 95%-CI [0.18, 1.47], $z=2.51$, $p=.012$. Most importantly, PEs again explained memory formation depending on their sign, $\beta=1.22$, 95%-CI [0.37, 2.08], $z=2.81$, $p=.004$. Further evidence that the model adding both outcomes and PEs fitted the data better compared with the prediction-only model came from a likelihood-ratio test, $\chi^2(9)=26.72$, $p=.002$. This was also reflected in a smaller AIC value for the model adding both outcomes and PEs (AIC=7366.7) compared with the prediction-only model (AIC=7375.4). Finally, as reported above, we did not find any improved recognition performance for items from the CS^{b+} category over items from the CS⁻ category, even though these were associated with significantly higher shock expectancies.

Closely related to both predictions and PEs, uncertainty about possible outcomes has been proposed to affect episodic memory formation as well (Stanek et al., 2019). In contrast to shock predictions, uncertainty is maximal when participants believe that the probability of a shock is 50% (i.e., maximal entropy). As before, we investigated effects of prediction uncertainty on memory formation using two mixed-effects logistic regression models with (1) linear uncertainty and (2) quadratic uncertainty as the sole independent variable to explain the binary recognition of an item. Because prediction uncertainty is independent of outcomes, these models included both unshocked and shocked trials. Results favored a linear model (AIC = 7408.2) over the quadratic model (AIC = 7418.2) and suggested a tendency towards a negative relationship between prediction uncertainty and memory formation for the associated item, $\beta=-0.29$, 95%-CI [-0.58, 0.008], $z=1.91$, $p=.056$ (Figure 3C). In other words, when participants were more uncertain in terms of their shock prediction, it tended to be less likely that they would later recognize the associated item.

Classic models of episodic memory formation in aversive contexts emphasized the memory promoting role of physiological arousal (Cahill & McGaugh, 1998; McGaugh, 2018). Therefore, we tested in a next step whether the PE-related memory changes that we observed here could be explained by physiological arousal. To this aim, we fitted a mixed-effects logistic regression model with standardized anticipatory SCRs and standardized outcome-related SCRs as the only two predictors for the binary recognition of an item. In this model, neither anticipatory SCRs ($\beta=-0.14$, 95%-CI [-0.58, 0.30], $z=0.63$, $p=.531$), nor outcome-related SCRs ($\beta=0.26$, 95%-CI [-0.11, 0.64], $z=1.37$, $p=.170$), had any significant effect on memory, suggesting that physiological arousal (expressed through SCR) did not drive long-term memory formation.

In a final mixed-effects logistic regression model, we included quadratic PEs, quadratic prediction uncertainty, anticipatory and outcome-related SCR in parallel to investigate whether previous results from simpler models would still hold after accounting for

other memory-modulating variables. Since this model was again applied to trials including both negative and positive PEs, we again entered PEs using the previously introduced inverted S-shaped transformation. Results confirmed our previous findings that PEs had memory-promoting effects in the case of unexpected shock omissions and memory-decreasing effects in the case of unexpected shocks, $\beta=0.44$, 95%-CI [0.20, 0.67], $z=3.62$, $p<.001$. Further, this combined model confirmed our previous findings of memory decreasing effects of prediction uncertainty, $\beta=-0.39$, 95%-CI [-0.68, -0.09], $z=2.56$, $p=.010$. As before, standardized anticipatory SCRs had no significant effect on episodic memory formation, $\beta=-0.17$, 95%-CI [-0.62, 0.29], $z=0.72$, $p=.47$. Notably, unlike in the simpler SCR model, outcome-related SCRs showed a positive effect on memory formation, $\beta=0.63$, 95%-CI [0.18, 1.08], $z=2.76$, $p=.006$. Therefore, only after accounting for effects of PEs and uncertainty on memory formation, additional arousal-related influences occurred.

Separate model comparisons using likelihood ratio tests confirmed that the full model including PEs, uncertainty, and physiological arousal (measured by both anticipatory and outcome-related SCRs) was the most appropriate. This suggests that all three components uniquely and additively contribute to long-term memory formation. Compared with a simple baseline model containing only a random intercept for each participant, adding any type of predictor from the full model (i.e., either PEs, uncertainty, or physiological arousal) significantly improved the fit (all $ps<.002$; Figure 3D). Critically, the full model containing all three types of predictions led to the lowest AIC value. The full model also significantly improved the fit compared with any model containing only a single type of predictor (all $ps<.001$; Figure 3D).

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 modelled 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 Methods). Based on the rich literature linking the medial temporal lobe with episodic memory formation (Alvarez & 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, $t(49)=4.90$, $p_{\text{svc}}=.001$ (FWE-corrected), $p_{\text{svc}}=.002$ (FWE- and Bonferroni-corrected; Figure 4), right posterior parahippocampal gyrus, $t(49)=4.16$, $p_{\text{svc}}=.006$ (FWE-corrected), $p_{\text{svc}}=.012$ (FWE- and Bonferroni-corrected) and, at trend level, in the right hippocampus, $t(49) = 3.89$, $p_{\text{svc}}=.031$ (FWE-corrected), $p_{\text{svc}}=.062$ (FWE- and Bonferroni-corrected) .

