Preliminary evidence indicates that selective maintenance of adverse events may explain conditioning phenomena attributed to fear generalization

Deepta Chandrasekhar (deepta@princeton.edu)
Princeton Neuroscience Institute
Princeton University, Princeton, NJ 08544 USA

Isabel M. Berwian (iberwian@princeton.edu)
Princeton Neuroscience Institute & Department of Psychology
Princeton University, Princeton, NJ 08544 USA

Abstract

Understanding how fear memories are maintained over time is crucial for improving the effectiveness of anxiety treatments. Previous work suggests that fear generalizes over time from the stimuli associated with aversive events to other, similar, stimuli, and that such stimuli are remembered better due to their association with potential aversive events. At the same time, a recent model showed that memory maintenance that is specific to the feared stimuli can explain phenomena such as why a fear response that is extinguished sometimes returns over time. Here, we test the prediction of this model that memory will be better for the specific stimuli associated with aversive events, and not for similar stimuli that were not followed by an aversive event. N=441 participants completed an online-administered fear-conditioning task with trial-unique stimuli from two categories (animals and objects) and a subsequent surprise recognition memory test. Preliminary results indicate that participants had better memory for the category of stimuli associated with aversive events compared to the stimulus category that was not associated with an aversive event. However, this effect was mainly driven by memory of the specific stimuli from trials with the aversive event. These results support the idea that memories in this paradigm are primarily organized according to emotional rather than semantic similarity, as has been shown in other domains as well.

Keywords: spontaneous recovery of fear; memory; fear conditioning; computational psychiatry

Introduction

Fear is often acquired through learning that a neutral cue (conditional stimulus, "CS+") is associated with an aversive outcome (unconditional stimulus, "US"). Such learning can be studied in fear conditioning paradigms, which often include a comparison neutral cue that is never associated with the US ("CS-"). The acquired fear can be reduced through extinction, which involves presenting the CS+ without the US.

Two widely observed phenomena are fear generalization and spontaneous recovery of fear after extinction. In fear generalization, individuals also show fear of stimuli that are similar or related to the CS+, even if those were never experienced with aversive outcomes. This generalization is also seen in memory. Interestingly, memory tests immediately after fear conditioning show no (Dunsmoor et al., 2018) or very small (Dunsmoor et al., 2015) differences between memory for CS+ (e.g., animals) vs. CS- (e.g., objects) images. However, after 24 hours, recognition memory for CS+ images is significantly better than for CS- images.

Spontaneous recovery of fear (Rescorla, 2004) is the well-studied phenomenon of return of fear of the CS+ with the passage of time after successful extinction. We have recently explained spontaneous recovery as resulting from selective maintenance of memories associated with aversive outcomes. This computational model captures the behavior observed in a fear conditioning paradigm better than a set of alternative models (Berwian et al., 2024), revealing that spontaneous recovery can only occur when aversive and non-aversive events are stored in distinct memories. Selective maintenance of aversive memories then strengthens these memories with time, giving them a competitive advantage during later retrieval, and therefore rekindling fear. This model predicts that memory for the specific CS+ items accompanied by the US should be better than memory of other CS+ items. However, previous studies did not report such memory differences.

In this preliminary analysis of a dataset from a category-conditioning task, we teased apart whether the difference in memory between CS+ and CS- images was driven by enhancement of memories along categories (which would predict better memory for all CS+ images) or by better memory of aversive events and their specific associated CS+ images.

