Reviewed Preprint
v1 • July 8, 2024
Not revised

Reviewed Preprint
v2 • March 12, 2025
Revised by authors

Reviewed Preprint
v3 • December 4, 2025
Revised by authors

**For correspondence:** leekin@gmail.com

* Equal contribution

**Competing interests:** No competing interests declared

**Funding:** See page 19

**Reviewing editor:** Mihaela D Iordanova, Concordia University, Canada



# Dissociable memory modulation mechanisms facilitate fear amnesia at different timescales

**Yinmei Ni[1,*], Ye Wang[1,*], Zijian Zhu[2], Jingchu Hu[3], Daniela Schiller[4], Jian Li[1,5]**

[1]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behavior and Mental Health, Peking University, Beijing, China • [2]School of Psychology, Shaanxi Normal University, Xi'an, China • [3]Department of Anxiety Disorders, Shenzhen Kangning Hospital, Shenzhen Mental Health Institute, Shenzhen, China • [4]Departments of Psychiatry and Neuroscience, and Friedman Brain Institute, Icahn School of Medicine at Mount Sinai, New York, United States • [5]IDG/McGovern Institute for Brain Research, Peking University, Beijing, China

## eLife Assessment

This manuscript presents **valuable** findings which reveal an intricate pattern of memory expression following retrieval extinction at different intervals from retrieval-extinction to test. The novel advance is in the demonstration that, relative to a standard extinction procedure, the retrieval-extinction procedure more effectively suppresses responses to a conditioned threat stimulus when testing occurs just minutes after extinction. While the data provide **solid** evidence that the "short-term" suppression of responding involves engagement of the dorsolateral prefrontal cortex, there are inconsistencies in the analyses reported which obscure the interpretation and leave some of the claims with limited evidence.

https://doi.org/10.7554/eLife.98652.3.sa3

## Abstract

Memory reactivation renders consolidated memory fragile and thereby opens the window for memory updates, such as memory reconsolidation. However, whether memory retrieval facilitates update mechanisms other than memory reconsolidation remains unclear. We tested this hypothesis in three experiments with healthy human participants. First, we demonstrate that memory retrieval-extinction protocol prevents the return of fear expression shortly after extinction training and this short-term effect is memory reactivation dependent (Study 1, $N$ = 57 adults). Furthermore, across different timescales, the memory retrieval-extinction paradigm triggers distinct types of fear amnesia in terms of cue-specificity and cognitive control dependence, suggesting that the short-term fear amnesia might be caused by different mechanisms from the cue-specific amnesia at a longer and separable timescale (Study 2, $N$ = 79 adults). Finally, using continuous theta-burst stimulation (Study 3, $N$ = 75 adults), we directly manipulated brain activity in the dorsolateral prefrontal cortex, and found that both memory reactivation and intact prefrontal cortex function were necessary for the short-term fear amnesia after the retrieval-extinction protocol. The differences in temporal scale, cue-specificity, and cognitive control ability dependence between the short- and long-term amnesia suggest that memory retrieval and extinction training trigger distinct underlying memory update mechanisms. These findings suggest the potential involvement of coordinated memory modulation processes upon memory retrieval and may inform clinical approaches for addressing persistent maladaptive memories.

## Introduction

Memory retrieval provides a unique window through which consolidated memories can be modified or even neutralized. Studies of memory reconsolidation suggest that memories are rendered fragile each time they are retrieved (Alberini, 2005; Dudai, 2006; Nader, Schafe, & LeDoux, 2000), possibly due to the disruption of the old memory trace. The reconsolidation of the new memory trace also requires *de novo* protein synthesis, which usually takes hours to complete. Pharmacological blockade of protein synthesis and behavioral interventions can both diminish the original fear memory expression in the long-term (24 hours later) memory test (Lee, 2008; Lee et al., 2017; Schiller et al., 2013; Schiller et al., 2010), resulting in the cue-specific fear memory deficit (Debiec et al., 2002; Lee, 2008; Nader, Schafe, & LeDoux, 2000). For example, during the reconsolidation window, retrieving a fear memory allows it to be updated through extinction training (i.e., the retrieval-extinction paradigm (Agren, 2014; Agren et al., 2012; Lee, 2008; Lee et al., 2017; Schiller et al., 2013; Schiller et al., 2010), but also see (Chalkia, Schroyens, et al., 2020; Chalkia, Van Oudenhove, & Beckers, 2020; Schiller, 2020).

However, the exact functional role of memory retrieval remains unclear. For example, while previous fear memory studies typically reported the reconsolidation effects emerging several hours later (the long-term effect) (Kindt and Soeter, 2018; Lee, 2008; Speer et al., 2021), the temporal dynamics of the amnesia effect through behavioral manipulation such as retrieval-extinction is often overlooked in the fear memory literature (Lee et al., 2017; Schiller et al., 2013; Schiller et al., 2010). It is possible that memory retrieval might facilitate memory modification at different temporal scales. This is in stark contrast to semantic and episodic memory studies where the behavioral manipulations such as the retrieval-induced forgetting (RIF) protocol or direct suppression result in short-term memory deficit within 30 minutes (Anderson et al., 2000; Anderson et al., 1994). More recently, behavioral strategies originally developed in the declarative memory studies such as direct suppression had also been successfully applied to the fear memory paradigms in humans and prevented the return of fear shortly afterward (30mins), suggesting that the fear memory might be amenable to modulation and this effect can manifest at a more immediate timescale, to which the memory reconsolidation theory remain agnostic (Wang et al., 2021).

Corresponding to the long-term behavioral manifestation, concurrent neural evidence supporting memory reconsolidation hypothesis emphasizes that synapse degradation and *de novo* protein synthesis are required for reconsolidation. These processes typically take several hours, if not longer, and are often aided by overnight sleep (Kindt and Soeter, 2018). On the contrary, previous behavioral manipulations engendering the short-term amnesia on declarative memory, such as the think/no-think paradigm, hinges on the intact activities in brain areas such as dorsolateral prefrontal cortex (cognitive control) and its functional coupling with specific brain regions such as hippocampus (memory retrieval) (Anderson and Green, 2001; Wimber et al., 2015). The declarative amnesia effect arises much earlier due to the more instant modulation of functional connectivity, rather than the slower processes of new protein synthesis in these brain regions (Wang et al., 2021). However, it remains unclear whether experimental paradigms such as memory retrieval-extinction, a method developed to induce long-term prevention of fear expression based on the memory reconsolidation hypothesis, might also trigger short-term amnesia for the fear memory.

To fully comprehend the temporal dynamics of the memory retrieval-extinction training effect, we first carried out a simple fear memory study by pairing a neutral cue (conditioned stimulus, CS) with the electric shock (unconditioned stimulus, US) to directly test whether the fear retrieval-extinction paradigm would yield a short-term effect on fear expression. We then performed another double cue CS-US fear memory experiment to study the temporal characteristics of fear memory malleability to behavioral extinction at the short-(30mins), medium- (6 hours) and long-term (24 hours) time scales.

We hypothesize that the labile state triggered by the memory retrieval may facilitate different memory deficits following extinction training, and these deficits can be further disentangled through the lens of temporal dynamics and cue-specificities. In theory, different cognitive mechanisms underlying specific fear memory deficits, therefore, can be inferred based on the difference between memory deficits. Specifically, memory reconsolidation effect, if exists, should only be evident in the long-term (24h) memory test due to the requirement of new protein synthesis and should be cue-dependent (Monfils et al., 2009). In contrast, if more immediate memory update mechanisms are involved, then the fear memory deficit should be observed relatively early (30mins) after extinction training (Anderson and Floresco, 2022; Anderson and Green, 2001). There may even exist a "limbo" state: when the immediate memory update effects dissipate and the reconsolidation updating is yet to take effect (6h), the conditioned stimuli should still elicit fear responses, regardless of the reminder (cue). Previous research in the short-term declarative and fear memory deficits highlighted the role of cognitive control. Subjects with stronger control over intrusive thoughts or stronger functional connectivity between brain regions such as the dorsolateral prefrontal cortex (dlPFC) and hippocampus exhibited more pronounced amnesia in the memory suppression tasks (Anderson et al., 1994; Wells and Davies, 1994; Wimber et al., 2015). We reason that if the retrieval related short-term fear amnesia is also related to cognitive control and the dlPFC activities, then the individual difference of such immediate fear memory deficits should be associated with individual subject's control abilities over intrusive thoughts (Küpper et al., 2014), which can be measured via the Thought-Control Ability Questionnaire (TCAQ) scale.

Indeed, through a series of experiments, we identified a short-term fear amnesia effect following memory retrieval-extinction training, in addition to the long-term amnesia effect that appeared much later. Finally, by applying the continuous theta-burst stimulation (cTBS) protocol over the dlPFC, the brain region critical for cognitive control and goal maintenance, we showed that both memory retrieval and intact dlPFC activity were necessary for the more immediate fear amnesia effect.

# Methods

## Participants

All participants were students recruited from Peking University. They were right-handed with normal vision and had not participated in electric shock-related experiments before. None of the participants had current or history of psychiatric illness, nor any contraindication on the application of transcranial magnetic stimulation (TMS) (Rossi et al., 2009; Rossini et al., 2015). All participants provided informed consent and were remunerated for their participation. This study was approved by the institutional review board of school of psychological and cognitive sciences at Peking University.

We first conducted a power analysis (G*Power) to determine the number of participants sufficient to detect a reliable effect (Faul et al., 2009). Based on the reported effect size of the reinstated fear memory between the treatment and control groups on fear memories reported in the previous literature ( $^{2}$ = 0.19, 0.216, 0.24; N = 40, 55, 30 of sample size, respectively (Kindt and Soeter, 2018; Kindt et al., 2009; Wang et al., 2021)), 52 participants for both groups (study 1: reminder group and no-reminder group) were enough to detect a significant interaction effect (α = 0.05, β = 0.9, 2 (groups) × 2 (phases) two-way ANOVA interaction effect). Similarly, a total of 72 participants for three groups (study 2: 30-min group, 6-h group and 24-h group) were required to detect a reliable effect (α = 0.05, β = 0.9, 3 (groups) × 2 (phases) × 2 (CS+) three-way ANOVA interaction effect) (Heo and Leon, 2010). For the rTMS study, based on the reported effect size of reinstated fear memory ( $^{2}$ = 0.144, 0.19, 0.216, 0.24; N =79, 40, 55, 30 total sample size, η respectively (Kindt and Soeter, 2018; Kindt et al., 2009; Wang et al., 2021)), a total of 72 participants (study 3) were needed to detect a significant interaction effect (α = 0.05, β = 0.8, 4 (groups) × 2 (phases) × 2 (CS+) three-way ANOVA interaction effect).