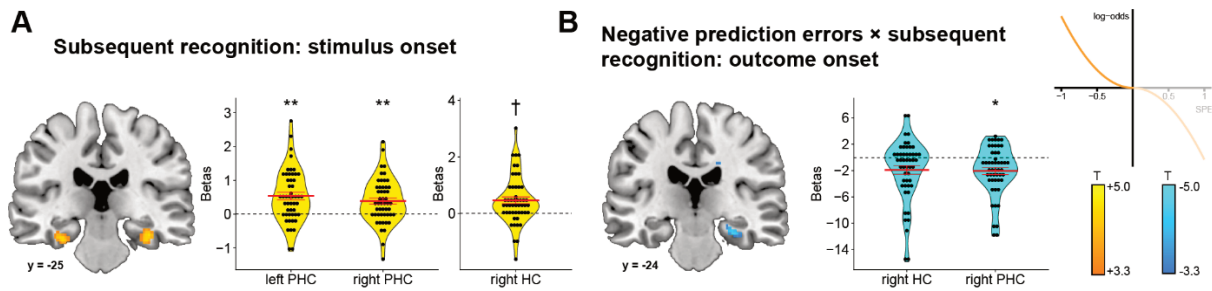


Figure 4. Univariate fMRI analysis of subsequent memory

(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 $p < .001$ (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. † $p_{\text{svc}} < .05$ (FWE-corrected) * $p_{\text{svc}} < .05$ (FWE- and Bonferroni-corrected), ** $p_{\text{svc}} < .01$ (FWE- and Bonferroni-corrected).

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 $p < .05$, 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; Figure 5A-B). In addition, negative PEs were associated with significant decreases in activation in large portions of the bilateral hippocampus and parahippocampal gyrus (Figure 5D). 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 first evidence that the PE-induced

memory enhancement that we observed here might involve a neural mechanism that is critically 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 (Figure 5C-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 first 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.

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 (Supplemental Figure 2A). Additionally, we observed decreased activation in the bilateral middle temporal gyrus (Supplemental Figure 2B), 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 (Supplemental Table 3). 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.

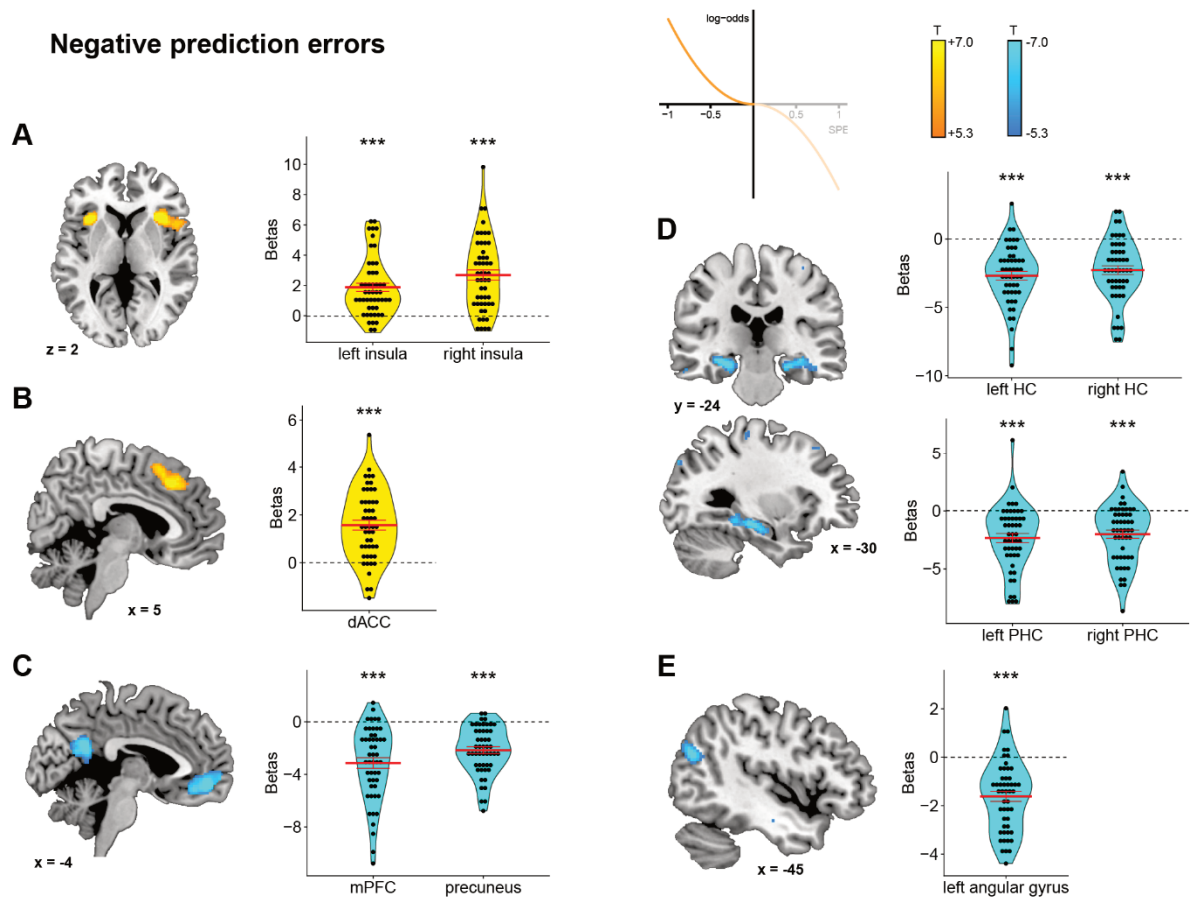


Figure 5. Univariate fMRI analysis to identify regions associated with negative PEs

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 $p < .05$ 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. *** $p_{FWE} < .001$.