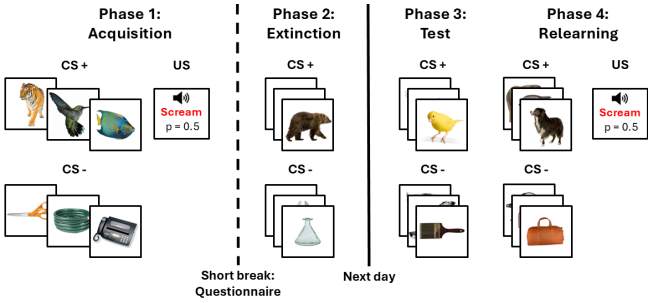

Figure 1: Fear conditioning task overview

Methods & Results

We designed an online-administered fear-conditioning paradigm (Berwian et al., 2024) with trial-unique stimuli used in category-conditioning studies (Hennings et al., 2021). Participants viewed images of animals (CS+) and images of objects (CS-), each image presented only once, across four phases: acquisition, extinction, spontaneous recovery test, and relearning (Fig. 1). Phases included 40, 48, 20 and 20 trials, respectively, with equal numbers of CS+ and CS- trials in each phase. On each trial, participants were asked to press a key for the image to flip around and reveal the outcome (US or no US). In acquisition and relearning, 50% of CS+ images were followed by an aversive but tolerable auditory scream (US), otherwise no outcome occurred. Every few trials, participants were asked to rate their expectations of how likely a scream would follow an animal or an object, on sliding scales ranging between 0-100% (Fig. 2A).

The experiment consisted of two sessions. The first session included acquisition, and after a short break (3-5 minute survey), extinction. On the following day, participants completed the spontaneous recovery test and relearning phases, followed by a surprise recognition memory test. In the memory test, participants viewed 150 images (75 each of animals and objects), of which 88 were previously seen in the acquisition and extinction phases, and 62 were novel 'foil' images. They were asked to make confidence judgments on whether

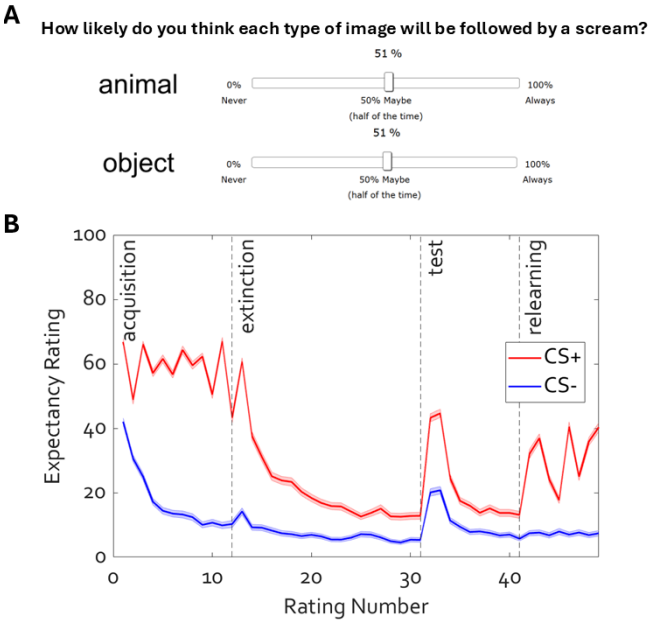

Figure 2: A) Expectancy rating scale for each category. B) Solid lines show average ratings of the CS+ animal category (red) and CS- object category (blue) across participants and shades show 95% bootstrapped confidence intervals. Dashed horizontal lines indicate the end of each phase.

the images were 'Definitely New', 'Maybe New', 'Maybe Old' or 'Definitely Old'. Trial order was identical for all participants. To ensure remote participants heard the US loudly enough, 7 auditory attention checks were spread throughout the task.

Procedure, participants and statistical effects.

The study was approved by the Institutional Review Board of Princeton University. Participants were recruited on Prolific, provided informed consent and were compensated for their time. Participants completed the experiment across two consecutive days. After excluding participants with incomplete datasets or failed attention checks, N=441 remained.

