Event segmentation protects emotional memories from competing experiences encoded close in time

Joseph E. Dunsmoor[1], **Marijn C.W. Kroes**[2], **Caroline M. Moscatelli**[3], **Michael D. Evans**[3], **Lila Davachi**[4], and **Elizabeth A. Phelps**[3,5,6]

[1]University of Texas at Austin, Department of Psychiatry, Austin, TX 78712, USA [2]Radboud University Medical Centre, Department of Cognitive Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands [3]New York University, Department of Psychology, New York, NY 10003, USA [4]Columbia University, Department of Psychology, New York, NY 10027, USA [5]New York University, Center for Neural Science, New York, NY 10003, USA [6]Nathan Kline Institute, Orangeburg, NY 10962, USA

SUMMARY

Fear memories are characterized by their permanence and a fierce resistance to unlearning by new experiences. We considered whether this durability involves a process of memory segmentation that separates competing experiences. To address this question, we used an emotional learning task designed to measure recognition memory for category exemplars encoded during competing experiences of fear-conditioning and extinction. Here we show that people recognized more fear-conditioned exemplars encoded during conditioning than conceptually related exemplars encoded immediately after a perceptual event boundary separating conditioning from extinction. Selective episodic memory depended on a period of consolidation, an explicit break between competing experiences, and was unrelated to within-session arousal or the explicit realization of a transition from conditioning to extinction. Collectively, these findings suggest that event boundaries guide selective consolidation to prioritize emotional information in memory—at the expense of related but conflicting information experienced shortly thereafter. We put forward a model whereby event boundaries bifurcate related memory traces for incompatible experiences. This stands in contrast to a mechanism that integrates related experiences for adaptive generalization[123], and reveals a potentially distinct organization by which competing memories are adaptively segmented to select and protect nascent fear memories from immediate sources of interference.

Correspondence. Correspondence concerning this article should be addressed to Joseph Dunsmoor, joseph.dunsmoor@austin.utexas.edu.

INTRODUCTION

Beyond being strongly remembered, fearful events are also fiercely resistant to being minimized and forgotten by new contradictory experiences. Consider a highly memorable experience like a surprise encounter with a terrifying snake on a hiking trial. This moment will likely be remembered, even if shortly after seeing the snake a trusted friend tells you it was a nonvenomous and nonaggressive species. In other words, although your experience transitioned from fear to safety, the nascent fear memory is immune to new countervailing knowledge of safety, and there remains a strong chance you will remember the snake on your next walk along the trail. While a considerable amount of research details how emotional events are preserved in long-term memory, it is far less clear how emotional memories resist disruption by new learning that directly contradicts a prior emotional experience. Knowledge in this domain can be sourced almost entirely to the field of Pavlovian threat ("fear") conditioning, which views extinction of conditioned threat as an active learning process that results in a separate (and more fragile) memory representation parallel to the original conditioned threat memory. This view dates back to Pavlov[4–6] and is supported by countless behavioral findings that conditioned behaviors reemerge following extinction[7], and neurobiological findings that threat and extinction memories are separately represented in discrete neural pathways[8–10] and neural ensembles in the amygdala[11 for review]. Why and how threat memories are segmented from—rather than overwritten by—incompatible experiences, however, is unclear[12]. This is especially vexing considering that other types of memory (e.g., motor[13] or word learning[14]) are susceptible to retroactively interfering experiences after being formed.

The field of episodic memory research provides a potential mechanism to explain how emotional memories are protected from immediate sources of interference. Specifically, event segmentation[15] has been shown to result in a loss of preceding event representations from active memory[16–19] and affects integration of long-term memory for information encountered across event boundaries[20–24]. Some recent work suggests that event boundaries might signal a transition from active encoding to consolidation[25,26]. A plausible and adaptive consequence of boundary-triggered consolidation for emotional memory could involve binding information encoded during the emotional experience into a single episode, whilst effectively isolating ongoing cellular and molecular processes involved in long-term memory consolidation from future sources of interference. An analogous process of event segmentation may occur by happenstance in laboratory conditioned threat extinction experiments also, as the transition from conditioning to extinction is often marked by a change in the spatial-temporal context that may bifurcate learning into distinct episodes[27]. An early initiation of consolidation triggered by event boundaries may in part explain why nascent conditioned threat memories are seemingly unaffected by sources of interference during the immediate post-training period[28–34], in stark contrast to non-emotional hippocampal or striatal-based memories which are most sensitive to disruption during this period[35–37].

Within the domain of episodic memory research, event segmentation has recently been shown to organize long-term memory into discrete episodes in humans[20]. Event boundaries may act as "anchors"[15] during episodic encoding that affects long-term episodic memory by

enhancing mnemonic integration (or clustering) of information learned within event boundaries[20,38,39]. This process can help bind individual episodic elements into a cohesive and meaningful unit for effective storage and retrieval[15,21,26,40]. Perhaps event segmentation also functions to select and protect emotionally relevant experiences. This may in part explain why fearful events fiercely resist retroactive interference by new contradictory experiences.

Here we developed a new emotional learning task in humans to test the hypothesis that event boundaries organize the long-term structure of emotional episodic memory. The task matches elements of Pavlovian conditioning with episodic learning by testing recognition memory for conditioned stimuli following Pavlovian conditioning and extinction. Subjects learned through experience that images from one object category (CS+, animals or tools) predicted an aversive electric shock to the right wrist (unconditioned stimulus, US) and images from another category (CS−, tools or animals, respectively) were not paired with shock (Figure 1A). We used a 50% reinforcement rate, whereby half of the items from the CS+ category encoded during conditioning were paired with shock. A subtle but explicit rest period (~10 s) separated early and late stages of threat conditioning, followed by another short explicit rest period between the late stage of conditioning and extinction. These explicit rest periods served to break up a series of conditioning trials and thus acted as de facto event boundaries. Participants returned 24 hours later for a surprise recognition memory test comprised of trial-unique CSs from conditioning, extinction, and category-related foils to control for false alarms. As in our prior work[2,41,42], the use of separate object categories allowed us to test selective memory for items from the CS+ versus a control category (CS−) in a within-subjects design. By combining threat conditioning with episodic learning, each trial is effectively isolated as a unique learning trial encoded at a specific moment in time during either conditioning or extinction. Then, at a surprise episodic memory test, we can measure recognition performance for each trial as a time-ordered function of *when* that trial was encoded. This allowed us to examine whether an explicit event boundary between threat and extinction affects the structure of long-term emotional memory.

Analysis of recognition memory tested 24-hours after encoding (N = 20; Figure 1B) revealed greater memory for conditioning-specific CS+ than CS− items. Given that half of the CS+ trials encoded during conditioning were *not* paired with shock (50% reinforcement rate), we could assess whether memory was better for trials from the same semantic category paired or unpaired with shock. Importantly, and in keeping with our prior findings[2,41,42], long-term episodic memory enhancements generalized to all CS+ trials encoded during the conditioning phase regardless of whether a CS+ trial was paired or unpaired with shock (Figure 1C). However, this memory benefit did not extend to conceptually-related unpaired CS+ exemplars encoded immediately after the break separating conditioning from extinction; a paired-samples t-test of mean recognition memory for CS+ extinction exemplars encoded during the first 20 trials of extinction was weaker than for conceptually related CS+ threat conditioning exemplars encoded during the last 20 trials of threat conditioning (t_{19}=2.987, P<0.008, d_{av}=.49, 95% CI [.030, .174]). Finally, whereas recognition memory for conceptually-related CS+ exemplars abruptly dropped for trials encoded after the break separating conditioning from extinction, trial-by-trial measures of

explicit learning of threat potential (i.e., shock expectancy) and conditioned autonomic arousal (i.e., skin conductance responses) obtained at the time of encoding (see **Methods**) only gradually declined as subjects began to learn through experience that CS+ trials no longer predicted shock (Figure 1D). A change point analysis[43,44] was used to detect a significant change in slope of the time-ordered behavioral data, visually depicted in Figure 1E as the cumulative record (sum) of the difference between behavior at each time point (averaged in blocks of 4 trials) and the overall average (see **Methods**). These analyses allowed us to detect the point at which behavioral performance significantly changed. This revealed that the significant decline in 24-hour recognition memory for CS+ exemplars preceded the change in shock expectancy and SCR at the time of encoding. Thus, online measures of learning about the probability of shock cannot explain the segmentation of subsequent memory.