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 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, $t(49)=3.87$, $p_{\text{svc}}=.015$ (FWE-corrected), $p_{\text{svc}}=.030$ (FWE- and Bonferroni-corrected; Figure 4B). Additionally, there was a similar non-significant trend in right hippocampus, $t(49)=3.65$, $p_{\text{svc}}=.062$ (FWE-corrected), $p_{\text{svc}}=.124$ (FWE- and Bonferroni-corrected). 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 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; Ham et al., 2013; Menon, 2011; Metereau & Dreher, 2013; Seeley et al., 2007), (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 8 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 Methods), we based our analysis on the implemented network atlas consisting of several ROIs each to compute within- and between-network correlations (Figure 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 vs. small negative PEs we obtained significantly increased functional connectivity between the salience network and both the schema network ($t(49)=2.68$, $p_{\text{corr}}=.030$ (Bonferroni-corrected), $d_{\text{av}}=0.344$) and, at trend level, the medial-temporal encoding network ($t(49)=2.18$, $p=.034$, $p_{\text{corr}}=.10$ (Bonferroni-corrected), $d_{\text{av}}=0.355$; Figure 6B); the connectivity between the schema network and the medial-temporal network did not depend on PEs in unshocked trials, $t(49)=0.29$, $p=.773$, $p_{\text{corr}}=1$ (Bonferroni-corrected), $d_{\text{av}}=0.046$). 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, $r=0.320$, $t(48)=2.34$, $p_{\text{corr}}=.048$ (Bonferroni-corrected; Figure 6C); salience-MTEN correlation with hit rates for high negative PE items: $r=0.147$, $t(48)=1.03$, $p=.31$, $p_{\text{corr}}=.616$ (Bonferroni-corrected). Furthermore, within-network connectivity tended to be decreased for large compared with small negative PEs in the medial-temporal encoding network ($t(49)=2.44$, $p=.018$, $p_{\text{corr}}=.055$ (Bonferroni-corrected), $d_{\text{av}}=0.307$), but not in the salience network ($t(49)=1.60$, $p=.115$, $p_{\text{corr}}=.346$ (Bonferroni-corrected) $d_{\text{av}}=0.218$), nor in the schema network ($t(49)=1.24$, $p=.221$, $p_{\text{corr}}=.664$ (Bonferroni-corrected), $d_{\text{av}}=0.221$; Figure 6D).

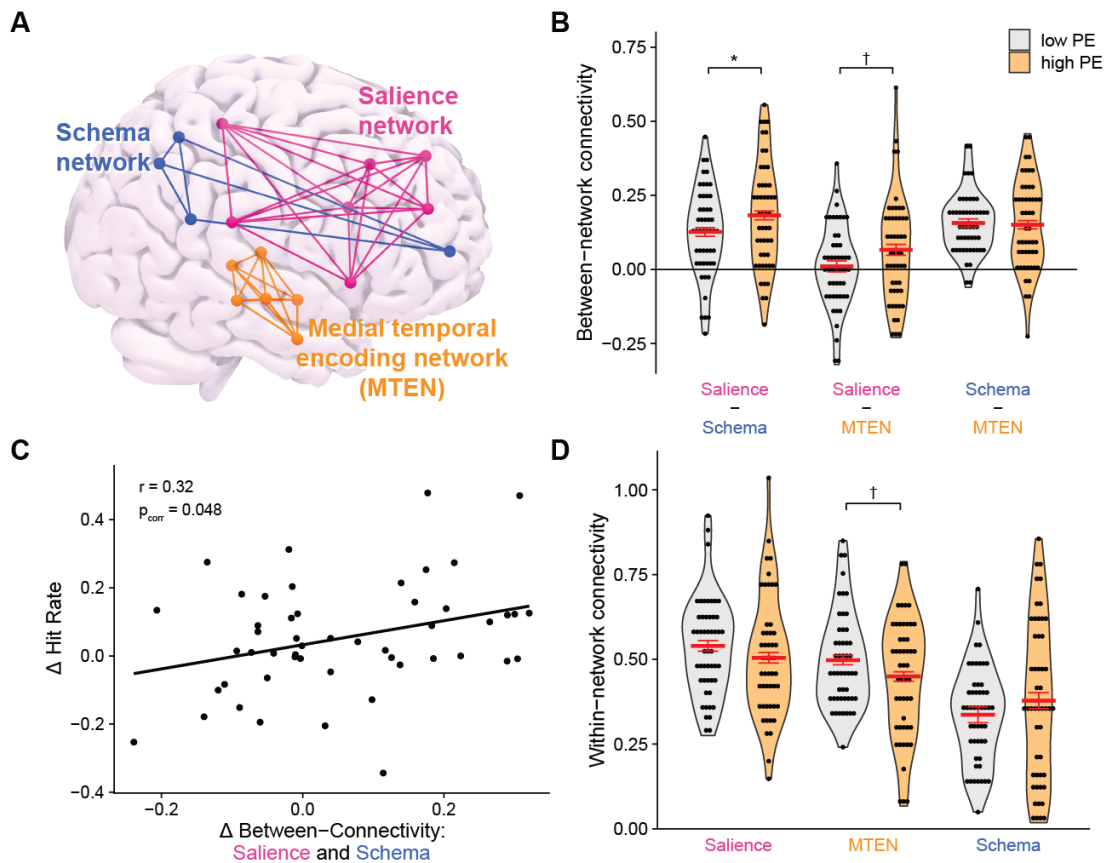


Figure 6. Negative prediction error magnitude is associated with altered within- and between-network connectivity in memory-relevant networks

(A) We investigated PE-associated changes in the activity within and between the salience network (rostral prefrontal cortex, supramarginal 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.

† $p < .05$, * $p_{\text{corr}} < .05$ (Bonferroni-corrected).

Activity patterns predictive for both PEs and item recognition

Our above analyses showed increased activity for negative PEs in regions of the salience network (dACC and bilateral anterior insula). Regions of the schema network (mPFC, angular gyrus, and precuneus) and the medial temporal lobe, however, showed

decreases in activation with larger negative PEs, with the latter being directly linked to improved subsequent recognition. To further elucidate the mechanism through which (negative) PEs facilitate memory formation, we used multivoxel pattern analysis based on activity patterns of areas identified in the univariate analysis to investigate whether a single region would contain both (1) pattern information that can be used to decode PEs and (2) pattern information that predicts subsequent recognition memory (see Methods).

Our results showed that regions associated with negative PEs in the univariate analysis also contained pattern information that enabled us to decode the magnitude of negative PEs (see Supplemental Figure 3). Although PEs were decoded above chance already before the outcome presentation which may be due to uncertainty effects, PEs could be decoded best around the time window when the outcome of a trial was revealed and shortly thereafter. The subsequent recognition of an item could only be decoded significantly above chance level with patterns of activity from the insula that occurred around the time the outcome (and by implication, the PE) of trial was revealed (see Supplemental Figure 3). However, this effect did not survive a correction for multiple comparisons and therefore needs to be interpreted with great caution.