In study 1, we first recruited a total of 75 human subjects (41 females; mean age= 21.3, SD= 2.21). Eight participants discontinued after Day 1 due to their non-measurable spontaneous conditioned stimulus (CS) signal (the skin conductance response, SCR) towards different CSs (non-responders, SCR response to any CS was less than 0.02 μS, n = 0 & 8; reminder group and no-reminder group, respectively). Ten participants (7 females; mean age = 20.6, SD = 1.79) who were "non-learners" during fear acquisition (n = 3 & 1 in both groups; respectively) or extinction (n = 3 & 3, respectively) phase were also excluded from further analysis. The criteria for "non-learners" in the fear acquisition include smaller mean SCR of CS+ (CS that was partially paired with the electric shock) than that of the CS- (CS that was never paired with shock) in the trials in the latter half of acquisition, and a smaller mean SCR difference between CS+ and CS- in the latter than the first half trials of fear acquisition on day 1. Similarly, the criteria for "non-learners" in the fear extinction are larger mean SCRs of CS+ than the CS-responses in both the last trial and the latter-half trials of extinction and a larger mean SCR difference between CS+ and CS-in the latter than the first half trials during fear extinction on day 2 (see the exclusion criteria Table S1 in the Supplemental Material). Therefore, we had a total of 57 subjects for data analysis in study 1 (30 for the reminder group, 17 females, mean age = 21.4, SD = 2.37, and 27 for the no-reminder group, 15 females, mean age = 21.62, SD = 2.13). The group × gender interaction effect on participants' age was not statistically significant ($F_{1,49}$ = 0.137, $P$ = 0.713, $\eta^2$ = 0.003, two-way ANOVA), and gender ($x^2$ = 0.007, $P$ = 0.933) and age ($t_{55}$ = -0.376, $P$ = 0.708) were not significantly different between two groups.

In study 2, we first recruited 119 participants (56 females; mean age= 20.85, SD= 2.59). Thirty-seven participants discontinued after Day 1 for they were "non-responders" (n = 19, 5 and 13; 30-min group, 6-h group and 24-h group, respectively). Additionally, two participants (n = 1&1; 30-min group and 24-h group; respectively) failed to show the evidence of fear acquisition and one participant (30-min group) failed to show the evidence of fear extinction on day 2 and were also excluded from further analysis. Finally, we had a total of 79 subjects for data analysis in study 2 (27 for the 30-min group, 15 females, mean age = 21.4, SD = 2.46; 26 for the 6-h group, 12 females, mean age = 20.54, SD = 1.74; and 26 for the 24-h group, 12 females, mean age = 20.96, SD = 2.990). The interaction effect of gender × group on participants' age was not statistically significant ($F$ = 0.213, $P$ = 0.809, $^2$ = 0.006, two-way ANOVA), and gender ($x^2$ = 0.021, $P$ = 0.990) or age ($F$ = 0.661, $P$ = 0.519, $^2$ = 0.017) was not significantly different across three groups.

In study 3, we recruited a total of 90 participants (47 females; mean age = 21.01, SD = 2.44). Fifteen participants (8 females; mean age = 21.2, SD = 1.22) were excluded from further analysis since they failed to show SCR response to any stimulus (CS1+, CS2+ or CS-) (non-responders, n = 3,4,4,4; R rdlPFC group, R vertex group, NR rdlPFC group and NR vertex group, respectively, see Supplemental Table S1) (de Voogd and Phelps, 2020; Dunsmoor et al., 2015; Dunsmoor et al., 2017; Hu et al., 2018; Raio et al., 2017; Schiller et al., 2013; Schiller et al., 2010; Schiller et al., 2012). Our final sample included 75 participants for data analysis (19 for the Reminder right dlPFC group, 10 females, mean age = 22.05, SD = 2.44; 18 for the Reminder vertex group, 9 females, mean age = 20.67, SD = 2.68; 18 for the No-reminder right dlPFC group, 10 females, mean age = 20.28, SD = 2.85; and 20 for the No-reminder vertex group, 10 females, mean age = 20.85, SD = 2.43).

## Conditioned stimuli and psychophysiological stimulation

For study 1, we employed two squares with different colors (yellow and blue) that served as CS+ and CS-. During fear acquisition, the CS+ was paired with a mild electrical shock (US) on a 37.5% partial reinforcement scheme (6 CS+ paired with electric shock and 10 CS+ with no shock), and the CS-was never paired with a shock (10 CS-). A pseudo-random CS delivery order was generated for the fear acquisition, extinction and test phase for two groups with the rule that no same trial-type (CS+ and CS-) repeated more than twice. Assignment of square colors to the CS (CS+ and CS-) was counterbalanced across participants.

For study 2 & 3, three squares with different colors (yellow, red and blue) were used as CS1+, CS2+ and CS-, respectively. During the fear acquisition phase, both CS1+ and CS2+ were paired with electric shocks (US) on a 37.5% partial reinforcement scheme, and the CS-was not paired with shocks (10 non-reinforced presentations of CS1+, CS2+ and CS-each, intermixed with an extra of 6 CS1+ and 6 CS2+ that co-terminated with the shock). Again, a pseudo-random stimulus order was generated for fear acquisition and extinction phases of three groups with the rule that no same trial-type (CS1+, CS2+ and CS-) repeated more than twice. In the test phase, to exclude the possibility that the difference between CS1+ and CS2+ was simply caused by the presentation sequence of CS1+ and CS2+, half of the participants completed the test phase using a pseudo-random stimuli sequence and the identities of CS1+ and CS2+ reversed in the other half of the participants. For all three groups, the CS colors were counterbalanced across participants.

The US was a mild electrical shock delivered to the right wrist of participants via a DS-5 Isolated Bipolar Constant Current Stimulator (Digitimer, Welwyn Garden City, UK). The US levels were set by the participants themselves, starting with a low shock level (5v) and gradually increased and settled to a level that they described as 'uncomfortable, but not painful' (with the maximum level of 10v). The duration of all electric shocks was 200ms with 50 pulses per second.

SCR measurements were collected using two Ag-AgCI electrodes attached to the tips of the index and middle fingers of each subject's left hand. All skin conductance data were recorded via the Biopac® MP160 BioNomadix System and analyzed using the Acknowledgement 5.0 software. The SCR level for each trial was calculated as the amplitude difference (in micro siemens) from peak to trough during the 0.5 to 4.5s window after the CS (CS SCR) or US (US SCR) onset and responses below 0.02μS were encoded as zero. The raw SCR scores were divided by each subject's mean US responses and then square-root-transformed to normalize SCR across individuals and groups (Wang et al., 2021).

Continuous theta burst stimulation (cTBS), a specific form of repetitive transcranial magnetic stimulation (rTMS), was applied with a Magstim Rapid 2 system (Magstim, Whitland, Wales, UK). Before the experiments, we determined the stimulation intensity of the cTBS protocol that was at 80% of the individual active motor threshold (AMT). Each transcranial magnetic stimulation (TBS) burst consisted of three stimuli pulses at 50HZ, with each train being repeated every 200ms (5Hz). In this cTBS protocol, a 40s train of uninterrupted TBS is given (600 pluses) (Borgomaneri et al., 2020a; Su et al., 2022). During the stimulation, the cTBS was applied either over the right dlPFC or the vertex, which was determined by standard F4 and Cz locations based on international electroencephalogram 10-20 system measurements (Censor et al., 2010).

The control ability over intrusive thoughts was measured by the 25-item Thought-Control Ability Questionnaire (TCAQ) scale (Wells and Davies, 1994). Participants were asked to rate on a five-point Likert-type scale the extent to which they agreed with the statement from 1 (*completely disagree*) to 5 (*completely agree*). At the end of the experiments, all participants completed the TCAQ scale to assess their perceived control abilities over intrusive thoughts in daily life (Küpper et al., 2014). The internal consistency (Cronbach's alpha) of TCAQ scale was 0.92, and the reliability coefficient was satisfactory ($r$ = 0.88).

To measure the extent to which fear returns after the presentation of unconditioned stimuli (US, electric shock) in the test phase, we defined the fear recovery index as the SCR difference between the first test trial and the last extinction trial for a specific CS for each subject. Similarly, in studies 2 and 3, differential fear recovery index was defined as the difference between fear recovery indices of CS+ and CS-for both CS1+ and CS2+.

## Experimental procedure

Study 1, 2 and 3 were carried out for two or three consecutive days with three phases. Before the experiments, all participants gave informed consent. During the experiments, subjects were required to stay relaxed and still, focus on the computer screen and pay attention to the relationship between color stimuli (CS) and the electric shock (US).

Day 1: Acquisition. For study 1, two groups of participants underwent a fear-conditioning paradigm (acquisition) where the CS+ was partially (37.5%) paired with the US and the CS-was never paired with the shock. For study 2 & 3, participants underwent fear conditioning using three color squares as the CSs (CS1+, CS2+ and CS-). The CSs were presented for 4s each with a 10-12s variable ITI in all the studies (1, 2 and 3).

Day 2: Retrieval-extinction phase. For study 1, the fear memory was reactivated before extinction in the reminder group. During reactivation, the CS+ was presented once without the delivery of the US. Followed by a 10-min break, the reminder group underwent extinction training with 10 CS+ and 11 CS-presentation non-reinforced. In the no-reminder group, the fear memory was not reactivated (no CS+ reminder). To match with the timeline of the reminder group, the no-reminder group subjects still needed to wait for 10-min after the beginning of the experiment on day 2 and then underwent extinction training with 11 CS+ and 11 CS+ non-reinforced presentation. During the 10-min break, all participants were asked to rest and stay still.

For study 2, only one of the CS+ (CS1+) was presented once without the US during reactivation in all three groups. After a 10-min break, all three groups underwent extinction training with 10 CS1+, 11CS2+ and 11CS-non-reinforced presentation.