These results raise the intriguing possibility that the long-term memory strength for conceptually-related CS+ conditioning and CS+ extinction exemplars were determined by the short break separating conditioning from extinction. We emphasize that this effect could only be revealed by a memory test that simultaneously probes item memory strength for trial-unique CSs as a function of when each CS had been encoded. Further, because the abrupt drop in CS+ item memory preceded the online decrease in expectancy and arousal, these memory results cannot be ascribed simply to an error-correcting learning process e.g., [45], which would predict that CS+ item memory strength should track the underlying rate of associative learning.

To test whether selective long-term enhancement of episodic CS+ conditioning memory, the drop in CS+ extinction memory, or both required a period of consolidation to emerge, a separate group underwent the same protocol with an immediate memory test (N = 20). To test the role of explicit episodic boundaries, a third group was tested after a period of consolidation but there were no breaks at any point during encoding (N = 20), such that conditioning and extinction consisted of one lengthy unbroken session. Repeated measures ANOVA on recognition memory of all three groups (Conditioning / Extinction — 24-hr retrieval; Conditioning / Extinction / Immediate retrieval; Conditioning and Extinction without breaks — 24-hr retrieval) revealed a significant CS type (2: CS+, CS−) by Phase (3: early conditioning and late conditioning, and the first half of extinction) by Group (3) interaction ($F_{4,114} = 2.527$, $P = .044$, partial eta squared = .081). For the immediate retrieval group (Figure 1F, left), item memory was slightly enhanced for conditioning-specific CS+ exemplars encoded during early conditioning, but there was no drop in item memory for CS+ extinction exemplars encoded across the break separating late conditioning from early extinction (paired-samples t-test: $t_{19}=1.022$, $P=0.317$, $d_{av}=.28$, 95% CI [−.029, .084]), indicating that the drop in CS+ extinction memory at 24-hrs was consolidation-dependent. Subjects who underwent conditioning and extinction without any breaks (Figure 1F, right) also showed a slight enhancement in memory for CS+ versus CS− exemplars encoded during the second run of conditioning, but no drop in CS+ extinction memory (paired-samples t-test: $t_{19}=1.292$, $P=0.212$, $d_{av}=.21$, 95% CI [−.026, .111]).

An additional group revealed that subjects for whom conditioning continued after the 2nd break (N = 20; Figure 1G, left), and thus never underwent extinction, showed prioritized 24-

hr memory for CS+ versus CS− exemplars encoded throughout the entire task. This shows that extinction, per se, is necessary to produce a drop in CS+ item memory for trials encoded after a break. This also effectively rules out that the drop in memory from study 1 was merely the result of subjects having poorer memory for trials encoded in the latter half of the task. Finally, the drop in CS+ memory for items encoded after conditioning is time-dependent; when extinction (N = 18) or a second conditioning session (N = 20) was separated from conditioning by 24-hours, there was no drop in CS+ memory (tested on the 3rd day) for trials encoded at the end of Day 1 and the start of Day 2: paired samples t-tests t_{17}=.403, P=0.692, d_{av}=.14, 95% CI [−.059, .087] (Figure 1G, middle) and t_{19}=1.560, P=0.135, d_{av}=.29, 95% CI [−.019, .135] (Figure 1G, right), respectively. Together, these additional experiments show that the explicit breaks separating conditioning from extinction only derived their utility to segment item memory strength for conditioning versus extinction if these phases occurred close in time, i.e., within a consolidation window[36].

We next reasoned that a strong test for the importance of the explicit punctate event boundary is to occasionally reinforce the CS+ at a gradually decreasing rate cf.,[46] after the 10s rest period (N = 17; Figure 2A). Such a procedure delays the moment when subjects explicitly or internally realize extinction until well after the last explicit break separating conditioning from extinction. This procedure also provides a strong test for whether the drop in CS+ item memory strength during extinction is merely due to separate affective encoding states: in gradual extinction, arousal and attention to CS+ trials will be maintained throughout the early phase of extinction due to occasional CS-US pairing. An additional 40 CS+ (and 40 CS−) trials were added to the end of extinction to ensure subjects eventually realized safety, but memory for these additional trials were not tested the following day.

Remarkably, 24-hour CS+ item recognition memory was again sharply divided at the explicit transition point separating the end of conditioning from the start of (gradual) extinction (paired samples t-test: t_{16}=3.031, P=0.008, d_{av}=.44, 95% CI [.027, .155]) (Figure 2B, left bars), despite the fact that these CS+ extinction trials were intermittently paired with shock. Indeed, occasional CS-US pairings successfully maintained within-session arousal and expectancy to the CS+ at the time these items were encoded (Figure 2C). A change point analysis (Figure 2D) confirmed that the drop in item memory occurred immediately at the transition point preceding within-session measures of conditioned learning (note—change point analyses for expectancy and SCR did not incorporate the additional 40 CS+ extinction trials). Again, this effect was consolidation dependent, as an immediate memory test following gradual extinction (N = 20; Figure 2B, right bars) showed neither selective prioritization for CS+ conditioning items nor a drop in recognition for CS+ extinction items.

These episodic memory results show that what is remembered from extinction is more fragile than what is remembered from conditioning. Importantly, a number of emerging behavioral strategies show potential as approaches to strengthen extinction learning and prevent post-extinction return of defensive behavior[12]. We tested some promising extinction strategies here to see whether techniques that diminish post-extinction behavioral expression of a threat-memory (e.g., freezing in rats or conditioned arousal in humans) have any retroactive effects on selective emotional episodic memory for CS+ trials encoded during conditioning. In other words, do optimized extinction techniques diminish or interfere with

episodic details of the original threat memory? We included (i) massive extinction training (N = 20) with double the number of extinction-specific CS+ items cf.[47], (ii) an augmented form of extinction in which the shock is replaced on every CS+ trial by a non-aversive outcome (N = 20) referred to as novelty-facilitated extinction[48], and (iii) a combination of approaches that included multiple context shifts cf.[49] (represented by a change in the background color) during massive novelty-facilitated extinction (N = 21).

None of these techniques successfully reduced 24-hour conditioning-specific episodic memory enhancements, and each technique left the abrupt drop in CS+ memory across the break separating conditioning from extinction intact (see Figure 3A, B and Supplemental Materials for statistical analysis). Change point analysis confirmed that the drop in CS+ memory occurred abruptly at the transition point, and that once again this drop preceded the decrease in arousal and expectancy (Figure 3C). Thus, emotional episodic memories are prioritized at the expense of related memories of safety that occur within the penumbra of emotional learning. In effect, these data represent a different way of showing that a variety of extinction learning strategies that attenuate the return of defensive behaviors likely do not operate by retroactively interfering with the original episodic emotional memory, in line with conventional models of fear extinction. This of course leaves open the question of how optimized extinction techniques do operate to prevent future threat expression, and whether the original episodic threat memory can be persistently altered. One potential result of strengthening new inhibitory learning could have been to improve long-term episodic memory for stimuli encoded during extinction, but such a result was not observed.