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 supplemental table S1).

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 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 $p_{\text{svc}} > .05$, FWE corrected). Even under a very liberal threshold of $p < .001$ (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 $p_{\text{FWE}} > .05$).

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 $p_{\text{FWE}} > .05$). 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 (Supplemental Table 3). 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 (Cohen, 2008; Schultz, 1998; Sutton & Barto, 1981). 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 modelling 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 is 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 (Davachi & Wagner, 2002; Eichenbaum, 2004; Fernández et al., 1999; Mayes et al., 2007; Reed & Squire, 1997; 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 (Bastin et al., 2016; Fazeli & Büchel, 2018; Garrison et al., 2013; Preuschoff et al., 2008; Taylor et al., 2007; Ullsperger et al., 2010). Moreover, both the anterior insula and the dACC are key

regions of the salience network (Ham et al., 2013; Menon, 2011), which signals biologically relevant events and the need for a behavioral or cognitive change (Dosenbach et al., 2006; Kerns, 2004). 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 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 & 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. Moreover, only in a combined model featuring additionally uncertainty and PEs, larger outcome-related SCRs were linked to improved recognition performance. Even in this combined model, we still found clear evidence for complementary effects of PEs (and uncertainty) beyond 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. Further, although greater shock expectancy was by itself linked to enhanced memory formation, we found that effects of PEs, which contrast such expectations with observed outcomes, explained recognition beyond main effects of shock expectations and observed outcomes. 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 (Cycowicz & Friedman, 2007; Strange & Dolan, 2004). 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 & 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 (Ergo et al., 2020; Jang et al., 2019; Rouhani et al., 2018). 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 (Starkweather et al., 2018; Yoshida & Ishii, 2006).

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 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. 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.

Funding

This work was supported by funding provided by the Universität Hamburg.

Acknowledgements

We gratefully acknowledge the support of Friederike Baier, Jan-Ole Großmann, Vincent Kühn, and Ricarda Vielhauer during data collection.

Declaration of interests

The authors declare no competing financial interests.

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 & 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

- Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., & Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics, 8*.
<https://doi.org/10.3389/fninf.2014.00014>
- Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences, 91*(15), 7041–7045. <https://doi.org/10.1073/pnas.91.15.7041>
- 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*(3), 255–278. <https://doi.org/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., Kahane, P., Lachaux, J.-P., & Jerbi, K. (2016). Direct Recordings from Human Anterior Insula Reveal its Leading Role within the Error-Monitoring Network. *Cerebral Cortex, bhv352*.
<https://doi.org/10.1093/cercor/bhv352>
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1–48.
<https://doi.org/10.18637/jss.v067.i01>
- Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods, 190*(1), 80–91.
<https://doi.org/10.1016/j.jneumeth.2010.04.028>
- Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. *Trends in Neurosciences, 21*(7), 294–299. [https://doi.org/10.1016/S0166-2236\(97\)01214-9](https://doi.org/10.1016/S0166-2236(97)01214-9)

- Cohen, M. X. (2008). Neurocomputational mechanisms of reinforcement-guided learning in humans: A review. *Cognitive, Affective, & Behavioral Neuroscience*, 8(2), 113–125.
<https://doi.org/10.3758/CABN.8.2.113>
- Cycowicz, Y. M., & Friedman, D. (2007). Visual novel stimuli in an ERP novelty oddball paradigm: Effects of familiarity on repetition and recognition memory. *Psychophysiology*, 44(1). <https://doi.org/10.1111/j.1469-8986.2006.00481.x>
- Davachi, L., & Wagner, A. D. (2002). Hippocampal Contributions to Episodic Encoding: Insights From Relational and Item-Based Learning. *Journal of Neurophysiology*, 88(2), 982–990. <https://doi.org/10.1152/jn.2002.88.2.982>
- 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 B: Biological Sciences*, 363(1511), 3787–3800.
<https://doi.org/10.1098/rstb.2008.0161>
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (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.
<https://doi.org/10.1016/j.neuroimage.2006.01.021>
- Dosenbach, N. U. F., Visscher, K. M., Palmer, E. D., Miezin, F. M., Wenger, K. K., Kang, H. C., Burgund, E. D., Grimes, A. L., Schlaggar, B. L., & Petersen, S. E. (2006). A Core System for the Implementation of Task Sets. *Neuron*, 50(5), 799–812.
<https://doi.org/10.1016/j.neuron.2006.04.031>
- Eichenbaum, H. (2001). The hippocampus and declarative memory: Cognitive mechanisms and neural codes. *Behavioural Brain Research*, 127(1–2), 199–207.
[https://doi.org/10.1016/S0166-4328\(01\)00365-5](https://doi.org/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.
<https://doi.org/10.1016/j.neuron.2004.08.028>
- Ergo, K., De Loof, E., & Verguts, T. (2020). Reward Prediction Error and Declarative Memory. *Trends in Cognitive Sciences*, *24*(5), 388–397.
<https://doi.org/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. *The Journal of Neuroscience*, *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, C. E. (1999). Real-time tracking of memory formation in the human rhinal cortex and hippocampus. *Science (New York, N.Y.)*, *285*(5433), 1582–1585. <https://doi.org/10.1126/science.285.5433.1582>
- 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). Psychology Press.
- Fouragnan, E., Retzler, C., & Philiastides, M. G. (2018). Separate neural representations of prediction error valence and surprise: Evidence from an fMRI meta-analysis. *Human Brain Mapping*, *39*(7), 2887–2906. <https://doi.org/10.1002/hbm.24047>
- Garrison, J., Erdeniz, B., & Done, J. (2013). Prediction error in reinforcement learning: A meta-analysis of neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *37*(7), 1297–1310. <https://doi.org/10.1016/j.neubiorev.2013.03.023>
- Gershman, S. J., & Daw, N. D. (2017). Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annual Review of Psychology*, *68*(1), 101–128. <https://doi.org/10.1146/annurev-psych-122414-033625>