For study 3, for the reminder groups (reminder-dlPFC and reminder-vertex groups), only one of the CS+ (CS1+) was presented once (without the US delivery) as the retrieval trial. No CS+ reminder was delivered for the no-reminder groups (no-reminder dlPFC and no-reminder vertex groups). Additionally, during the 10-min break (8 mins after the supposed CS+ reminder delivery), cTBS was applied either over the right dlPFC (dlPFC groups: PFC) or the vertex (vertex groups: VER), resulting in a 2 (reminder vs. no-reminder) x 2 (dlPFC vs. vertex) of 4 different groups (reminder-dlPFC: R-PFC, reminder-vertex: R-VER, no-reminder dlPFC: NR-PFC, no-reminder vertex: NR-VER).

Day 2 or Day 3: Reinstatement and test phases. For study 1, 30 minutes after the extinction training, both groups received four un-signaled electric shocks with 10s to 12s ITI (reinstatement). Eighteen seconds after the reinstatement phase, all subjects were presented with CS+ and CS-10 times each without electric shocks and their SCRs were recorded (test phase).

For study 2, participants were assigned into three groups: the 30min group, the 6h group and the 24h group. All participants received four un-signaled electric shocks with 10s to 12s ITI (reinstatement) after their corresponding time delays (30min, 6h or 24h) between extinction and reinstatement. Eighteen seconds after the reinstatement phase, all subjects were tested via the presentation of CS1+, CS2+ and CS-10 times each without electric shocks associated.

For study 3, the reinstatement and testing procedures were the same as in study 2 except that they were conducted one hour after the extinction training.

# Results

## The short-term effect of the fear memory retrieval-extinction procedure

To test the memory-retrieval-induced fear memory deficits at different time scales in humans, we designed two experiments to examine whether the fear memory retrieval-extinction procedure would block the return of fear memory.

In the first study, we aimed to test whether there is a short-term amnesia effect following the fear retrieval-extinction paradigm. We measured subjects' SCRs for the fear acquisition, extinction and test phases to assess the associative fear memory and the recovery of fear memory. At each phase, the differential fear response was calculated as the difference between SCRs to the CS+ and CS-.

To assess fear acquisition across groups (Figure 1B and C), we conducted a mixed two-way ANOVA of group (reminder vs. no-reminder) x time (early vs. late part of the acquisition; first 5 and last 5 trials, correspondingly) on the differential fear SCR. Our results showed a significant main effect of time (early vs. late; $F_{1,55} = 6.545$, $P = 0.013$, $\eta^2 = 0.106$), suggesting successful fear acquisition in both groups. There was no main effect of group (reminder vs. no-reminder) or the

group x time interaction (group: $F = 0.057$, $P = 0.813$, $^2 = 0.001$; interaction: $F = 0.066$, $P = 0.798$, $\eta = 0.001$), indicating similar levels of fear acquisition between two groups. Post-hoc *t*-tests confirmed that the fear responses to the CS+ were significantly higher than that of CS-during the late part of acquisition phase in both groups (reminder group: $t_{29} = 6.642$, $P < 0.001$; no-reminder group: $t_{26} = 8.522$, $P < 0.001$; Figure 1C). Importantly, the levels of acquisition were equivalent in both groups (early acquisition: $t_{55} = -0.063$, $P = 0.950$; late acquisition: $t_{55} = -0.318$, $P = 0.751$; Figure 1C).

To ensure equivalent and successful extinction across groups (Figure 1B and D), we performed similar mixed two-way ANOVA (group x trial) on the differential SCR in the extinction phase. Our results showed significant main effect of trial ($F_{10,550} = 2.825$, $P = 0.010$, $\eta = 0.049$), but no effect of group ($F = 0.063$, $P = 0.802$, $^2 = 0.001$) or their interaction ($F = 1.77$, $P = 0.101$, $\eta = 0.031$). Follow-up *t*-tests further confirmed that in the last trial of extinction there was no difference between fear responses to CS+ and CS-within group (reminder group: $t_{29} = -1.147$, $P = 0.261$; no-reminder group: $t_{26} = 0.261$, $P = 0.796$; Figure 1D). Importantly, the levels of extinction did not show significant difference between two groups ($t_{55} = -1.074$, $P = 0.288$; Figure 1D).

Fear recovery was assessed using a two-way ANOVA with main effects of group (reminder vs. no-reminder) and time (last trial of extinction vs. first trial of test) and it showed a significant time × group interaction ($F_{1,55} = 4.087$, $P = 0.048$, $\eta = 0.069$). We defined the fear recovery index as the SCR difference between the first test trial and the last extinction trial for a specific CS. The interaction effect was confirmed by the significant difference between the differential fear recovery indexes between CS1+ and CS2+ in the reminder and no-reminder groups ($t_{55} = -2.022$, $P = 0.048$; Figure 1E, also see Supplemental Material for the direct test of the test phase). Post hoc *t*-tests showed that fear memories were resilient after regular extinction training, as demonstrated by the significant difference between fear recovery indexes of the CS+ and CS- for the no-reminder group ($t_{26} = 7.441$, $P < 0.001$; Bonferroni correction, Figure 1E), while subjects in the reminder group showed no difference of fear recovery between CS+ and CS- ($t_{29} = 0.797$, $P = 0.432$, Figure 1E). Also, our results remain robust even with the “non-learners” included in the analysis (Figure supplement 1 in the Supplemental Material).

Together, these results indicate that the retrieval-extinction procedure prevents the return of fear expression in the short-term test, a phenomenon distinct from the hypothesized fear reconsolidation effect at a much later time (Schiller et al., 2013). In the current study, fear recovery was tested 30 minutes after extinction training, whereas the effect of memory reconsolidation was generally evident only several hours later and possibly with the help of sleep (Kindt and Soeter, 2018; Speer et al., 2021), leaving open the possibility of a different cognitive mechanism for the short-term fear amnesia related to the retrieval-extinction procedure. Importantly, such a short-term effect is also retrieval dependent, suggesting the labile state of memory may be necessary for the short-term memory update to take effect (Figure 1E).

## Temporal and generalizing characteristics of the retrieval-extinction effect

Given the findings of short-term fear amnesia following the retrieval-extinction paradigm, we set out to map the temporal dynamics of fear amnesia as well as its cue-specificities - whether the amnesia triggered by a specific CS+ reminder (for example CS1+) generalizes to the fear memory associated with the other CS+ (CS2+). We adopted a double-cue fear learning paradigm and examined fear recovery in the short- (30 mins), medium- (6 hours) and long-term (24 hours) fear memory tests in three separate groups of participants after the extinction training (Figure 2A).

Fear acquisition was assessed using a mixed three-way ANOVA with group (30min, 6h and 24h), CS+ (CS1+ vs. CS2+, defined as the mean SCR differences between each CS+ and CS-) and trial numbers (10 trials) as main factors. This analysis showed a significant main effect of trial ($F_{9,684} = 12.782$, $P < 0.001$, $\eta^2 = 0.144$), but no effect of group ($F_{2,76} = 2.121$, $P = 0.127$, $\eta^2 = 0.053$), CS+ ($F = 0.576$, $P = 0.45$, $^2 = 0.008$) or their interactions (all *Ps* > 0.05; Figure 2B). Furthermore, to ensure equivalent fear acquisition across three groups, a mixed two-way ANOVA showed no effect of group (30min, 6h and 24h; $F_{2,76} = 1.79$, $P = 0.174$, $\eta = 0.045$), CS+ (CS1+ vs. CS2+; $F_{1,76} = 1.337$, $P =$