People automatically accumulate new information in the aftermath of a fearful or threatening event. This new information might be relevant for predicting and avoiding threat in the future, and should therefore be selectively preserved in memory along with details directly associated with the experience[50]. But this new information might also invalidate prior emotional learning by disconfirming a learned threat association. An adaptive memory system should ensure that contradictory experiences of safety are stored as separate memory traces that receive lower priority in long-term memory, since the risk of forgetting about danger can be much more costly than forgetting about safety[51].

These results support the idea that a subtle break separating threat conditioning from extinction segments these competing memory traces. One explanation for better memory performance for CS+ than CS− items from conditioning is that stress or arousal induced by the anticipation for shock strengthened encoding, akin to emotional enhancements for episodic memory seen using intrinsically evocative stimuli[52]. It is possible that the memory benefit for CS+ trials encoded during conditioning that were *not* paired with shock was due to more elaborate encoding as subjects tried to predict the outcome during the anticipation period. This 'Levels of Processing' explanation, however, cannot explain why memory performance was weaker for CS+ items encoded immediately after the transition to extinction, given that shock expectancy and arousal were maintained as subjects eventually learned extinction. From the Pavlovian conditioning[53,54] and human clinical research literature[55], it is now widely appreciated that the magnitude of within-session defensive behavior is an unreliable indicator of between-session performance. This dissociation between arousal and attention at the time of encoding from long-term behavioral

performance at test implies that other (perhaps latent) factors determine the long-term structure of conditioned memory.

To elaborate on this point, it is especially noteworthy that CS+ exemplars encoded during gradual extinction received lower priority in long-term memory than conceptually-related CS+ exemplars from before the transition. That is, subjects could only realize retrospectively at some point after the transition to gradual extinction that the previous break represented a transition toward a period of safety. This perhaps indicates that event segmentation is an automatic process that sets anchors at event boundaries that can later be utilized (i.e., retroactively) once a meaningful change in the environment is detected. In this case, information related to threat (e.g., pictures of animals both paired and unpaired with shock) encoded within the boundaries of threat conditioning aggregate collectively, over a period of consolidation, as a unitized episode that receives high priority in long-term memory. Simultaneously, CS+ trials encoded after the boundary (e.g., different animal pictures) also aggregate collectively as a separate unitized episode that receives *lower* priority in memory. What is striking is that CS+ trials encoded after the transition to gradual extinction presumably have more in common with CS+ trials encoded before the transition via association with shock; yet subjects still remembered fewer items from (gradual) extinction than from threat conditioning. Thus, even if extinction is only gradually realized, the most recent explicit boundary is still relied on as a landmark to retroactively segment the memory trace for selective consolidation. This result also makes clear that long-term emotional memory is not predicted by emotional arousal or expectancies for shock at the time of encoding, which were both maintained well past the point at which subjects would later show the drop in memory.

Future research will need to assess whether these episodic memory results correspond with long-term physiological threat responses. For instance, there are contradictory findings in the literature on the return of fear when extinction follows on the heels of conditioning ("immediate extinction"). Although there is some debate as to whether a quick transition from conditioning to extinction causes unlearning through putative depotentiaton of synaptic activity in the amygdala during a labile period[56], most empirical evidence points to *stronger* reemergence of defensive behavior if extinction occurs soon after conditioning e.g.,[33,34]. Our results fits with the idea of an 'immediate extinction deficit,' as extinction on the heels of conditioning resulted in a weaker episodic memory trace of extinction training. That is, immediate extinction that involved an explicit (or perceptible) break appeared to segment memory. We found that immediate extinction with *no* breaks, however, resulted in an overall diminution of CS+ episodic memory performance with no decrement in extinction-specific memory, suggesting that the absence of a break was the most effective procedure to reduce selective emotional memory enhancement.

However, whether a truly immediate extinction session affects the return of physiological threat responses is surprisingly unclear from the literature: "immediate extinction" paradigms in animals incorporated a ~10 minute break and changed the context between conditioning to extinction[33,56]. An immediate extinction paradigm in which the animal is removed from the conditioning environment even briefly would, according to the model we put forth here, serve as a sufficient event boundary to segment threat memories from

extinction memories. There is some evidence that immediate extinction without any delay (versus 24-hour delayed extinction) does diminish post-extinction return of fear-potentiated startle responses in humans[57], which is generally in line with the recognition memory results provided here. Likewise, we found evidence of event segmentation for gradual extinction despite some evidence that a gradual extinction paradigm reduces the return of physiological threat behaviors (freezing in rats[46], startle in humans[58]). Overall, the relationship between segmentation of episodic memories related to threat and extinction, and post-extinction recovery of physiological threat behaviors, warrants exploration. We envision a rather straightforward test in animal models of whether a truly unbroken conditioning-to-extinction session diminishes post-extinction recovery. Such evidence might support a model whereby the absence of an event boundary serves to integrate, rather than segment, competing memory traces, which might consequently minimize selective consolidation of a threat memory.

Another question raised by these results is what, precisely, constituted an event boundary for segmenting memory. Event boundaries have been broadly construed as perceptual or conceptual change in the environment that breaks up a continuous stream of experience[15,39]. In human memory research, these changes can be obvious, such as the end of a video clip[40], or more subtle, such as transition phrases in a narrative[20]. Here, the explicit nature of the rest periods—in which subjects were specifically informed that the task would resume after the break—appeared sufficient to break up a series of CS trials. The explicit nature of the rest period likely distinguishes it from other potential event boundaries, such as the offset of each CS trial. But whether the qualitative nature of an event boundary has different effects on the organization of emotional memory (for instance, leading to more or less event segmentation) is an important question.

Finally, according to most contemporary associative learning models, memories of threat and extinction co-exist as parallel memory traces. Bouton's theory of extinction e.g.,[59] has been widely influential in describing the contextual factors that promote retrieval of one memory over the other. Here, by effectively tagging each trial as a unique learning episode, we could track item memory in a time-ordered fashion as a function of when the item was encoded: conditioning or extinction. This revealed that extinction memories might not simply co-exist parallel to threat memories, but that extinction leaves a weaker long-term memory trace overall. Put simply, threat is easier to remember.

In conclusion, event boundaries serve as useful anchors in our otherwise continuous stream of episodic experience[15], and may provide a landmark to retroactively cleave and protect our most prominent experiences from being consolidated in the same memory trace as similar but contradictory information. Here we provide evidence that an event boundary allows threat and safety to be segmented into separate episodic memory traces during consolidation, thereby prioritizing emotionally relevant memories from interference. Specifically, we found that a subtle transition from threat to safety gains significance as an event boundary, perhaps triggering separate consolidation mechanisms of conditioning and extinction memories[31]. The idea that event boundaries set a learning tag that gets utilized during consolidation (Figure 4D) is in keeping with our prior work showing that meaningful events retroactively enhance memory consolidation for conceptually related items[2,41]. The present findings

extend this 'behavioral tagging' hypothesis[60,61] to event boundaries as a means to separate, rather than integrate, memory traces for related events. This mechanism may have the effect of protecting emotional memories from incompatible experiences during a sensitive period of memory formation, thus ensuring that we remember an emotional event regardless of what happens next. Unfortunately, a putative automatic event segmentation mechanism would cost the ability to easily rid ourselves of unwanted emotional memories.