- Ghosh, V. E., & Gilboa, A. (2014). What is a memory schema? A historical perspective on current neuroscience literature. *Neuropsychologia*, *53*, 104–114.
<https://doi.org/10.1016/j.neuropsychologia.2013.11.010>
- Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*(Supplement_3), 15647–15654. <https://doi.org/10.1073/pnas.1014269108>
- Golfinopoulos, E., Tourville, J. A., Bohland, J. W., Ghosh, S. S., Nieto-Castanon, A., & Guenther, F. H. (2011). fMRI investigation of unexpected somatosensory feedback perturbation during speech. *NeuroImage*, *55*(3), 1324–1338.
<https://doi.org/10.1016/j.neuroimage.2010.12.065>
- Green, P., & MacLeod, C. J. (2016). SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498.
<https://doi.org/10.1111/2041-210X.12504>
- 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. <https://doi.org/10.1016/j.jml.2016.11.001>
- Ham, T., Leff, A., de Boissezon, X., Joffe, A., & Sharp, D. J. (2013). Cognitive Control and the Salience Network: An Investigation of Error Processing and Effective Connectivity. *Journal of Neuroscience*, *33*(16), 7091–7098.
<https://doi.org/10.1523/JNEUROSCI.4692-12.2013>
- Henson, R. N., & Gagnepain, P. (2010). Predictive, interactive multiple memory systems. *Hippocampus*, *20*(11), 1315–1326. <https://doi.org/10.1002/hipo.20857>
- Hermans, E. J., Battaglia, F. P., Atsak, P., de Voogd, L. D., Fernández, G., & Rozenendaal, B. (2014). How the amygdala affects emotional memory by altering brain network properties. *Neurobiology of Learning and Memory*, *112*, 2–16.
<https://doi.org/10.1016/j.nlm.2014.02.005>

- Jang, A. I., Nassar, M. R., Dillon, D. G., & Frank, M. J. (2019). Positive reward prediction errors during decision-making strengthen memory encoding. *Nature Human Behaviour*, 3(7), 719–732. <https://doi.org/10.1038/s41562-019-0597-3>
- Kalbe, F., & Schwabe, L. (2020). Beyond arousal: Prediction error related to aversive events promotes episodic memory formation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2), 234–246. <https://doi.org/10.1037/xlm0000728>
- Kerns, J. G. (2004). Anterior Cingulate Conflict Monitoring and Adjustments in Control. *Science*, 303(5660), 1023–1026. <https://doi.org/10.1126/science.1089910>
- Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage*, 56(2), 411–421. <https://doi.org/10.1016/j.neuroimage.2011.01.061>
- Mayes, A., Montaldi, D., & Migo, E. (2007). Associative memory and the medial temporal lobes. *Trends in Cognitive Sciences*, 11(3), 126–135. <https://doi.org/10.1016/j.tics.2006.12.003>
- McGaugh, J. L. (2018). Emotional arousal regulation of memory consolidation. *Current Opinion in Behavioral Sciences*, 19, 55–60. <https://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(2), 205–210. [https://doi.org/10.1016/S0959-4388\(02\)00306-9](https://doi.org/10.1016/S0959-4388(02)00306-9)
- McHugh, S. B., Barkus, C., Huber, A., Capitao, L., Lima, J., Lowry, J. P., & Bannerman, D. M. (2014). Aversive Prediction Error Signals in the Amygdala. *Journal of Neuroscience*, 34(27), 9024–9033. <https://doi.org/10.1523/JNEUROSCI.4465-13.2014>
- Menon, V. (2011). Large-scale brain networks and psychopathology: A unifying triple network model. *Trends in Cognitive Sciences*, 15(10), 483–506. <https://doi.org/10.1016/j.tics.2011.08.003>

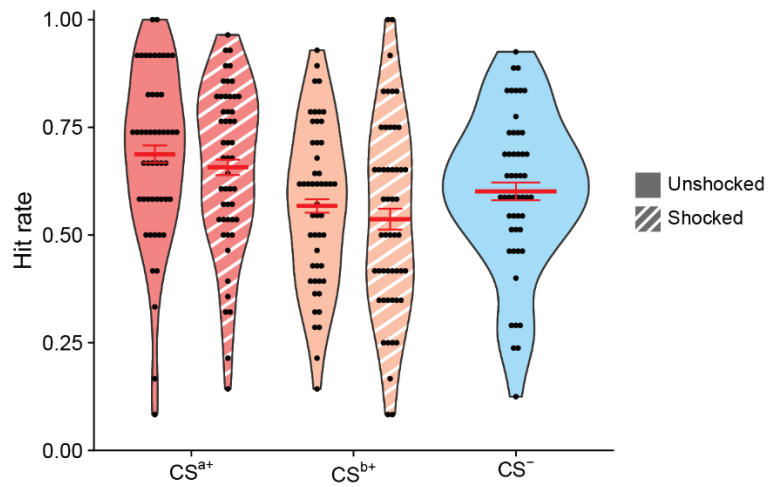
- Metereau, E., & Dreher, J.-C. (2013). Cerebral Correlates of Salient Prediction Error for Different Rewards and Punishments. *Cerebral Cortex*, *23*(2), 477–487.
<https://doi.org/10.1093/cercor/bhs037>
- Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118. <https://doi.org/10.1016/j.neuroimage.2010.05.051>
- Mizumori, S. J. Y. (2013). Context Prediction Analysis and Episodic Memory. *Frontiers in Behavioral Neuroscience*, *7*. <https://doi.org/10.3389/fnbeh.2013.00132>
- Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25.
<https://doi.org/10.1002/hbm.1058>
- Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*(3), 139–154. <https://doi.org/10.1016/j.jmp.2008.12.005>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.
- 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.
<https://doi.org/10.1016/j.cortex.2018.10.010>
- Preuschoff, K., Quartz, S. R., & Bossaerts, P. (2008). Human Insula Activation Reflects Risk Prediction Errors As Well As Risk. *Journal of Neuroscience*, *28*(11), 2745–2752.
<https://doi.org/10.1523/JNEUROSCI.4286-07.2008>