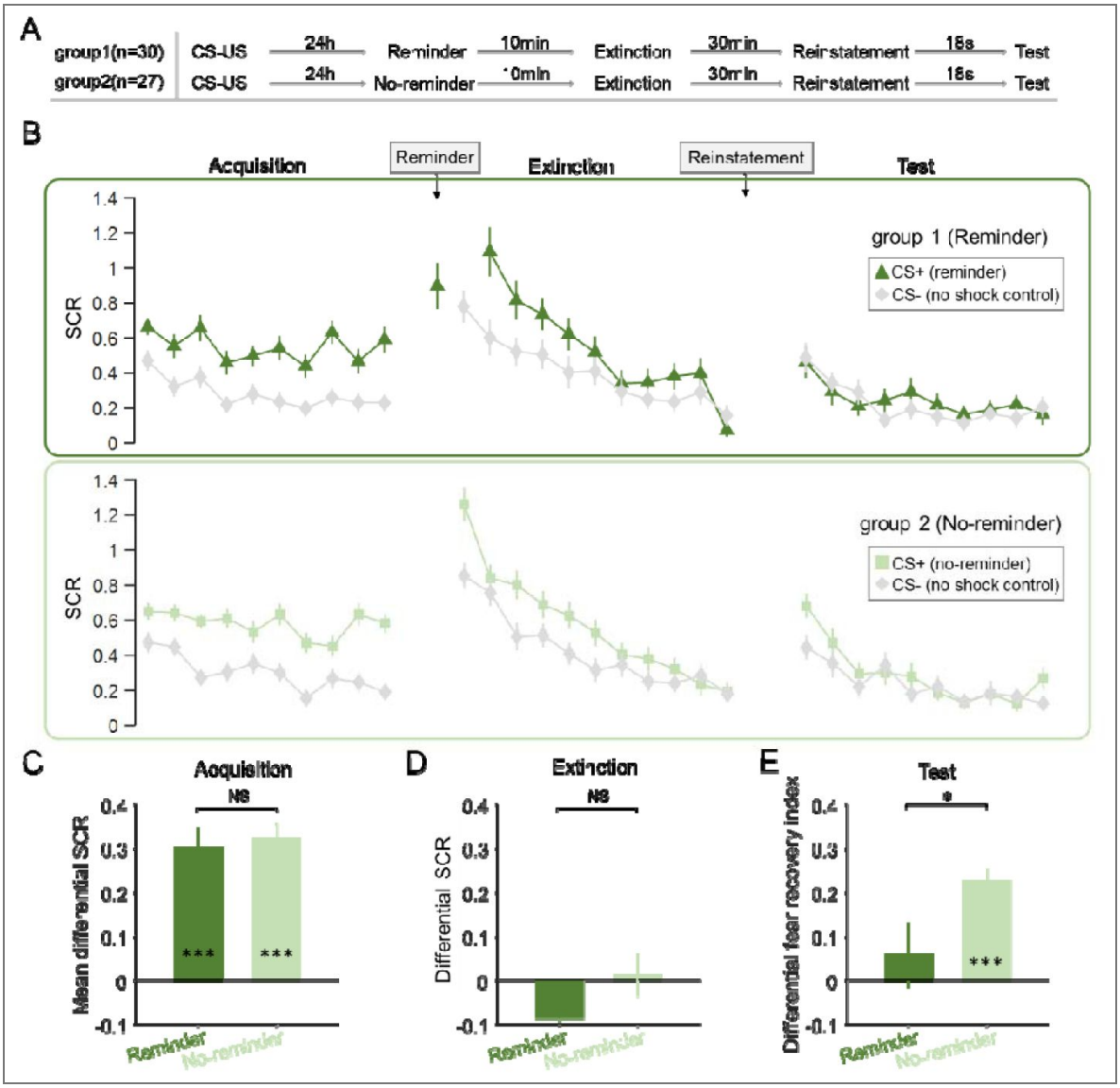



**Figure 1. SCR responses to conditioned stimuli in two groups of study 1 indicating short-term memory amnesia.**

(**A**) Experimental design and timeline. (**B**) SCRs of fear conditioned stimuli (CS+) and the control stimulus (CS-) across fear acquisition, extinction and test phases for two groups (reminder and no-reminder). (**C**) Mean differential SCRs (CS+ minus CS-) in the acquisition phase (trials in the latter half). (**D**) Differential SCRs (CS+ minus CS-) in the extinction phase (last trial). **(E)** Differential fear recovery index between CS+ and CS-in the test phase. ***$P < 0.001$, *$P < 0.05$. NS: Non-significant. Error bars represent standard errors.

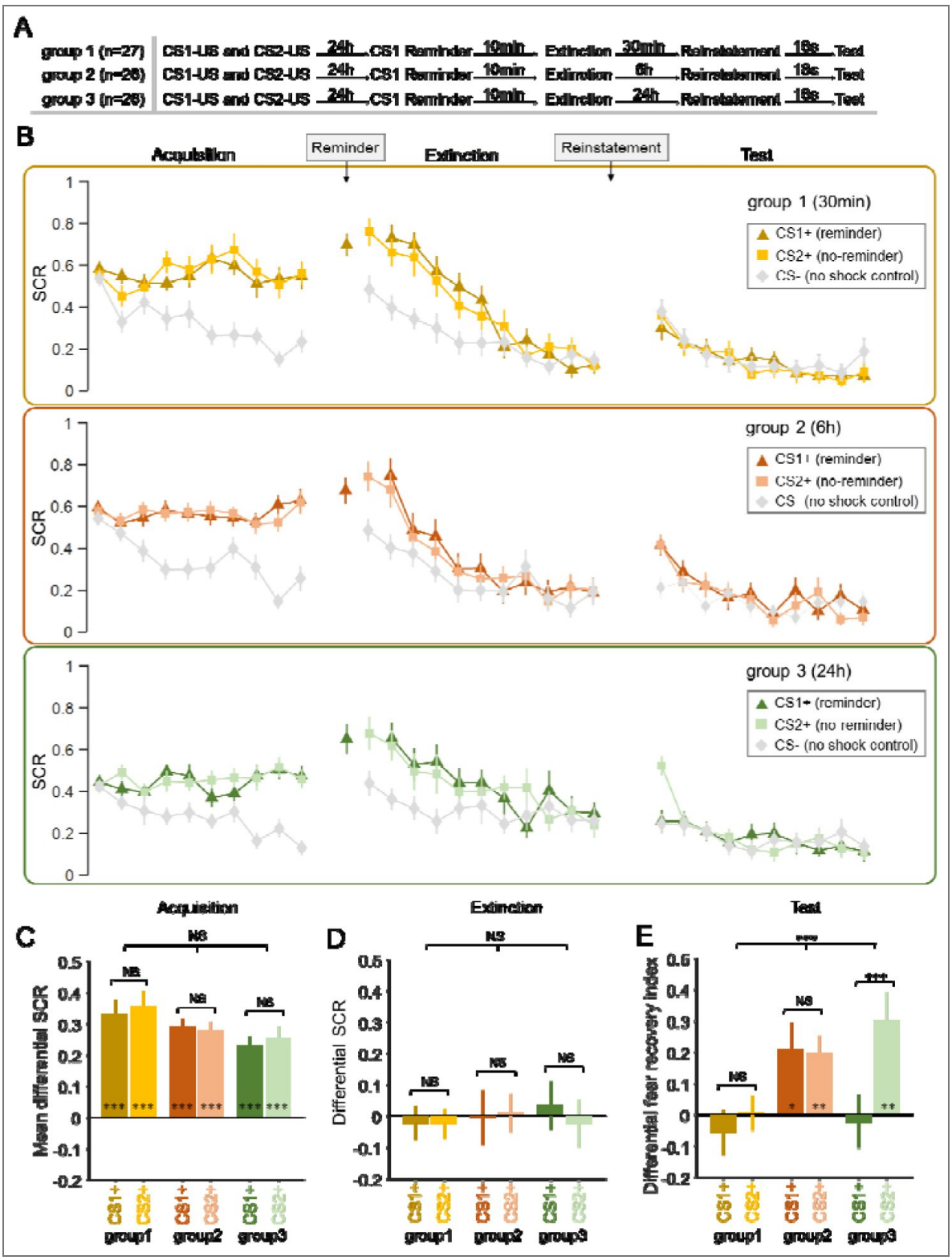



**Figure 2. SCR responses to conditioned stimuli in three groups of study 2 indicating amnesia at different timescales.**

(**A**) Experimental design and timeline. (**B**) SCRs of fear conditioned stimuli CS1+ (reminder) and CS2+ (No-reminder), and the control stimulus (CS-) across the fear acquisition, extinction and test phases for each group (30min, 6h and 24h groups). (**C**) Mean differential SCRs (CS1+ minus CS- and CS2+ minus CS-) in the acquisition phase (trials in the latter half). (**D**) Differential SCRs in the extinction phase (last trial). (**E**) Differential fear recovery index between CS+ and CS- in the test phase. ***$P$ < 0.001. **$P$ < 0.01. *$P$ < 0.05. NS: Non-significant. Error bars represent standard errors.

0.251, η = 0.017) or their interaction ($F_{2,76}$ = 0.959, $P$ = 0.388, η = 0.025) on the late phase of acquisition (last 5 trials). Post-hoc $t$-tests confirmed that the SCRs elicited by CS1+ and CS2+ were significantly higher than that of CS-among three groups on the late phase of acquisition (CS1+: $t_{26}$ = 6.806, $P$ < 0.001; CS2+: $t_{26}$ = 6.854, $P$ < 0.001 for the 30min group; CS1+: $t_{25}$ = 9.822, $P$ < 0.001; CS2+: $t_{25}$ = 9.841, $P$ < 0.001 for the 6h group and CS1+: $t_{25}$ = 6.909, $P$ < 0.001; CS2+: $t_{25}$ = 6.956, $P$ < 0.001 for the 24h group), indicating successful and similar fear acquisition across all groups (Figure 2C).

We then set out to examine whether participants underwent successful extinction training on day 2. The similar mixed three-way ANOVA showed a significant effect of trial ($F_{10,760}$ = 8.220, $P$ < 0.001, η = 0.098), but no effect of group ($F_{2,76}$ = 1.822, $P$ = 0.169, η = 0.046), CS+ ($F_{1,76}$ = 0.91, $P$ = 0.343, $\eta^2$ = 0.012) or their interactions (all $P$s > 0.1; Figure 2B). To test whether participants across all groups achieved similar levels of extinction, a mixed two-way ANOVA showed no effect of group (30min, 6h and 24h; $F_{2,76}$ = 0.059, $P$ = 0.943, η = 0.002), CS+ (CS1+ vs. CS2+; $F_{1,76}$ = 0.258, $P$ = 0.613, η = 0.003) or their interaction ($F_{2,76}$ = 0.504, $P$ = 0.606, η = 0.013) on the last trial of extinction training. Post-hoc $t$-tests showed that there was no difference between CS+ and CS-responses in all groups on the last trial of extinction (CS1+: $t_{26}$ = -0.397, $P$ = 0.695; CS2+: $t_{26}$ = -0.505, $P$ = 0.618 for the 30min group; CS1+: $t_{25}$ = -0.054, $P$ = 0.957; CS2+: $t_{25}$ = 0.140, $P$ = 0.890 for the 6h group and CS1+: $t_{25}$ = 0.435, $P$ = 0.668; CS2+: $t_{25}$ = -0.317, $P$ = 0.754 for the 24h group; Figure 2D).

To assess the effects of the retrieval-extinction procedure as a function of the time delay between fear extinction and test phases, we conducted a 3-way ANOVA with the within-subject factor CS+ (CS1+ vs. CS2+, defined by the mean SCR differences between each CS+ and CS-), time (last trial of extinction vs. first trial of test) and between-subject factor group (30min, 6h and 24h) and found a significant three-way interaction ($F_{2,76}$ = 6.376, $P$ = 0.003, $\eta^2$ = 0.144). To disentangle this interaction, we further examined the effects in each group.

In the 30min group, there was no significant effect of time ($F$ = 0.208, $P$ = 0.652, $^2$ = 0.008), CS+ ($F_{1,26}$ = 0.888, $P$ = 0.355, η = 0.033) or their interaction ($F_{1,26}$ = 1.039, $P$ = 0.318, η = 0.038). Post-hoc $t$-tests showed that the retrieval-extinction training diminished fear responses to both the reminded CS+ (differential fear recovery index between CS1+ and CS-, $t_{26}$ = -0.787, $P$ = 0.438; Figure 2C) and the non-reminded CS+ (differential fear recovery index between CS2+ and CS-, $t_{26}$ = 0.088, $P$ = 0.93; Figure 2E), suggesting a cue-independent short-term amnesia effect. This result indicates that the short-term amnesia effect observed in Study 2 is not reminder-cue specific and can generalize to the non-reminded cues. In addition, there was no significant difference between the fear recovery index associated with CS1+ and CS2+ ($t_{26}$ = -1.019, $P$ = 0.318; Figure 2E, also see the supplemental material for direct test of the test phase SCRs).