Methods

Subjects

Based on our prior work[2,41,42], we sought 20 subjects per group as described below and in the Supplemental Materials. No statistical method was used to predetermine sample size. A total of 238 subjects were recruited to participate. 22 subjects were removed from the final analysis for quitting the task early or failure to return for the memory test (n = 14), failure to understand or follow the task instructions (n = 2), equipment failures with stimulus presentation software (n = 4), or a failure to show any evidence of recognition memory above chance (n = 2). The final sample included 216 subjects assigned to 11 unique experiments as described generally below and in more detail in the Supplemental Materials and Supplemental Table 1. All subjects provided written informed consent. The procedures and consent were approved by the University Committee on Activities Involving Human Subjects at New York University (IRB#2016-2).

Task Design: General methods

Stimulus materials—Memoranda for incidental encoding and the recognition memory test were color photographs of animals and tools presented on a white background and obtained from the website http://www.lifeonwhite.com or from publicly available resources on the internet. Each picture was a unique basic level object with a different name; for example, there were not two different pictures of an elephant. Stimulus order was pseudorandomized across participants such that no more than 3 pictures from the same category occurred in a row. During encoding, pictures were never repeated.

Encoding Session—All experiments included an associative conditioning phase in which pictures from one category (animals or tools, counterbalanced between subjects; referred to as CS+) were paired with an aversive electrical shock to the right wrist (unconditioned stimulus, US). Pictures from the other category (tools or animals, respectively) were never paired with a shock (referred to as CS−). Each trial was 4.5 s and followed by an 8–10 s variable duration waiting period that included a fixation cross on a blank background.

Unless noted otherwise in the details for each study (Supplemental Materials), encoding consisted of 2 phases: associative conditioning and extinction. Associative conditioning (i.e., Pavlovian fear or threat conditioning) included 80 trials, 40 CS+ and 40 CS−. Half of the CS + trials were paired with the US (50% reinforcement rate) during conditioning. Midway through conditioning (i.e., after 20 CS+ and 20 CS− trials), there was a short explicit break in which subjects were informed that the experiment would resume shortly. The break was ~10 s and served as a perceptual event boundary. Another break occurred after the next 20

CS+ (10 paired with the US) and 20 CS− trials. For several of the extinction paradigms, the 2nd break was followed by an unbroken session of extinction that included 40 CS+ and 40 CS− trials, all unpaired with the US. The extinction session was modified across studies (see Supplemental Materials) by, for example, presenting occasional shocks (gradual extinction), replacing the shock with the tone (novelty-extinction), or adding 40 additional CS+ and 40 CS− trials (massive extinction and gradual extinction). In all studies, the shock leads remained attached and subjects were not explicitly informed that they would no longer receive any shocks.

On each trial, subjects were asked whether or not they expected a shock using two alternative-forced-choice (Yes or No). Subjects were instructed that their button presses did not affect whether or not the shock would occur, thus mitigating the chance that subjects mistakenly attributed the outcome to their action. At the end of the session subjects were asked to rate how intense the shock felt, and how much they had feared the shock on scales ranging from 1 to 9. These data are reported below for each experiment (Supplemental Table 1), and there were no differences between groups on these ratings.

Recognition memory test—Memory was tested using a surprise recognition memory test. The test occurred 24 hours after encoding with four exceptions detailed in the Supplemental Materials (immediate test after conditioning/extinction; immediate test after conditioning/gradual extinction; test 48 hours after conditioning/24 hours after delayed extinction; test 48 hours after conditioning part 1/24 hours after conditioning part 2). The test included a total of 320 trials: 40 CS+ and 40 CS− from acquisition, 40 CS+ and 40 CS− from extinction, and an equal number of new category related lures that were not shown at encoding (80 animals and 80 tools). The recognition memory test was self-paced and subjects rated each trial as old or new and their level of confidence (definitely old, maybe old, maybe new, and definitely new). The trial order was pseudorandomized during the memory test to ensure a mostly balanced presentation of old and new CS+ and CS− trials (i.e., pseudo-randomization ensured that subjects did not encounter a long string of old or new CS+ or CS− trials in a row). Encoding and memory tests occurred in the same test room around the same time of day, with the exception of studies with an immediate retrieval test.

To assess whether subjects expected the surprise memory test, subjects were asked whether they had any expectations for the experiment just prior to the memory test[2]. Subjects either indicated that they had no knowledge or expectations for the upcoming task (i.e., the surprise memory test), or indicated that they expected a continuation of the earlier experiment. No subjects indicated that they expected to have their memory tested for the pictures they had seen earlier.

All data are presented in the Supplemental Methods for each experiment. We collapsed across confidence for recognition memory, because subjects showed stronger than chance (corrected recognition) memory at both confidence levels in most cases. Analysis focusing only on high-confidence responses yielded similar results. For memory analyses we used corrected recognition (hits minus false alarms) to account for differences in response criteria and control for any bias to endorse items from either the CS+ or CS− category as "old." Notably, the false alarm rate was generally low and there were no consistent differences in

false alarms between the CS+ and CS− category. In all, data were normally distributed and variance was similar across studies.

For repeated measures ANOVA and t-tests (two-tailed), trials were binned in blocks of 20 for analysis *a priori*. This constituted the first half of conditioning, the second half of conditioning, and the first half of extinction. Analysis were considered significant at $P < .05$. Effect size for repeated measures ANOVA are partial eta squared. Effect size for paired samples t-tests were calculated using Cohen's *d*. Figures plotting trials in blocks of 10 are for visualization purposes.

For the change point analysis, average corrected recognition memory, SCRs, and expectancy ratings were binned into blocks of 4 trials to provide a more fine-scale analysis of the time-ordered data. To visualize the data, we plotted the cumulative sum of the differences between the mean values (blocks of 4) and the average of the data. As detailed in Taylor[62], this cumulative sum line will be increasingly positive as values above the overall average are added to the cumulative sum (upward slope), and then decline as values below the overall average are added to the cumulative sum (downward slope). Stable lines indicate that the running average does not change, whereas a sudden change in slope indicates a change in the running average. The Change-Point Analyser[62] was used to detect significant changes in the time series occur using 10,000 bootstrap iterations with 95% confidence.

Code availability—The experiments were coded in E-Prime 2.0 ® (Psychology Software Tools) and SPSS (IBM) was used for statistical analysis of ANOVA and paired samples t-tests. No custom code was generated for these experiments.

Shock and psychophysiology—Autonomic arousal was measured by skin conductance responses collected from pre-gelled snap electrodes (BIOPAC EL509) connected to the BIOPAC MP100 System (Goleta, CA). Electrodes were attached to the hypothenar eminence of the left palm and SCRs were calculated according to our previous criteria[2]. In brief; an SCR was considered related to stimulus presentation if the trough-to-peak deflection occurred within 0.5–4.5 s following CS onset, lasted between 0.5 and 5.0s, and was greater than 0.02 microsiemens. If an SCR did not meet these criteria, then the trial was scored as a zero. Responses were obtained using a custom Matlab (The Mathworks, Inc.) script that extracted SCRs for each trial using these criteria[63]. Although not the primary dependent measure of the present experiment, SCRs were used as a manipulation check that fear-conditioning induced higher autonomic arousal on CS+ than CS− trials during acquisition. Results from each group are presented in the Supplemental Methods. SCRs were similar between groups during the fear-conditioning phase, as expected since this phase was the same for every group.