- Reed, J. M., & Squire, L. R. (1997). Impaired recognition memory in patients with lesions limited to the hippocampal formation. *Behavioral Neuroscience, 111*(4), 667–675. <https://doi.org/10.1037//0735-7044.111.4.667>
- Richardson, M. P., Strange, B. A., & Dolan, R. J. (2004). Encoding of emotional memories depends on amygdala and hippocampus and their interactions. *Nature Neuroscience, 7*(3), 278–285. <https://doi.org/10.1038/nn1190>
- 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*(9), 1430–1443. <https://doi.org/10.1037/xlm0000518>
- Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition, 203*, 104269. <https://doi.org/10.1016/j.cognition.2020.104269>
- Schultz, W. (1998). Predictive Reward Signal of Dopamine Neurons. *Journal of Neurophysiology, 80*(1), 1–27. <https://doi.org/10.1152/jn.1998.80.1.1>
- Seeley, W. W., Menon, V., Schatzberg, A. F., Keller, J., Glover, G. H., Kenna, H., Reiss, A. L., & Greicius, M. D. (2007). Dissociable Intrinsic Connectivity Networks for Salience Processing and Executive Control. *Journal of Neuroscience, 27*(9), 2349–2356. <https://doi.org/10.1523/JNEUROSCI.5587-06.2007>
- Shohamy, D., & Adcock, R. A. (2010). Dopamine and adaptive memory. *Trends in Cognitive Sciences, 14*(10), 464–472. <https://doi.org/10.1016/j.tics.2010.08.002>
- Shrager, Y., Kirwan, C. B., & Squire, L. R. (2008). Activity in Both Hippocampus and Perirhinal Cortex Predicts the Memory Strength of Subsequently Remembered Information. *Neuron, 59*(4), 547–553. <https://doi.org/10.1016/j.neuron.2008.07.022>
- Sridharan, D., Levitin, D. J., & Menon, V. (2008). A critical role for the right fronto-insular cortex in switching between central-executive and default-mode networks.

- Proceedings of the National Academy of Sciences*, 105(34), 12569–12574.
<https://doi.org/10.1073/pnas.0800005105>
- Stanek, J. K., Dickerson, K. C., Chiew, K. S., Clement, N. J., & Adcock, R. A. (2019). Expected Reward Value and Reward Uncertainty Have Temporally Dissociable Effects on Memory Formation. *Journal of Cognitive Neuroscience*, 31(10), 1443–1454. https://doi.org/10.1162/jocn_a_01411
- Starkweather, C. K., Gershman, S. J., & Uchida, N. (2018). The Medial Prefrontal Cortex Shapes Dopamine Reward Prediction Errors under State Uncertainty. *Neuron*, 98(3), 616–629.e6. <https://doi.org/10.1016/j.neuron.2018.03.036>
- Strange, B. A., & Dolan, R. J. (2004). -Adrenergic modulation of emotional memory-evoked human amygdala and hippocampal responses. *Proceedings of the National Academy of Sciences*, 101(31), 11454–11458. <https://doi.org/10.1073/pnas.0404282101>
- Summerfield, C., & Egnér, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, 13(9), 403–409. <https://doi.org/10.1016/j.tics.2009.06.003>
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 88(2), 135–170.
<https://doi.org/10.1037/0033-295X.88.2.135>
- Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural Systems for Error Monitoring: Recent Findings and Theoretical Perspectives. *The Neuroscientist*, 13(2), 160–172.
<https://doi.org/10.1177/1073858406298184>
- Ullsperger, M., Harsay, H. A., Wessel, J. R., & Ridderinkhof, K. R. (2010). Conscious perception of errors and its relation to the anterior insula. *Brain Structure and Function*, 214(5–6), 629–643. <https://doi.org/10.1007/s00429-010-0261-1>
- van Kesteren, M. T. R., Ruiters, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, 35(4), 211–219.
<https://doi.org/10.1016/j.tins.2012.02.001>

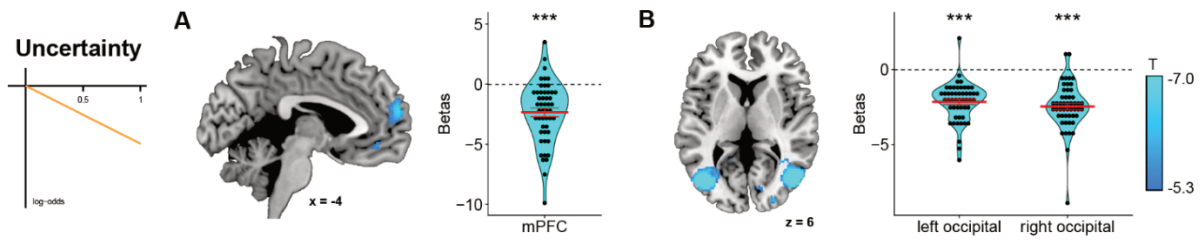
- Vogel, S., Klüen, L. M., Fernández, G., & Schwabe, L. (2018a). Stress leads to aberrant hippocampal involvement when processing schema-related information. *Learning & Memory*, 25(1), 21–30. <https://doi.org/10.1101/lm.046003.117>
- Vogel, S., Klüen, L. M., Fernández, G., & Schwabe, L. (2018b). Stress affects the neural ensemble for integrating new information and prior knowledge. *NeuroImage*, 173, 176–187. <https://doi.org/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 Connectivity*, 2(3), 125–141. <https://doi.org/10.1089/brain.2012.0073>
- Wolbers, T. (2003). Contralateral Coding of Imagined Body Parts in the Superior Parietal Lobe. *Cerebral Cortex*, 13(4), 392–399. <https://doi.org/10.1093/cercor/13.4.392>
- Wolpert, D. M., Goodbody, S. J., & Husain, M. (1998). Maintaining internal representations: The role of the human superior parietal lobe. *Nature Neuroscience*, 1(6), 529–533. <https://doi.org/10.1038/2245>
- Xia, C., Touroutoglou, A., Quigley, K. S., Feldman Barrett, L., & Dickerson, B. C. (2017). Salience Network Connectivity Modulates Skin Conductance Responses in Predicting Arousal Experience. *Journal of Cognitive Neuroscience*, 29(5), 827–836. https://doi.org/10.1162/jocn_a_01087
- Yoshida, W., & Ishii, S. (2006). Resolution of Uncertainty in Prefrontal Cortex. *Neuron*, 50(5), 781–789. <https://doi.org/10.1016/j.neuron.2006.05.006>
- Young, C. B., Raz, G., Everaerd, D., Beckmann, C. F., Tendolkar, I., Hendler, T., Fernández, G., & Hermans, E. J. (2017). Dynamic Shifts in Large-Scale Brain Network Balance As a Function of Arousal. *The Journal of Neuroscience*, 37(2), 281–290. <https://doi.org/10.1523/JNEUROSCI.1759-16.2016>