In contrast to the short-term effect, there was a significant effect of time ($F_{1,25}$ = 9.654, $P$ = 0.005, η = 0.279), but no effect of CS+ ($F_{1,25}$ = 0.037, $P$ = 0.85, η = 0.001) or their interaction ($F_{1,25}$ = 0.028, $P$ = 0.869, $^2$ = 0.001) in the 6h (medium-term) group. Post-hoc $t$-tests showed that fear memory recovered for the reminded CS+ (differential fear recovery index between CS1+ and CS-, $t_{25}$ = 2.382, $P$ = 0.025), as well as the non-reminded CS+ (differential fear recovery index between CS2+ and CS-, $t_{25}$ = 3.525, $P$ = 0.002) in the medium-term test of the retrieval-extinction procedure effect, suggesting the failure to suppress the return of extinguished fear memory at such a time scale (Figure 2E). There was also no significant difference measured by the fear recovery index between CS1+ and CS2+ ($t_{25}$ = 0.166, $P$ = 0.870; Figure 2E).

Finally, in line with previous research on fear memory reconsolidation, significant SCR difference of the fear reinstatement effect between the reminded CS+ and non-reminded CS+ (time × CS+ interaction: $F_{1,25}$ = 16.924, $P$ < 0.001, $\eta^2$ = 0.404) was detected in the 24h group via two-way ANOVA. More specifically, the retrieval-extinction procedure only diminished fear responses to the reminded CS+ (differential fear recovery index between CS1+ and CS-, $t_{25}$ = -0.25, $P$ = 0.805), whereas the fear response to the non-reminded CS+ remained significant (differential fear recovery index between CS2+ and CS-, $t_{25}$ = 3.269, $P$ = 0.003; Figure 2E). Post-hoc t-test showed that the fear recovery index between CS1+ and CS2+ was significantly different ($t_{25}$ = -4.114, $P$ < 0.001; Figure 2E), aligning well with previous literature that suggested the cue-specificity of the

fear reconsolidation effect (Lee et al., 2017; Monfils et al., 2009). Similarly, these results remain robust even with the "non-learners" included in the analysis (Figure supplement 2 in the Supplemental Material).

Altogether, in line with our hypothesis, the return of fear memory measured by the reinstatement effect showed distinct patterns at different time scales. The retrieval-extinction procedure diminished fear SCRs for both the reminded and non-reminded CS+, engendering cue-independent amnesia in the short-term test (30min). On the other hand, at the long-term (24h) fear memory test, fear responses were only diminished for the reminded CS+ (but not the non-reminded CS+), consistent with previous literature on memory reconsolidation. Interestingly, SCR measures of the fear memory re-appeared in the medium-term (6h) test and there was no difference between fear responses elicited by the reminded and non-reminded CS+, suggesting that the 6-hour timescale may be at a "limbo" stage where both the immediate and the delayed (reconsolidation) memory update mechanisms are not at work.

## Fear amnesia as a function of the thought-control ability

Finally, to assess whether the fear reinstatement effects at different time scale were related to subjects' thought-control ability, we first ran a one-way ANOVA to confirm that subjects among different groups did not differ in their thought-control abilities measured by the TCAQ scores ($F$ = 0.154, $P$ = 0.857, $^2$ = 0.004). We then ran separate linear regressions between the differential fear recovery index (differential fear recovery index between CS+ and CS-) and thought-control abilities (TCAQ scores) in the three groups. This analysis yielded significant negative correlation between TCAQ score and fear recovery index for both CS+ (CS1+ and CS2+) in the 30min group (Figure 3A, main effect of thought-control ability, $t_{51}$ = -3.446, $P$ = 0.003, Bonferroni correction), indicating that subjects endowed with higher thought-control abilities were less likely to experience the return of fear response. However, no significant correlation was observed in the 6h or the 24h group (Figure 3B and C; $P$s > 0.4). Post-hoc analysis of the 30min group showed significant correlation between the fear recovery index of both CS1+ and CS2+ with subjects' thought-control abilities (Figure 3A, CS1+: r = -0.461, $P$ = 0.015; CS2+: r = -0.413, $P$ = 0.032). No significant TCAQ and CS+ interactions were found in the three groups ($P$s > 0.4).

It is worth noting that correlations between TCAQ scores and fear responses in all three groups were not significant in the late phase of acquisition or the last trial of extinction (all $P$s > 0.10), indicating that the effect of thought-control ability was specifically reflected in the fear recovery index.

In conclusion, thought-control abilities were associated with fear recovery indices only in the short-term (30min) test for both the reminded and non-reminded CS+ but not at longer time scales (6h and 24h). The 30min group results are consistent with previous research which suggested that people with better capability to resist intrusive thoughts also performed better in motivated forgetting in both declarative and associative memories (Küpper et al., 2014; Wang et al., 2021).

## dlPFC dependent short-term fear memory amnesia

In study 3 (Figure 4A), we examined whether intact activity of the right dlPFC (rdlPFC) was also necessary for short-term fear amnesia, in addition to the memory retrieval cue (reminder). To test this hypothesis, we set up a reminder CS+ (CS1+ vs. CS2+) x reminder (reminder vs. no-reminder) x cTBS (rdlPFC vs. vertex) x trial number four-way ANOVA against the differential SCRs (CS+ minus CS-) for the acquisition and extinction phases. In the acquisition phase, as expected, there were no significant main effects of reminder ($F$ = 1.291, $P$ = 0.260, $^2$ = 0.018), cTBS ($F$ < 0.001, $P$ = 0.990, η < 0.001), CS+ ($F_{1,71}$ = 1.927, $P$ = 0.169, η = 0.026) or their interactions (all $P$s > 0.1; Figure 4B and C) but a significant main effect of trial number ($F_{9,639}$ = 8.054, $P$ < 0.001, η = 0.102), indicating successful learning of CS+ stimuli. Indeed, post-hoc tests confirmed that the SCRs of CS+ was significantly higher than that of the CS- on the late phase (last 5 trials) of acquisition in all four groups (CS1+: $t_{18}$ = 7.735, $P$ < 0.001; CS2+: $t_{18}$ = 8.546, $P$ < 0.001 for the R-PFC group; CS1+: $t_{17}$ = 8.568, $P$ < 0.001; CS2+: $t_{17}$ = 8.478, $P$ < 0.001 for the R-VER group; CS1+: $t_{17}$ = 6.032, $P$ < 0.001; CS2+:

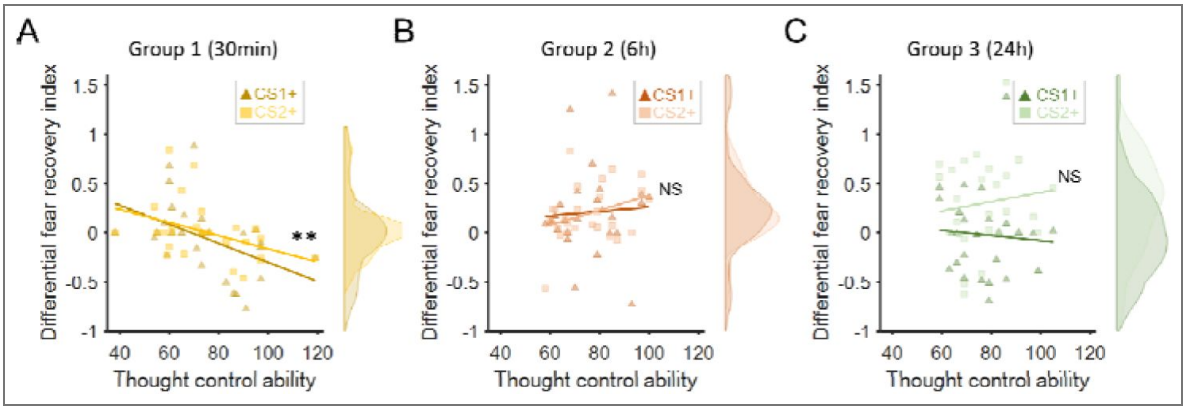



**Figure 3. Fear recovery as a function of thought-control abilities.**

(**A**)Thought control ability was significantly correlated with fear recovery index in the 30min group ($P$ = 0.003, Bonferroni correction), but not in the 6h (**B**) or 24h group (**C**, $P$s > 0.7). The violin graphs indicate the distribution of fear recovery index across subjects (Figure 2E). $^{**}P$ < 0.01. NS: Non-significant.

$t_{17}$ = 7.722, $P$ < 0.001 for the NR-PFC group and CS1+: $t_{19}$ = 9.110, $P$ < 0.001; CS2+: $t_{19}$ = 6.538, $P$ < 0.001 for the NR-VER group, Figure 4C) . A similar mixed-effect four-way ANOVA was also performed on the SCRs in the extinction phase, which showed significant effects of trial ($F_{10,710}$ = 8.567, $P$ < 0.001, η = 0.108), cTBS ($F_{1,71}$ = 6.269, $P$ = 0.015, $\eta^2$ = 0.081) and cTBS × trial interaction ($F_{10,710}$ = 2.292, $P$ = 0.030, $\eta^2$ = 0.031), but no effect of reminder ($F$ = 1.894, $P$ = 0.173, $^2$ = 0.026), CS+ ($F$ = 0.398, $P$ = 0.530, $^2$ = 0.006) or other interactions (all $Ps$ > 0.1; Figure 4B). To test whether participants across all groups achieved similar levels of extinction, a mixed-effect three-way ANOVA was performed on the last trial of extinction training and there was no significant effect of reminder (reminder and no-reminder; $F_{1,71}$ = 3.112, $P$ = 0.082, η = 0.042), cTBS ($F_{1,71}$ = 0.929, $P$ = 0.338, η = 0.013), CS+ (CS1+ vs. CS2+; $F_{1,71}$ = 1.623, $P$ = 0.207, η = 0.022) or their interactions (all $Ps$ > 0.1; Figure 4D). Furthermore, post-hoc $t$-tests showed that there was no difference between CS+ and CS-responses in all groups on the last trial of extinction (CS1+: $t_{18}$ = 1.310, $P$ = 0.207; CS2+: $t_{18}$ = 1.298, $P$ = 0.211 for the R-PFC group; CS1+: $t_{17}$ = 0.839, $P$ = 0.413; CS2+: $t_{17}$ = 1.090, $P$ = 0.291 for the R-VER group; CS1+: $t_{17}$ = 0.380, $P$ = 0.709; CS2+: $t_{17}$ = 0.624, $P$ = 0.541 for the NR-PFC group and CS1+: $t_{19}$ = -1.434, $P$ = 0.168; CS2+: $t_{19}$ = -0.535, $P$ = 0.599 for the NR-VER group; Figure 4D).