The electric shock was a 200-ms stimulation delivered to the right wrist using pre-gelled snap electrodes (BIOPAC EL508) connected to a Grass Medical Instruments stimulator (West Warwick, Rhode Island). Prior to the start of encoding, intensity was calibrated for each subject to reach a level they deemed "highly annoying but not painful" in keeping with protocols from our lab. Shock intensity was scored on a modified pain assessment scale from 1 (=no sensation) to 9 (=very high intensity) at the end of calibration and once again at the

end of encoding to ensure that the shock did not drastically decrease or increase in subjective intensity (see Supplemental Table 1).

Data availability—The datasets generated during and/or analysed during the current study are available from the corresponding author upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Cai DJ, et al. A shared neural ensemble links distinct contextual memories encoded close in time. Nature. 2016; 534:115–118. [PubMed: 27251287]

2. Dunsmoor JE, Murty VP, Davachi L, Phelps EA. Emotional learning selectively and retroactively strengthens memories for related events. Nature. 2015

3. Rashid AJ, et al. Competition between engrams influences fear memory formation and recall. Science. 2016; 353:383–387. [PubMed: 27463673]

4. Pavlov IP. Conditioned Reflexes. Oxford University Press; 1927.

5. Pearce JM, Hall G. A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychological Review. 1980; 87:532–552. [PubMed: 7443916]

6. Konorski J. Integrative activity of the brain: An interdisciplinary approach. University of Chicago Press; 1967.

7. Bouton ME. Context, ambiguity, and unlearning: Sources of relapse after behavioral extinction. Biological Psychiatry. 2002; 52:976–986. [PubMed: 12437938]

8. Senn V, et al. Long-range connectivity defines behavioral specificity of amygdala neurons. Neuron. 2014; 81:428–437. [PubMed: 24462103]

9. Milad MR, Quirk GJ. Neurons in medial prefrontal cortex signal memory for fear extinction. Nature. 2002; 420:70–74. [PubMed: 12422216]

10. Sotres-Bayon F, Sierra-Mercado D, Pardilla-Delgado E, Quirk GJ. Gating of fear in prelimbic cortex by hippocampal and amygdala inputs. Neuron. 2012; 76:804–812. [PubMed: 23177964]

11. Tovote P, Fadok JP, Lüthi A. Neuronal circuits for fear and anxiety. Nature Reviews Neuroscience. 2015; 16:317–331. [PubMed: 25991441]

12. Dunsmoor JE, Niv Y, Daw ND, Phelps EA. Rethinking extinction. Neuron. 2015; 88:47–63. [PubMed: 26447572]

13. Brashers-Krug T, Shadmehr R, Bizzi E. Consolidation in human motor memory. Nature. 1996; 382:252. [PubMed: 8717039]

14. Müller GE, Pilzecker A. Experimentelle beiträge zur lehre vom gedächtniss. Barth JA, editorVol. 1. 1900.

15. Kurby CA, Zacks JM. Segmentation in the perception and memory of events. Trends in cognitive sciences. 2008; 12:72–79. [PubMed: 18178125]

16. Speer NK, Zacks JM. Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. Journal of Memory and Language. 2005; 53:125–140.

17. Rinck M, Bower GH. Temporal and spatial distance in situation models. Memory & Cognition. 2000; 28:1310–1320. [PubMed: 11219959]

18. Zacks JM, Tversky B. Event structure in perception and conception. Psychological bulletin. 2001; 127:3. [PubMed: 11271755]

19. Morrow DG, Greenspan SL, Bower GH. Accessibility and situation models in narrative comprehension. Journal of Memory and language. 1987; 26:165–187.

20. Ezzyat Y, Davachi L. What constitutes an episode in episodic memory? Psychological Science. 2011; 22:243–252. [PubMed: 21178116]

21. Davachi L, DuBrow S. How the hippocampus preserves order: the role of prediction and context. Trends in Cognitive Sciences. 2015; 19:92–99. [PubMed: 25600586]

22. Ezzyat Y, Davachi L. Similarity breeds proximity: pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. Neuron. 2014; 81:1179–1189. [PubMed: 24607235]

23. DuBrow S, Davachi L. The influence of context boundaries on memory for the sequential order of events. Journal of Experimental Psychology: General. 2013; 142:1277. [PubMed: 23957281]

24. DuBrow S, Davachi L. Temporal binding within and across events. Neurobiology of learning and memory. 2016; 134:107–114. [PubMed: 27422018]

25. Clewett D, Davachi L. The Ebb and Flow of Experience Determines the Temporal Structure of Memory. Current Opinion in Behavioral Sciences. 2017; 17:186–193. [PubMed: 29276730]

26. Dudai Y, Karni A, Born J. The consolidation and transformation of memory. Neuron. 2015; 88:20–32. [PubMed: 26447570]

27. Gershman SJ, Monfils M-H, Norman KA, Niv Y. The computational nature of memory modification. eLife. 2017; 6:e23763. [PubMed: 28294944]

28. Lee HJ, Berger SY, Stiedl O, Spiess J, Kim JJ. Post-training injections of catecholaminergic drugs do not modulate fear conditioning in rats and mice. Neuroscience Letters. 2001; 303:123–126. [PubMed: 11311508]

29. Wilensky AE, Schafe GE, LeDoux JE. Functional inactivation of the amygdala before but not after auditory fear conditioning prevents memory formation. J Neurosci. 1999; 19:RC48. [PubMed: 10594092]

30. McGowan J, et al. Prophylactic Ketamine Attenuates Learned Fear. Neuropsychopharmacology. 2017

31. Schiff HC, et al. β-Adrenergic Receptors Regulate the Acquisition and Consolidation Phases of Aversive Memory Formation Through Distinct, Temporally Regulated Signaling Pathways. Neuropsychopharmacology. 2017; 42:895–903. [PubMed: 27762270]

32. Bush DE, Caparosa EM, Gekker A, LeDoux J. Beta-adrenergic receptors in the lateral nucleus of the amygdala contribute to the acquisition but not the consolidation of auditory fear conditioning. Frontiers in behavioral neuroscience. 2010; 4:154. [PubMed: 21152344]

33. Maren S, Chang CH. Recent fear is resistant to extinction. Proceedings of the National Academy of Sciences of the United States of America. 2006; 103:18020–18025. [PubMed: 17090669]

34. Huff NC, Hernandez JA, Blanding NQ, LaBar KS. Delayed Extinction Attenuates Conditioned Fear Renewal and Spontaneous Recovery in Humans. Behavioral Neuroscience. 2009; 123:834–843. [PubMed: 19634943]

35. McGaugh JL. Consolidating memories. Annual review of psychology. 2015; 66:1–24.

36. Dudai Y. The neurobiology of consolidations, or, how stable is the engram? Annu. Rev. Psychol. 2004; 55:51–86. [PubMed: 14744210]

37. Robertson EM, Pascual-Leone A, Miall RC. Current concepts in procedural consolidation. Nature Reviews Neuroscience. 2004; 5:576–582. [PubMed: 15208699]

38. Swallow KM, Zacks JM, Abrams RA. Event boundaries in perception affect memory encoding and updating. Journal of Experimental Psychology: General. 2009; 138:236. [PubMed: 19397382]

39. Zacks JM, Swallow KM. Event segmentation. Current Directions in Psychological Science. 2007; 16:80–84. [PubMed: 22468032]

40. Ben-Yakov A, Eshel N, Dudai Y. Hippocampal immediate poststimulus activity in the encoding of consecutive naturalistic episodes. Journal of Experimental Psychology: General. 2013; 142:1255–1263. [PubMed: 23815458]

41. Patil A, Murty VP, Dunsmoor JE, Phelps EA, Davachi L. Reward retroactively enhances memory consolidation for related items. Learning & Memory. 2017; 24:65–69. [PubMed: 27980078]

42. Dunsmoor JE, Martin A, LaBar KS. Role of conceptual knowledge in learning and retention of conditioned fear. Biological Psychology. 2012; 89:300–305. [PubMed: 22118937]

43. Herry C, et al. Switching on and off fear by distinct neuronal circuits. Nature. 2008; 454:600–606. [PubMed: 18615015]

44. Gallistel CR, Fairhurst S, Balsam P. The learning curve: implications of a quantitative analysis. Proc Natl Acad Sci U S A. 2004; 101:13124–13131. [PubMed: 15331782]

45. Rescorla RA, Wagner AR. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Appleton-Century-Crofts; 1972.

46. Gershman SJ, Jones CE, Norman KA, Monfils M-H, Niv Y. Gradual extinction prevents the return of fear: implications for the discovery of state. Frontiers in behavioral neuroscience. 2013; 7

47. Denniston JC, Chang RC, Miller RR. Massive extinction treatment attenuates the renewal effect. Learning and Motivation. 2003; 34:68–86.