SUPPLEMENTARY MATERIAL



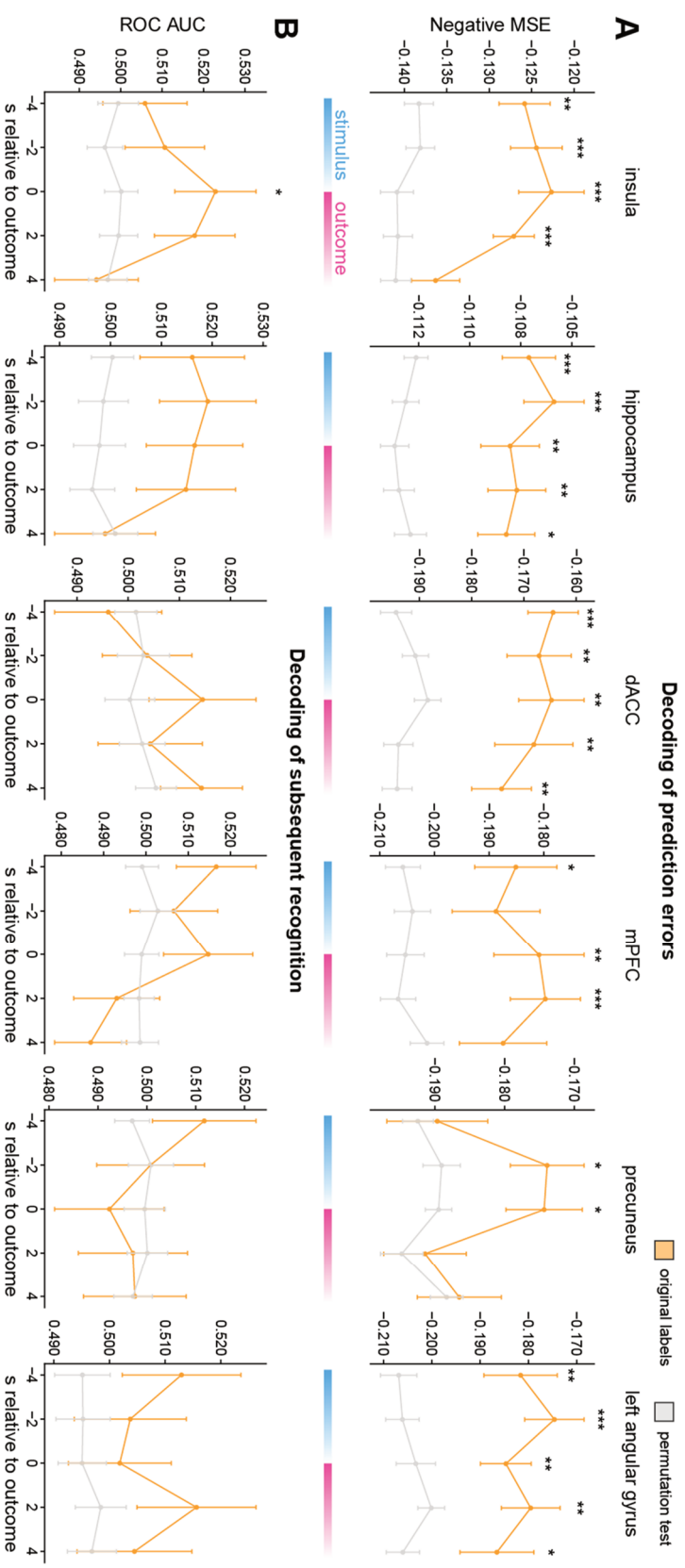
Supplemental Figure 1. Hit rates by CS category and trial outcome.

Hit rates for items from the CS^{a+} category were larger compared with both CS^{b+} and CS⁻ items. Importantly, the greater number of aversive shocks for CS^{a+} items could not explain this memory difference, as participants descriptively even showed slightly decreased hit rates for shocked (compared with unshocked) items in both the CS^{a+} and CS^{b+} category.



Supplemental Figure 2. Univariate fMRI analysis to identify regions associated with uncertainty in unshocked trials

For prediction uncertainty in unshocked trials, we obtained decreased BOLD responses in the mPFC and bilateral occipital areas (A, B). Only voxels significant at 5%-level 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. Thin red bars show ± 1 standard error of the mean. *** $p_{FWE} < .001$.



Supplemental Figure 3. Decoding of prediction errors and subsequent recognition using multivoxel pattern analysis (MVPA)

(A) Activity patterns from regions associated with negative PEs in the univariate analysis could be used to decode the magnitude of negative PEs. Best decoding performance was generally achieved around the time the outcome was revealed.

(B) Subsequent recognition of an item could be decoded significantly above chance level using patterns of activity from the insula specifically when the outcome of a trial was revealed.