To assess the effects of the right dlPFC on the memory retrieval related short-term amnesia, we conducted a mixed-effect 3-way ANOVA with the within-subject factors CS+ (CS1+ vs. CS2+), between-subject factors reminder (reminder vs. no-reminder) and cTBS (dlPFC vs. vertex) on fear recovery index (the difference between the first test trial and the last extinction trial for CS1+ and CS2+), which revealed a significant effect of reminder ($F_{1,71}$ = 8.920, $P$ = 0.004, $\eta^2$ = 0.112), cTBS ($F_{1,71}$ = 7.180, $P$ = 0.009, $\eta^2$ = 0.092) and reminder × cTBS interaction ($F_{1,71}$ = 10.613, $P$ = 0.002, $\eta^2$ = 0.130), but no effect of CS+ ($F_{1,71}$ = 0.518, $P$ = 0.474, $\eta^2$ = 0.007) or other interactions (all $Ps$ > 0.1; Figure 4C). Post-hoc $t$-tests showed that there were significant fear reinstatement responses in the R-PFC group (CS1+: $t_{18}$ = 5.182, $P$ < 0.001; CS2+: $t_{18}$ = 5.258, $P$ < 0.001), the NR-PFC group (CS1+: $t_{17}$ = 3.496, $P$ = 0.003; CS2+: $t_{17}$ = 4.628, $P$ < 0.001) and the NR-VER group (CS1+: $t_{19}$ = 4.885, $P$ < 0.001; CS2+: $t_{19}$ = 3.262, $P$ = 0.004), but not in the R-VER group (CS1+: $t_{17}$ = -1.640, $P$ = 0.119; CS2+: $t_{17}$ = -0.744, $P$ = 0.467) (Figure 4E), suggesting that both memory retrieval and intact dlPFC function are necessary for the short-term fear amnesia.

Finally, we went on to examine whether the fear reinstatement response was related to the thought-control ability. A two-way ANOVA of reminder (reminder vs. no-reminder) x cTBS (rdlPFC vs. vertex) confirmed that there were no significant effects of reminder ($F_{1,71}$ = 0.366, $P$ = 0.547, $\eta^2$ = 0.005), cTBS ($F_{1,71}$ = 0.606, $P$ = 0.439, $\eta^2$ = 0.008) or their interaction ($F_{1,71}$ = 1.216, $P$ = 0.274, η = 0.017) on the TCAQ scores. We then conducted simple linear regressions on different groups (as in study 2) and found a significant correlation between the thought-control ability and differential fear recovery index in the R-VER group (Figure 5B, $t_{33}$ = -3.283, $P$ = 0.008, Bonferroni correction across 4 groups), consistent with the results of the short term (30min) amnesia effect in study 2 (Figure 3A). Further post-hoc analysis of the individual CS+ showed that the correlations between thought-control ability and fear recovery index were significant for both CS+ (CS1+: $r$ = -0.557, $P$ = 0.016; CS2+: $r$ = -0.483, $P$ = 0.042). However, there was no such correlation in the R-PFC group (Figure 5A, $t_{35}$ = 0.719, $P$ = 0.477), the NR-PFC group (Figure 5C, $t_{33}$ = -1.334, $P$ = 0.192) or the NR-VER group (Figure 5D, $t_{37}$ = -0.025, $P$ = 0.980) and the interaction effects of thought-control ability and CS+ (CS1+ and CS2+) were not significant in all 4 groups ($P$s > 0.3). Importantly, there was no significant correlation between thought-control ability and SCR in the late phase of acquisition or the last trial of extinction in four groups (all $Ps$ > 0.10), suggesting a specific relationship between thought-control ability and short-term fear recovery.

# Discussion

Our results provide direct evidence to support the hypothesis that the memory reactivation might engage distinct cognitive mechanisms for fear memory modulation, which can be separated by their temporal dynamics, cue specificity and the dependence of thought-control ability. In line with previous research on fear memory reconsolidation (Liu et al., 2014), our research replicated the reported results that the retrieval-extinction paradigm blocked fear responses to the reminded CS+ while leaving non-reminded CS+ response intact (cue-specificity) in the long-term

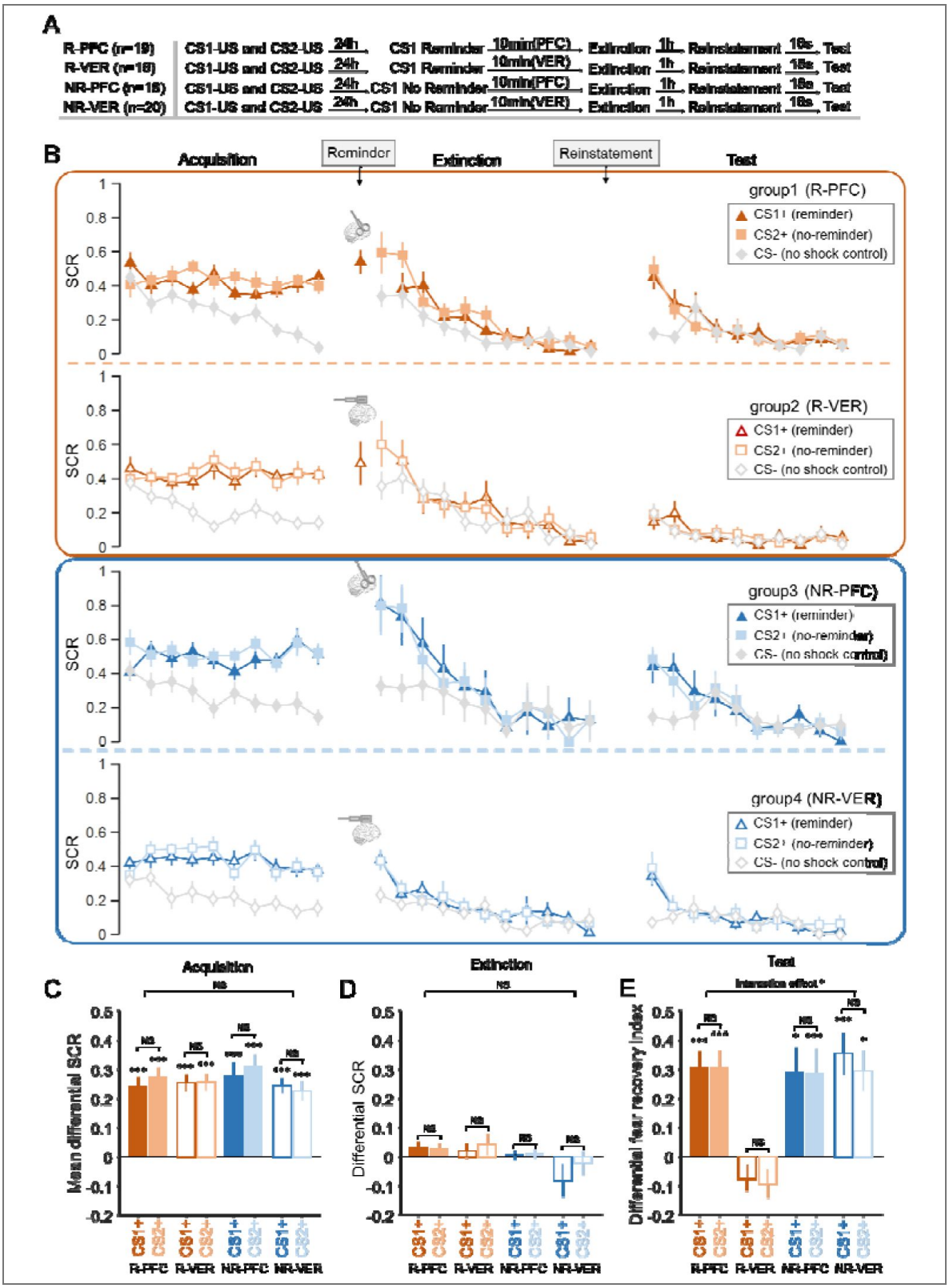



**Figure 4. SCR responses to conditioned stimuli in the cTBS study (study 3) indicating that both intact dlPFC and cue reminder are required for short-term memory amnesia.**

(**A**) Experimental design and timeline. (**B**) SCRs of fear conditioned stimuli CS1+ (reminder) and CS2+ (No-reminder), and the control stimulus (CS-) across the fear acquisition, extinction and test phases for each group (R-PFC, R-VER, NR-PFC and NR-VER groups). (**C**) Mean differential SCRs (CS1+ minus CS- and CS2+ minus CS-) in the acquisition phase (trials in the latter half). (**D**) Differential SCRs in the extinction phase (last trial). **(E)** Differential fear recovery index between CS+ and CS- in the test phase. ***$P < 0.001$. *$P < 0.05$. NS: Non-significant. Error bars represent standard errors.