48. Dunsmoor JE, Campese VD, Ceceli AO, LeDoux JE, Phelps EA. Novelty-facilitated extinction: providing a novel outcome in place of an expected threat diminishes recovery of defensive responses. Biological Psychiatry. 2015; 78:203–209. [PubMed: 25636175]

49. Gunther LM, Denniston JC, Miller RR. Conducting exposure treatment in multiple contexts can prevent relapse. Behaviour research and therapy. 1998; 36:75–91. [PubMed: 9613018]

50. Shohamy D, Adcock RA. Dopamine and adaptive memory. Trends Cogn Sci. 2010; 14:464–472. [PubMed: 20829095]

51. Adolphs R. The Biology of Fear. Current Biology. 2013; 23:R79–R93. [PubMed: 23347946]

52. Talmi D. Enhanced emotional memory: Cognitive and neural mechanisms. Current Directions in Psychological Science. 2013; 22:430–436.

53. Plendl W, Wotjak CT. Dissociation of within-and between-session extinction of conditioned fear. The Journal of Neuroscience. 2010; 30:4990–4998. [PubMed: 20371819]

54. Gershman SJ, Hartley CA. Individual differences in learning predict the return of fear. Learning & behavior. 2015:1–8. [PubMed: 25488021]

55. Craske MG, et al. Optimizing inhibitory learning during exposure therapy. Behaviour Research and Therapy. 2008; 46:5–27. [PubMed: 18005936]

56. Myers KM, Ressler KJ, Davis M. Different mechanisms of fear extinction dependent on length of time since fear acquisition. Learning & Memory. 2006; 13:216–223. [PubMed: 16585797]

57. Golkar A, Öhman A. Fear extinction in humans: Effects of acquisition–extinction delay and masked stimulus presentations. Biological psychology. 2012; 91:292–301. [PubMed: 22898744]

58. Shiban Y, Wittmann J, Weißinger M, Mühlberger A. Gradual extinction reduces reinstatement. Frontiers in behavioral neuroscience. 2015; 9

59. Bouton ME. Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. Psychological bulletin. 1993; 114:80–99. [PubMed: 8346330]

60. Ballarini F, Moncada D, Martinez MC, Alen N, Viola H. Behavioral tagging is a general mechanism of long-term memory formation. Proc Natl Acad Sci U S A. 2009; 106:14599–14604. [PubMed: 19706547]

61. Wang SH, Redondo RL, Morris RG. Relevance of synaptic tagging and capture to the persistence of long-term potentiation and everyday spatial memory. Proc Natl Acad Sci U S A. 2010; 107:19537–19542. [PubMed: 20962282]

62. Taylor WA. Change-point analysis: a powerful new tool for detecting changes. 2000. <http://www.variation.com/cpa/tech/changepoint.htm>

63. Green SR, Kragel PA, Fecteau ME, LaBar KS. Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. International Journal of Psychophysiology. 2013; 91
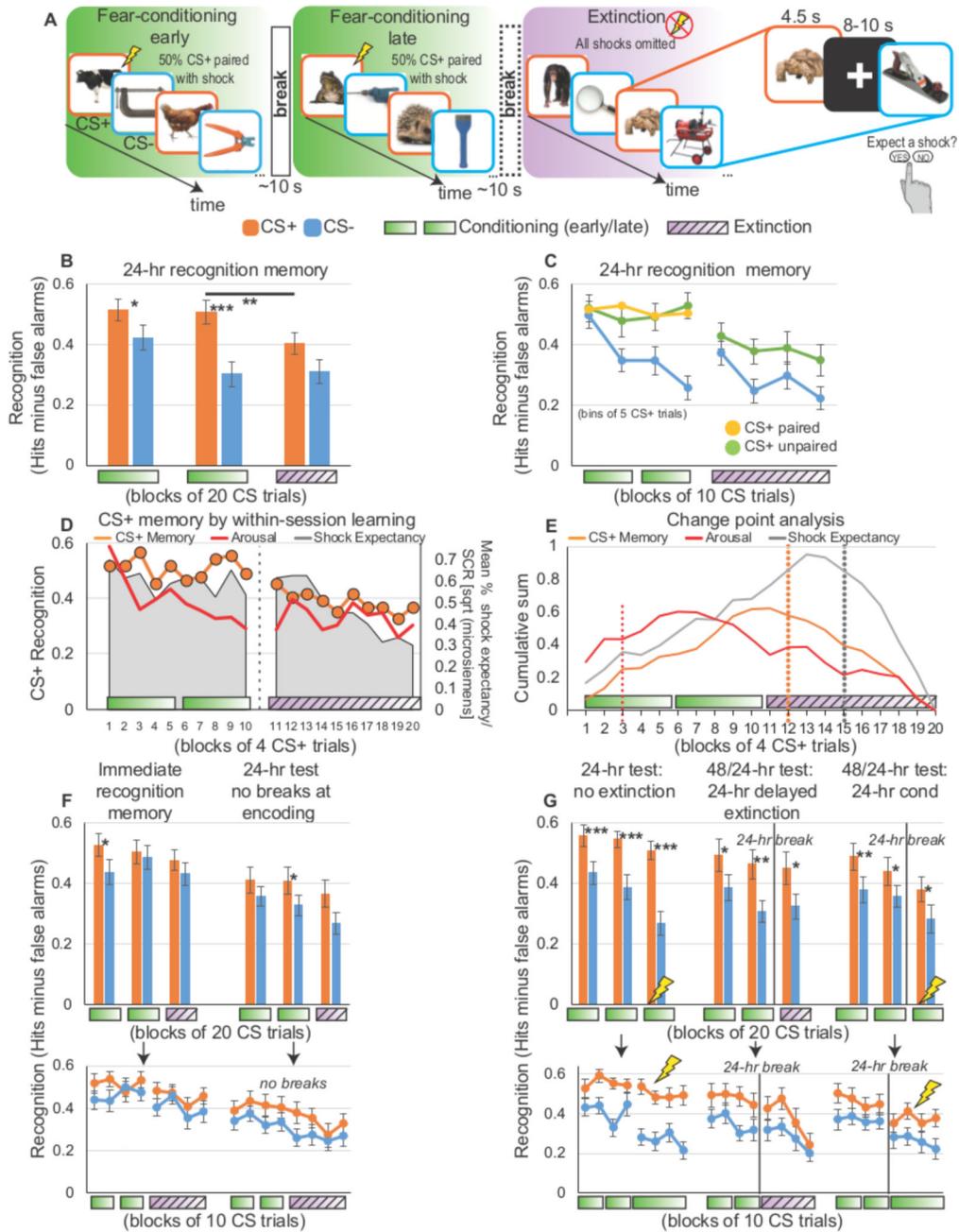
Figure 1. Emotional learning selectively prioritizes, and extinction diminishes, episodic memories for conceptually related items encoded close in time