We computed corrected recognition as the high-confidence hit rate minus the high-confidence false alarm rate (each defined as the number of images judged as 'Definitely Old' in the corresponding category or subset thereof). First, we compared corrected recognition for CS+ and CS- from the acquisition phase. In line with prior work (Dunsmoor et al., 2015), we found that recognition memory of the CS+ category was better $(t = 7.85, p < 0.001)$. Next, we teased apart whether this effect was driven by better memory for stimuli within the CS+ category that were followed by a US. Thus, we compared memory for "reinforced CS+" (those followed by a US), "non-reinforced CS+", and CS- images from the acquisition phase. We found that recognition performance for reinforced CS+ images significantly exceeded that of non-reinforced CS+ images $(t = 6.50, p < 0.001$; Fig. 3). Additionally, memory for non-reinforced CS+ images was not significantly different from memory for CS- images in the acquisition phase $(t = 1.22, p = 0.223)$. In contrast, in the extinction phase, memory for CS+ images (that were *not* reinforced) was significantly better than for CS- images $(t = 5.07, p < 0.001)$.

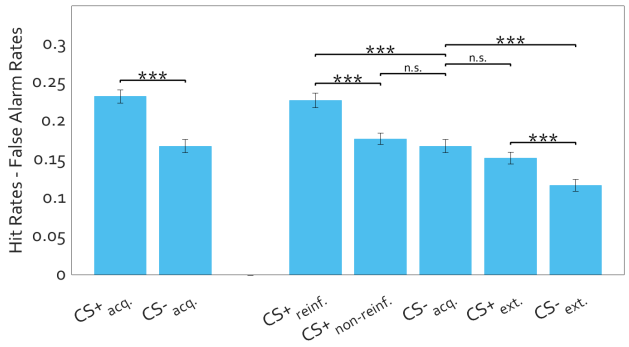

Figure 3: Corrected recognition for CS+ and CS- stimuli in the acquisition phase (2 bars on left), and for subcategories of stimuli (see text) throughout acquisition and extinction (right).

Discussion

Our preliminary results indicate that enhanced memory of CS+ images from acquisition was driven by enhanced memory of stimuli that were followed by aversive outcomes and not by enhanced memory of all CS+ category images. However, during extinction, memory was better for (non-reinforced) CS+ images than for CS- images.

We propose that these findings are consistent with selective maintenance of contexts (or memories) associated with aversive outcomes. For delayed memory enhancement of categories to occur (e.g., better memory of non-reinforced CS+ images in extinction), contexts of remembered events have to be distinguished based on semantic similarity. In contrast, for spontaneous recovery to take place, contexts for memories have to be determined based on emotional similarity. Importantly, both types of memory organization can coexist, but might also compete with each other.

Our task design and analyses have important limitations. First, we did not randomize categories or stimuli across participants, which could lead to biases in memory effects due to general differences in memorability of individual items. Second, our design differed from traditional category-conditioning experiments as we explicitly instructed participants about the categories and had them rate their expectations for each category (Fig. 2A) rather than for individual items, possibly interfering with implicit learning that could have driven category-based memory benefits in prior work. Finally, our enhanced memory findings can be explained by stronger encoding of reinforced stimuli rather than later selective maintenance of memories. We plan to address these limitations in future work.

Acknowledgement

We would like to thank Yael Niv and Augustin C. Hennings for their invaluable feedback on study design, interpretation of results, and manuscript.

References

Berwian, I. M., Pisupati, S., Chiu, J., Ren, Y., & Niv, Y. (2024). Selective maintenance of negative memories as a mechanism of spontaneous recovery of fear after extinction. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 46).

Dunsmoor, J. E., Kroes, M. C., 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.

Dunsmoor, J. E., Murty, V. P., Davachi, L., & Phelps, E. A. (2015). Emotional learning selectively and retroactively strengthens memories for related events. *Nature*, *520*.

Hennings, A. C., Lewis-Peacock, J. A., & Dunsmoor, J. E. (2021). Emotional learning retroactively enhances item memory but distorts source attribution. *Learning & Memory*, *28*(6), 178–186.

Rescorla, R. A. (2004). Spontaneous recovery. *Learning & Memory*, *11*(5), 501–509.