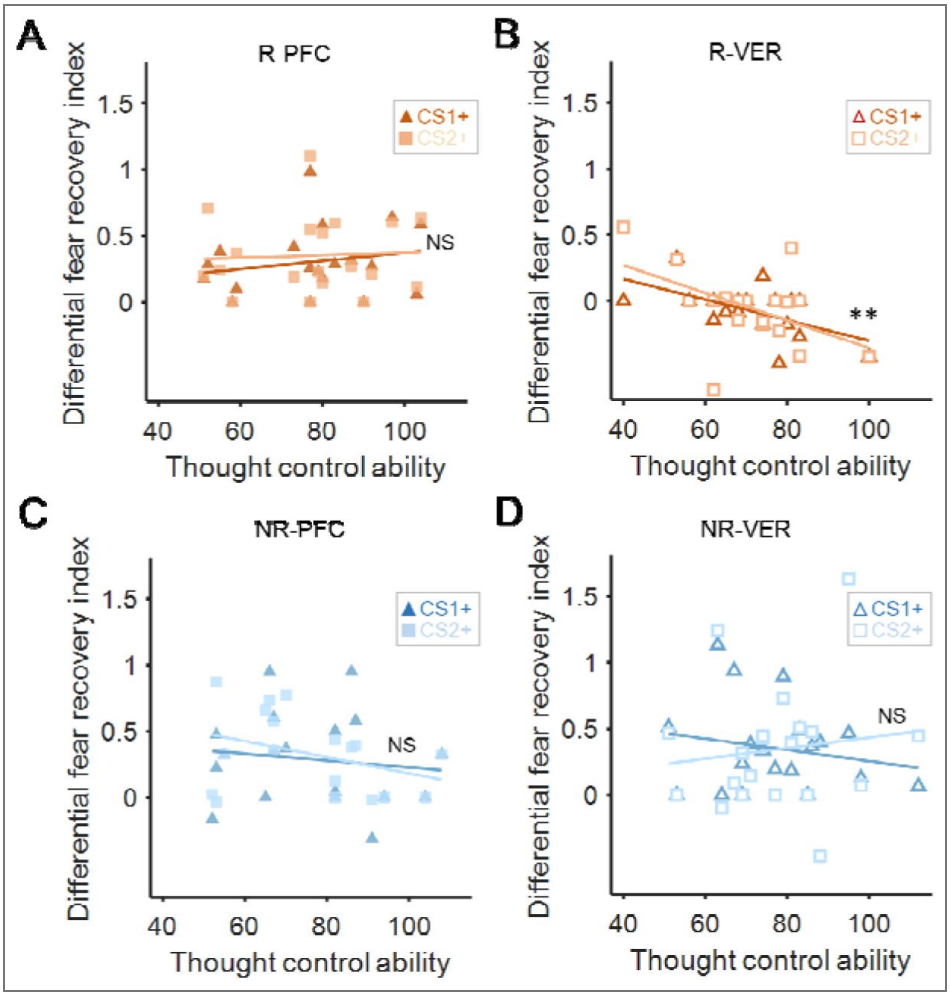



**Figure 5. The correlation between differential fear recovery index and thought-control ability in four cTBS groups.**

(**A**) There was no significant correlation between differential fear recovery index and thought-control abilities in the R-PFC group ($P > 0.4$). (**B**) However, in the R-VER group, the correlation between thought control ability scores and the differential fear recovery index was significant ($P = 0.008$, Bonferroni correction), with high thought-control ability participants showing less fear recovery for both CSs+. Such correlation was not observed in the (**C**) NR-PFC or the (**D**) NR-VER group (*Ps* >0.4). ** $P < 0.01$, NS: Non-significant.

test (Figure 6A and B). However, thought-control ability was not related to the long-term fear amnesia (Figure 6C). Therefore, findings of the short-term fear amnesia suggest that the reconsolidation framework falls short to accommodate this more immediate effect (Figure 6A and B).

Research in both declarative and associative emotional memories suggest that memory retrieval is not a simple act of obtaining a copy of stored information, but provides a critical time window where the old memory is eligible to be integrated with new information (Gershman, Schapiro, et al., 2013; Hupbach et al., 2007; Monfils et al., 2009). Research in the field of declarative memory has shown that the motivated forgetting effect can appear early (10-30 minutes) after the memory suppression practices such as the influential think/no-think (TNT) paradigm (Anderson and Green, 2001), suggesting that memory retrieval may also facilitate short-term memory update (Benoit and Anderson, 2012; Engen and Anderson, 2018). Unlike the slow protein synthesis process involved in memory reconsolidation, the motivated forgetting effect was accompanied by heightened dorsolateral prefrontal cortex (dlPFC) activity, diminished hippocampal activation and altered functional connectivity between these brain regions and hence more adaptive (Anderson et al., 2004; Wimber et al., 2015).

Previous studies indicate that a suppression mechanism can be characterized by three distinct features: first, the memory suppression effect tends to emerge early, usually 10-30 mins after memory suppression practice and can be transient (MacLeod and Macrae, 2001; Saunders and MacLeod, 2002); second, the memory suppression practice seems to directly act upon the unwanted memory itself (Levy and Anderson, 2002), such that the presentation of other cues originally associated with the unwanted memory also fails in memory recall (cue-independence); third, the magnitude of memory suppression effects is associated with individual difference in control abilities over intrusive thoughts (Küpper et al., 2014). The short-term fear amnesia effects we identified are fundamentally different from memory reconsolidation effects and fit closely with the characteristics of the memory suppression effect. Inspired by the similarities between our results and suppression-induced declarative memory amnesia (Gagnepain et al., 2017), we speculate that the retrieval-extinction procedure might facilitate a spontaneous memory suppression process and thus yield a short-term amnesia effect. Accordingly, the activated fear memory induced by the retrieval cue would be subjected to an automatic fear memory suppression through the extinction training (Anderson and Floresco, 2022).

In our experiments, subjects were not explicitly instructed to suppress their fear expression, yet the retrieval-extinction training significantly decreased short-term fear expression. These results are consistent with the short-term amnesia induced with the more explicit suppression intervention (Anderson et al., 1994; Kindt and Soeter, 2018; Speer et al., 2021; Wang et al., 2021; Wells and Davies, 1994). It is worth noting that although consciously repelling unwanted memory is a standard approach in memory suppression paradigm, it is possible that the engagement of the suppression mechanism can be unconscious. For example, in the retrieval-induced forgetting (RIF) paradigm, recall of a stored memory impairs the retention of related target memory and this forgetting effect emerges as early as 20 minutes after the retrieval procedure, suggesting memory suppression or inhibition can occur in a more spontaneous and automatic manner (Imai et al., 2014). Moreover, subjects with trauma histories exhibited more suppression-induced forgetting for both negative and neutral memories than those with little or no trauma (Hulbert and Anderson, 2018). Similarly, people with higher self-reported thought-control capabilities showed more severe cue-independent memory recall deficit, suggesting that suppression mechanism is associated with individual differences in spontaneous control abilities over intrusive thoughts (Küpper et al., 2014). It has also been suggested that similar automatic mechanisms might be involved in organic retrograde amnesia of traumatic childhood memories (Schacter et al., 2012; Schacter et al., 1996).

Therefore, it is theoretically possible that similar brain mechanisms, such as the dynamic brain activity and connectivity modulation, involved in memory suppression may be deployed spontaneously and act as a general cognitive mechanism to cope with intrusive memories in daily life. Indeed, recently researchers have proposed that the standard retrieval-extinction paradigm to

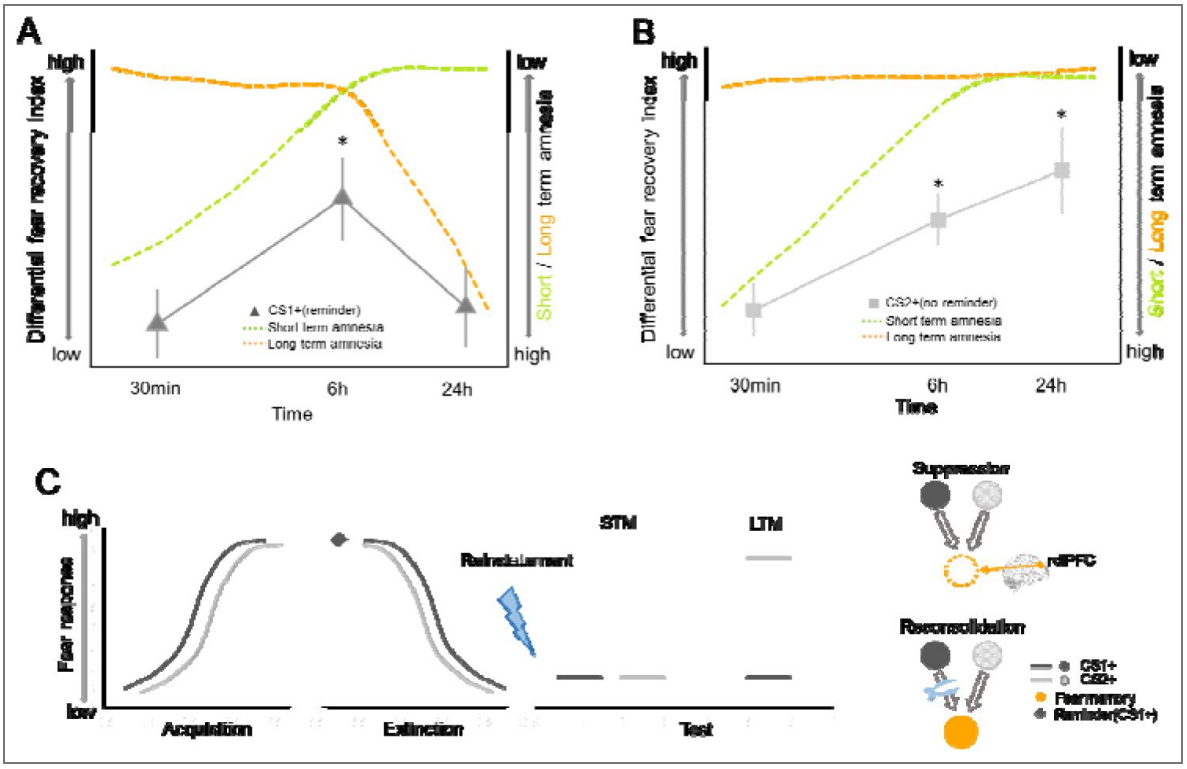



**Figure 6. Time courses of short- and long-term amnesia.**

(**A**) At the short interval (30min to 1h), fear recovery of the reminded CS (CS1+) is inhibited (green). As time progresses (from 6h to 24h), amnesia effect is likely due to the fear memory reconsolidation effect emerged later (orange). Actual SCR data in black solid line. (**B**) Fear amnesia of the non-reminded CS (CS2+) is only evident at the 30-min interval and such effect starts to decay as test interval increases (green). However, long-term amnesia does not generalize to CS2+ (orange, cue-dependence). The observed SCR data (black solid line) paralleled the prediction from both the short-term and the long-term effects as the interval length increased (from 30min to 24h). (**C**) Schematics depicting the effect of cue-reminder on fear memory retention. After both CS1+ (black) and CS2+ (grey) successfully elicit fear responses in the acquisition phase, CS1+ is reminded (black) before both CS+ go through the extinction training. The lack of both CS1+ and CS2+ fear responses in the short-term memory (STM) test (30 min) might be explained by the dlPFC dependent direct suppression effect (dotted circle of US representation). On the other hand, the cue-specific fear amnesia effect in the long-term memory (LTM) test (24h) of CS1+ but not CS2+ could be attributed to the reconsolidation effect specific to CS1+. * $P < 0.05$, Error bars represent standard errors.

study fear memory might also engage automatic memory suppression (Anderson and Floresco, 2022). Consistent with this hypothesis, we showed that the retrieval cue was necessary to produce the short-term amnesia effect (Figure 1) and critically, our results fitted well with the three key characteristics of the active suppression mechanism: temporal dynamics, cue-specificity and the dependence of thought-control ability (Figures 2 and 3).