A. Incidental encoding of animal and tool conditioned stimuli during fear conditioning and extinction. A brief (~10 s) explicit rest separated early and late conditioning and conditioning from extinction. The rest served as a perceptual event boundary. A surprise recognition memory test was administered 24-hours later. Color borders and shading are for illustrative purposes. **B.** A repeated-measures ANOVA on recognition memory for subjects who underwent immediate extinction and tested 24-hours later revealed (n = 20) a significant main effect of CS type (CS+, CS−; $F_{1,19} = 10.061$, $P = .009$, partial eta squared = .311) and Phase (early fear-conditioning, late fear-conditioning, early extinction; $F_{2,38} =$

7.071, P = .002, partial eta squared = .271) and a tendential CS type by Phase interaction, $F_{2,38}$ = 3.045, P = .059, partial eta squared = .138. **C.** There was no difference between CS+ trials paired or unpaired with shock. **D.** Shock expectancy and conditioned autonomic arousal on CS+ trials (gray area and red line, respectively) indicated that subjects still expected shock and were physiologically responding to the CS+ during the early trials of extinction; however, 24-hr recognition memory for these trials was low (orange circles). **E.** Change-point analysis revealed a significant decrease in memory soon after the boundary (dotted orange line), preceding the chance in expectancy (dotted gray line) at the time of encoding. **F.** At immediate recognition (n =20 **left bars**) and 24-hour recognition following encoding with no breaks (n = 20 **right bars**) there was significantly greater memory for CS+ than CS− from early or late conditioning (respectively), but no drop in CS+ memory following the transition to safety. See supplemental for full ANOVAs. **G.** Memory was selectively enhanced for CS+ versus CS− from each phase of learning, but there was no drop in CS+ memory across event boundaries at 24-hr retrieval following conditioning without extinction (**n = 20 left bars**); at retrieval 24-hours following 24-hour delayed extinction (n = 18 **middle bars**); or at retrieval 24-hours following additional conditioning trials 24-hours after the first half of conditioning (n = 20 **right bars**). See Supplemental Materials for full ANOVAs. Error bars reflect SEM; *** = P < .001; ** = P < .01; * = P < .05.
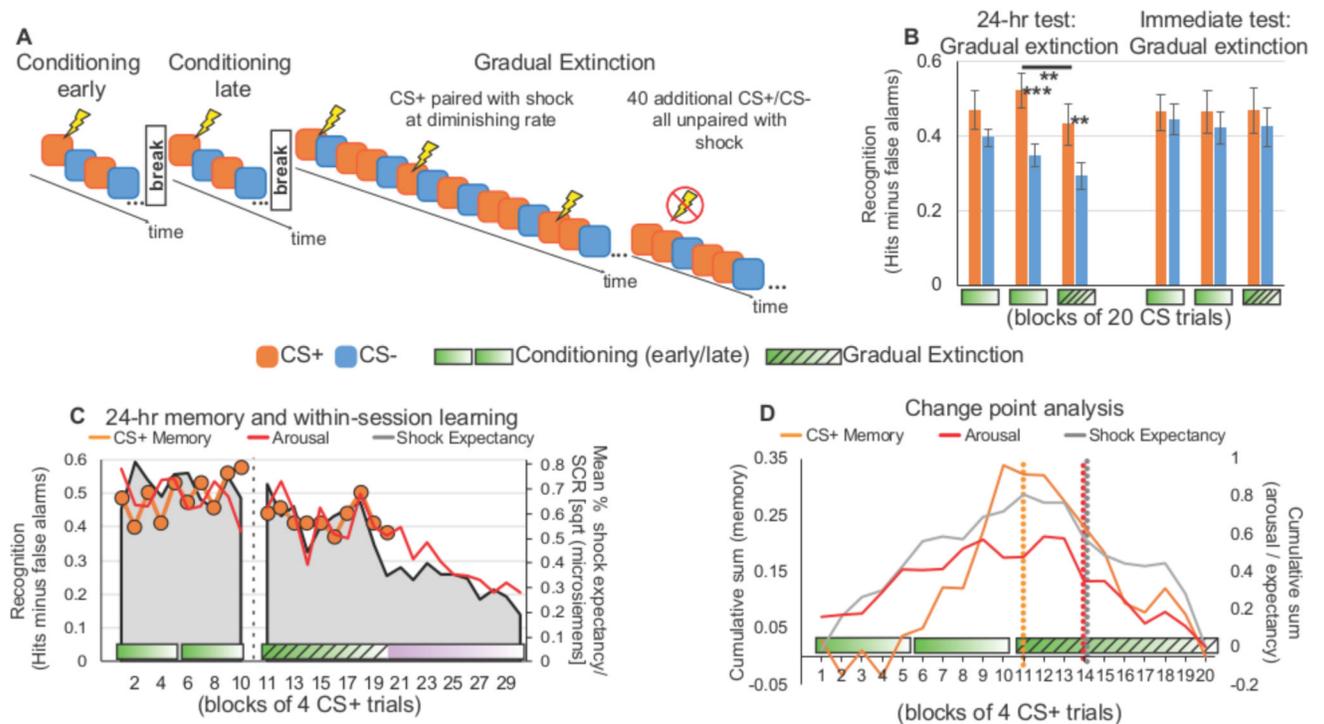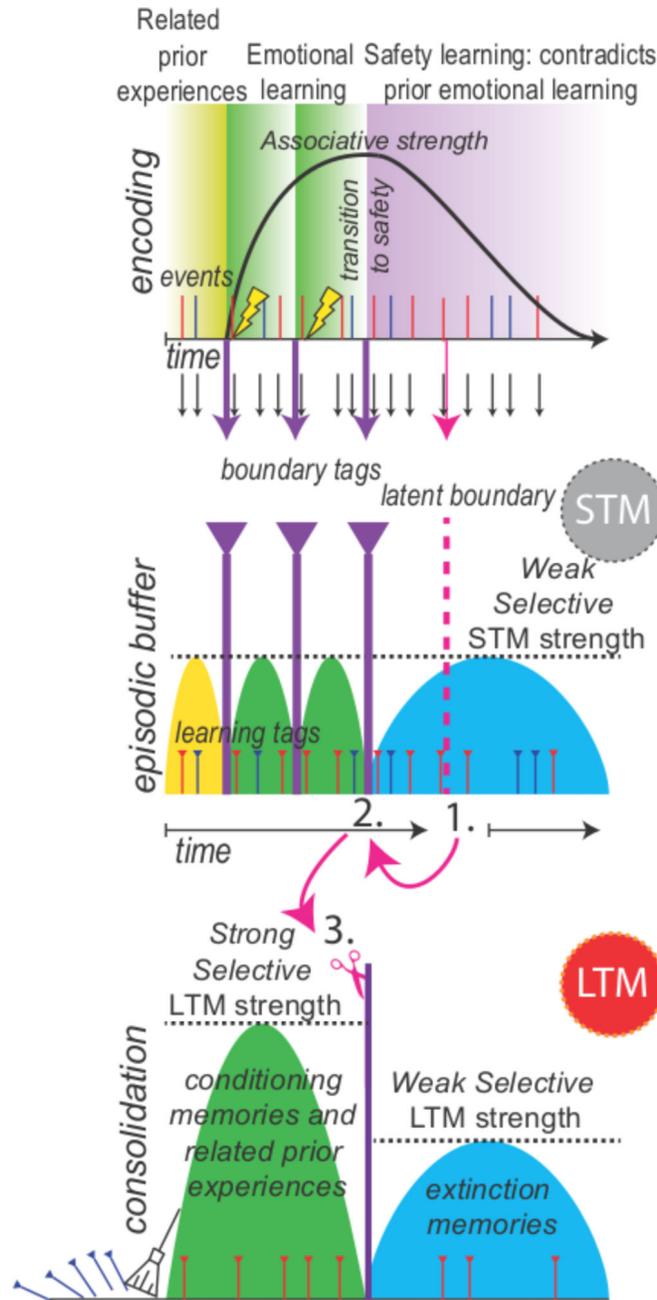
Figure 2. Event boundaries segment memory despite occasional shocks during extinction. A-right bars