* $p < .05$, ** $p < .01$, *** $p < .001$.

Supplemental Table 1. Univariate fMRI results for PEs

Effect	Direction	Region	x	y	z	T	p(FWE)
Negative PEs (cluster extend threshold: $k \geq 10$)	positive	right anterior insula	32	22	-8	8.51	2.09×10^{-6}
		left anterior insula	-34	18	4	8.01	9.97×10^{-6}
		dACC	8	26	42	7.63	3.39×10^{-5}
		right inferior frontal gyrus	48	26	22	5.68	1.43×10^{-2}
		right middle frontal gyrus	42	12	30	5.52	2.31×10^{-2}
		left hippocampus	-26	-20	-16	8.49	2.20×10^{-6}
	negative	precuneus	-12	-56	18	8.41	2.89×10^{-6}
		mPFC	-8	50	-6	8.27	4.45×10^{-6}
		left angular gyrus	-44	-70	26	7.86	1.64×10^{-6}
		right hippocampus	24	-22	-18	7.11	1.73×10^{-4}
		left middle temporal gyrus	-62	-12	-16	6.61	8.29×10^{-4}
		left superior frontal gyrus	-24	36	50	6.05	4.74×10^{-3}
Positive PEs (no cluster extend threshold)	positive	left primary motor cortex	-28	-28	60	5.85	8.66×10^{-3}
		left caudate	-8	24	10	5.74	1.20×10^{-2}
		right middle temporal gyrus	60	-4	-22	5.58	1.94×10^{-2}
		left superior parietal lobule	-32	-72	52	6.05	3.01×10^{-3}
		left lateral temporal cortex	-64	-40	-8	5.47	1.72×10^{-2}
	negative	left supramarginal gyrus	-60	-30	40	5.40	2.10×10^{-2}

All displayed peaks were significant at $p < .05$ after whole-brain voxel-level family-wise error correction. Additional minimal cluster extend thresholds of $k \leq 10$ were only applied where indicated.

Supplemental Table 2. Univariate fMRI results for uncertainty

Effect	Direction	Region	x	y	z	T	p(FWE)
Uncertainty (unshocked trials)	positive	right middle frontal gyrus	46	24	34	6.08	5.27×10^{-3}
	negative	right extrastriate cortex	50	-70	6	11.89	3.76×10^{-11}
		left extrastriate cortex	-40	-74	4	11.76	5.63×10^{-11}
		left postcentral	-44	-30	62	8.38	3.88×10^{-6}
		right cuneus	20	-82	40	7.49	6.53×10^{-5}
		right lingual gyrus	16	-62	-4	7.09	2.29×10^{-4}
		left lingual gyrus	-14	-72	0	7.01	2.90×10^{-4}
		mPFC	-4	62	18	6.76	6.46×10^{-4}
		left middle temporal gyrus	-64	-16	-10	6.71	7.61×10^{-4}
		dorsal posterior cingulate area	-10	-24	46	6.67	8.67×10^{-4}
		right primary visual cortex	12	-78	2	6.56	1.18×10^{-3}
		left temporal pole	-58	4	-24	6.39	2.06×10^{-3}
		right fusiform area	42	-28	-16	6.37	2.13×10^{-3}
		inferior temporal area	24	-72	28	6.34	2.35×10^{-3}
		right anterior cingulate area	8	-12	50	6.34	2.37×10^{-3}
		right superior parietal lobule	28	-54	62	6.21	3.55×10^{-3}
		left fusiform area	-40	-46	-24	5.85	1.07×10^{-2}

All displayed peaks were significant at $p < .05$ after whole-brain voxel-level family-wise error correction and clusters extended a threshold of $k = 10$ voxels. For shocked trials, there were no significant voxels under this criterion.

Supplemental Table 3. Univariate fMRI results for predictions

Effect	Direction	Region	x	y	z	T	p(FWE)
Prediction (unshocked trials)	positive	right lingual gyrus	12	-72	-6	13.45	$< 1 \times 10^{-12}$
		left precentral	-36	-26	54	6.95	2.85×10^{-4}
		left supplementary motor area	-14	-8	64	5.33	4.09×10^{-2}
Prediction (shocked trials)	negative	left lingual gyrus	-10	-76	-6	14.18	$< 1 \times 10^{-12}$
	positive	right lingual gyrus	10	-74	-2	8.85	6.83×10^{-7}
	negative	left lingual gyrus	-10	-78	-4	9.11	3.00×10^{-7}

All displayed peaks were significant at $p < .05$ after whole-brain voxel-level family-wise error correction and clusters extended a threshold of $k = 10$ voxels.



Erklärung gemäß *(bitte Zutreffendes ankreuzen)*

- § 4 (1c) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010
- § 5 (4d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003

Hiermit erkläre ich,

_____ (Vorname, Nachname),

dass ich mich an einer anderen Universität oder Fakultät noch keiner Doktorprüfung unterzogen oder mich um Zulassung zu einer Doktorprüfung bemüht habe.

Ort, Datum

Unterschrift

Eidesstattliche Erklärung nach *(bitte Zutreffendes ankreuzen)*

- § 7 (4) der Promotionsordnung des Instituts für Bewegungswissenschaft der Universität Hamburg vom 18.08.2010**
- § 9 (1c und 1d) der Promotionsordnung des Instituts für Psychologie der Universität Hamburg vom 20.08.2003**

Hiermit erkläre ich an Eides statt,

1. dass die von mir vorgelegte Dissertation nicht Gegenstand eines anderen Prüfungsverfahrens gewesen oder in einem solchen Verfahren als ungenügend beurteilt worden ist.
2. dass ich die von mir vorgelegte Dissertation selbst verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt und keine kommerzielle Promotionsberatung in Anspruch genommen habe. Die wörtlich oder inhaltlich übernommenen Stellen habe ich als solche kenntlich gemacht.

Ort, Datum

Unterschrift