A. Gradual extinction paradigm. **B.** A repeated-measures ANOVA on recognition memory for subjects who underwent gradual extinction and tested 24-hours later (n = 17 **left bars**) revealed a main effect of CS type (CS+, CS−; $F_{1,16}$ = 10.577, P = .005, partial eta squared = .398) and Phase (early fear-conditioning, late fear-conditioning, early extinction; $F_{2,32}$ = 4.363, P = .021, partial eta squared = .214) and a tendential CS type by Phase interaction, $F_{2,32}$ = 2.900, P = .070, partial eta squared = .153). There was a decrease in memory for CS+ items encoded within the boundary of fear-conditioning versus extinction (paired t-test, t_{16}=3.031, P=0.008). For subjects who underwent gradual extinction and tested immediately after encoding (n = 20 **right bars**), there was no effect of CS type ($F_{1,19}$ = 1.157, P = .296, partial eta squared = .057), Phase ($F_{2,38}$ = .105, P = .901, partial eta squared = .005) nor an interaction, $F_{2,38}$ = .209, P = .812, partial eta squared = .011). **C.** Mean shock expectancy (gray area, calculated as mean % expectancy) and skin conductance responses (red line, measured in microsiemens) on CS+ trials indicated that subjects still expected shock during gradual extinction; however, 24-hr recognition memory for these trials was low (orange circles). **D.** Change point analysis revealed that the decrease in CS+ memory (orange dotted line) occurred immediately after the boundary separating conditioning from extinction, preceding the decline in expectancy (gray dotted line) and arousal (red dotted lines) measured during encoding. Error bars reflect SEM; *** = P < .001; ** = P < .01; * = P < .05.
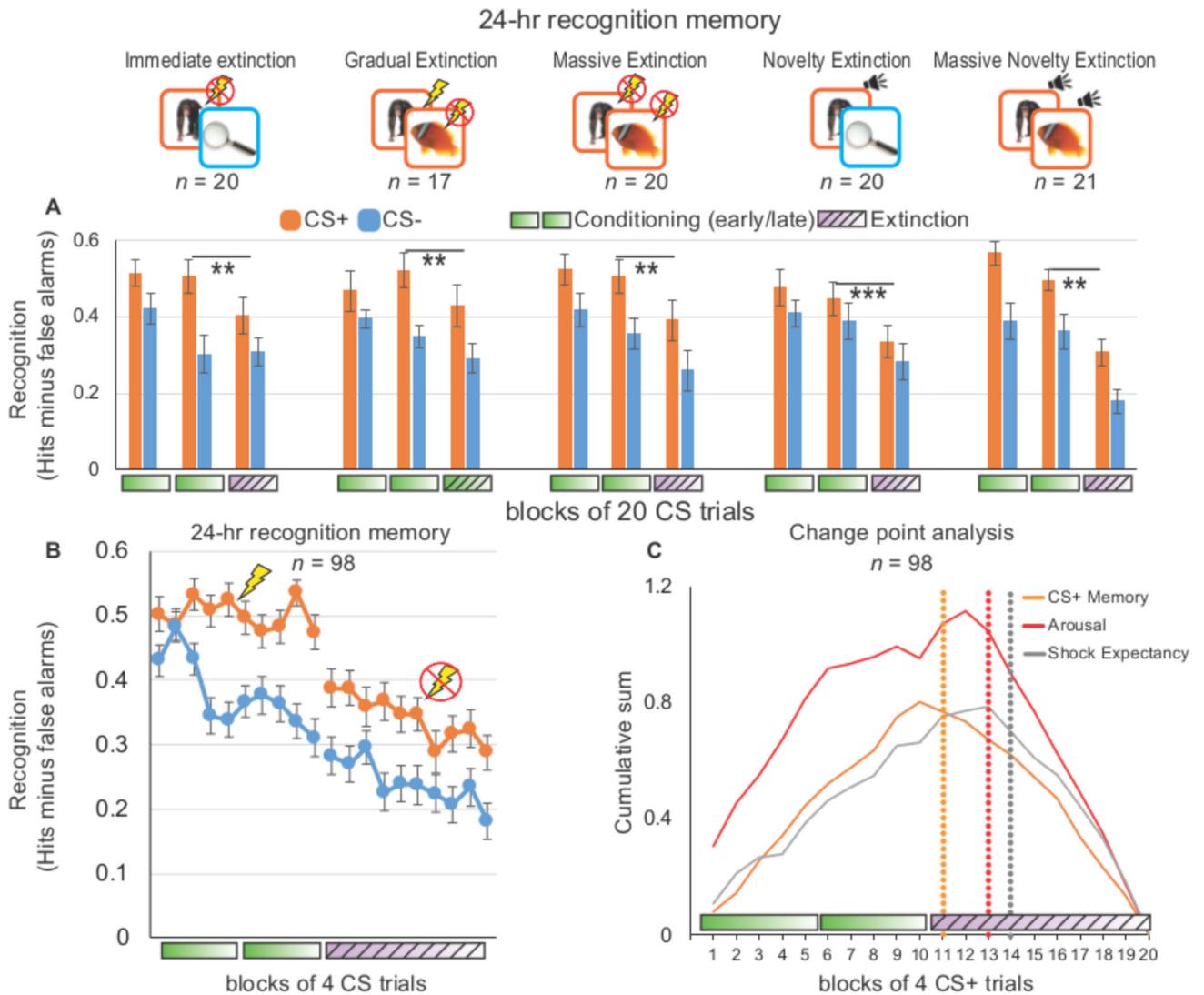
Figure 3. Selective episodic memory prioritization for threat withstands a variety of optimized extinction protocols

A. A variety of extinction protocols did not weaken original fear memories. **B.** When data is combined from all one-day extinction protocols with event boundaries and 24-hour retrieval, the drop in CS+ memory from conditioning to extinction memory is unambiguous. **C.** Change point analysis shows a significant decrease in memory for items encoded immediately after the boundary (orange dotted line), preceding the decline in expectancy (gray dotted line) and arousal (red dotted lines) measured during encoding. Error bars reflect SEM; *** = $P < .001$; ** = $P < .01$; * = $P < .05$.

Figure 4. A hypothesized mechanism by which perceptual event boundaries automatically segment long-term threat and extinction memories

A hypothesized mechanism by which event boundaries set a weak learning tag (purple lines). After the transition to safety, the last episodic boundary is credited as the relevant marker separating threat from safety once extinction is realized (**1.** Latent boundary). Memory segmentation selectively prioritizes memory for information preceding the event boundary (**2.**) at the expense of information encoded after the event boundary and preceding the latent boundary (**3.**). These *selective* memory effects (i.e., memory for threat greater than memory for safety) are enhanced following a period of consolidation.