To calibrate the temporal boundaries of the short and long-term effects, we also included a separate group (medium-term group) where fear expression was tested 6 hours after the retrieval-extinction training. Interestingly, fear returned at this time point (Figure 6A and B), confirming the 6-hour lower time boundary for the hypothesized memory reconsolidation in previous animal and human studies (Agren, 2014; Agren et al., 2012; Debiec et al., 2002; Kindt and Soeter, 2018; Lee, 2008; Nader, Schafe, & Le Doux, 2000; Speer et al., 2021). However, the return of fear expression in the medium-term group also indicates that the short-term fear amnesia is transient, laying out the upper time boundary for the hypothesized fear memory suppression effect. Although mixed results have been reported regarding the durability of suppression effects in the declarative memory studies (Meier et al., 2011; Storm et al., 2012), future research will be needed to investigate whether the short-term effect we observed is specifically related to associative memory or the spontaneous nature of suppression as in RIF (Figure 6C). Previous research also showed that separate mnemonic processes can be involved after emotional memory retrieval. For example, by varying the duration length and repetition number of memory cue re-exposure, different levels of memory persistence (reconsolidation, extinction or a “limbo” state in between) can be induced, suggesting a neural titration mechanism for memory de-stabilization and re-stabilization (Faliagkas et al., 2018; Merlo et al., 2014). These works identified important “boundary conditions” of memory retrieval in affecting the retention of the maladaptive emotional memories. In our study, however, we showed that even within a boundary condition previously thought to elicit memory reconsolidation, mnemonic processes other than reconsolidation could also be at work, and these processes jointly shape the persistence of fear memory.

Our results also corroborate with a recent study showing that blocking fear reconsolidation using post-retrieval repetitive transcranial magnetic stimulation (rTMS) did not influence the short-term fear memory expression (Borgomaneri et al., 2020b). However, as we demonstrated in studies 2 & 3, memory retrieval and intact rdlPFC function were both necessary for the short-term fear amnesia after the extinction training. Therefore, instead of “intact short-term fear expression by the rTMS over the dlPFC”, their results could also be interpreted such that the rTMS on the dlPFC demolished the otherwise short-term fear amnesia induced by the memory retrieval cue. Indeed, a series of neuroimaging studies have revealed that the prefrontal neural network including anterior cingulate cortex, anterior ventrolateral prefrontal cortex and the dlPFC is directly engaged in purging the unwanted declarative memory from consciousness (Anderson et al., 2004; Benoit and Anderson, 2012; Levy and Anderson, 2002). Moreover, previous research has established a causal relationship between dlPFC activity and the memory suppression mechanism underlying retrieval-induced forgetting using transcranial direct current stimulation (tDCS) (Penolazzi et al., 2014; Stramaccia et al., 2017; Valle et al., 2020). In our experiment (studies 2 and 3), when the dlPFC function was intact, subjects’ thought-control abilities were positively correlated with fear amnesia in the test phase. Such correlations disappeared in the no-reminder groups (NR-VER and NR-PFC) and the R-PFC group (Figures 3 and 5), establishing a causal link between the dlPFC activity and short-term fear amnesia.

It should be noted that while our long-term amnesia results were consistent with the fear memory reconsolidation literatures, there were also studies that failed to observe fear prevention (Chalkia, Schroyens, et al., 2020; Chalkia, Van Oudenhove, & Beckers, 2020; Schroyens et al., 2023). Although the memory reconsolidation framework provides a viable explanation for the long-term amnesia, more evidence is required to validate the presence of reconsolidation, especially at the neurobiological level (Elsey et al., 2018). While it is beyond the scope of the current study to discuss the discrepancies between these studies, one possibility to reconcile these results concerns the procedure for the retrieval-extinction training. It has been shown that the eligibility for old

memory to be updated is contingent on whether the old memory and new observations can be inferred to have been generated by the same latent cause (Gershman et al., 2017; Gershman and Niv, 2012). For example, prevention of the return of fear memory can be achieved through gradual extinction paradigm, which is thought to reduce the size of prediction errors to inhibit the formation of new latent causes (Gershman, Jones, et al., 2013). Therefore, the effectiveness of the retrieval-extinction paradigm might depend on the reliability of such paradigm in inferring the same underlying latent cause. Furthermore, other studies highlighted the importance of memory storage *per se* and suggested that memory retention was encoded in the memory engram cell ensemble connectivity whereas the engram cell synaptic plasticity is crucial for memory retrieval (Ryan et al., 2015; Tonegawa, Liu, et al., 2015; Tonegawa, Pignatelli, et al., 2015). It remains to be tested how the cue-independent short-term and cue-dependent long-term amnesia effects we observed could correspond to the engram cell synaptic plasticity and functional connectivity among engram cell ensembles (Figure 6). This is particularly important, since the cue-independent characteristic of the short-term amnesia suggest that either different memory cues fail to evoke engram cell activities, or the retrieval-extinction training transiently inhibits connectivity among engram cell ensembles. Finally, SCR is only one aspect of the fear expression, how the retrieval-extinction paradigm might affect subjects' other emotional (such as the startle response) and cognitive fear expressions such as reported fear expectancy needs to be tested in future studies since they do not always align with each other (Kindt et al., 2009; Sevenster et al., 2012, 2013).

These results help shed light on the dynamics of memory modulation after memory activation and designing novel treatment of psychiatric disorders caused by excessive fear or anxiety (Chen et al., 2021). Memory reconsolidation and suppression both have been studied thoroughly in separate fields. Specifically, aversive memories are usually associated with different retrieval cues, and interfering memory reconsolidation only blocks fear memory recall from the retrieval memory probe but not the others (Debiec et al., 2002; Kindt and Soeter, 2018; Lee, 2008; Nader, Schafe, & Le Doux, 2000; Speer et al., 2021). On the other hand, memory suppression is cue-independent and directly suppresses the fear memory trace yet only effective with limited durability (MacLeod and Macrae, 2001). Although future research is clearly needed to understand the brain mechanisms for the short-term amnesia and its potential connection to the memory suppression effect, especially in relevance to clinical populations such as posttraumatic stress disorder (PTSD) and phobia patients (Homan et al., 2019; Stramaccia et al., 2021), our results provide a general framework to highlight the potential of memory retrieval to triggering different memory updating mechanisms with unique temporal dynamics.

# Data availability

The data and analysis codes for all studies are publicly accessible at *https://osf.io/9agvk/*.

# Additional information

## Open Practices Statement

Neither of the studies reported in this article was preregistered. The data and analysis codes for all studies are publicly accessible at *https://osf.io/9agvk/*.

## Funding


| Funder | Grant reference number | Author |
|---|---|---|
| National Science and Technology Major Project (国家科技重大专项) | 2021ZD0203702 | Jian Li |
| MOST \| National Natural Science Foundation of China (NSFC) | 32441111 | Jian Li |

## Author ORCID iDs

**Yinmei Ni**: http://orcid.org/0000-0003-3614-1828
**Daniela Schiller**: http://orcid.org/0000-0002-0357-7724
**Jian Li**: http://orcid.org/0000-0002-3941-2622


# Additional files

***Supplemental file***

# References


1. **Agren T** (2014) Human reconsolidation: a reactivation and update. *Brain Res Bull* **105**:70-82 https://doi.org/10.1016/j.brainresbull.2013.12.010
2. **Agren T.**, Engman J., Frick A., Björkstrand J., Larsson E. M., Furmark T., Fredrikson M (2012) Disruption of reconsolidation erases a fear memory trace in the human amygdala. *Science* **337**:1550-1552 https://doi.org/10.1126/science.1223006
3. **Alberini C. M** (2005) Mechanisms of memory stabilization: are consolidation and reconsolidation similar or distinct processes?. *Trends Neurosci* **28**:51-56 https://doi.org/10.1016/j.tins.2004.11.001
4. **Anderson M. C.**, Bjork E. L., Bjork R. A (2000) Retrieval-induced forgetting: evidence for a recall-specific mechanism. *Psychon Bull Rev* **7**:522-530 https://doi.org/10.3758/bf03214366
5. **Anderson M. C.**, Bjork R. A., Bjork E. L (1994) Remembering can cause forgetting: retrieval dynamics in long-term memory. *Journal of experimental psychology: Learning, memory, and cognition* **20**:1063-1087 https://doi.org/10.1037//0278-7393.20.5.1063
6. **Anderson M. C.**, Floresco S. B (2022) Prefrontal-hippocampal interactions supporting the extinction of emotional memories: the retrieval stopping model. *Neuropsychopharmacology* **47**:180-195 https://doi.org/10.1038/s41386-021-01131-1
7. **Anderson M. C.**, Green C (2001) Suppressing unwanted memories by executive control. *Nature* **410**:366-369 https://doi.org/10.1038/35066572
8. **Anderson M. C.**, Ochsner K. N., Kuhl B., Cooper J., Robertson E., Gabrieli S. W., Glover G. H., Gabrieli J. D (2004) Neural systems underlying the suppression of unwanted memories. *Science* **303**:232-235 https://doi.org/10.1126/science.1089504
9. **Benoit R. G.**, Anderson M. C (2012) Opposing mechanisms support the voluntary forgetting of unwanted memories. *Neuron* **76**:450-460 https://doi.org/10.1016/j.neuron.2012.07.025
10. **Borgomaneri S.**, Battaglia S., Garofalo S., Tortora F., Avenanti A., di Pellegrino G. (2020a) State-Dependent TMS over Prefrontal Cortex Disrupts Fear-Memory Reconsolidation and Prevents the Return of Fear. *Curr Biol* **30**:3672-3679.e3674. https://doi.org/10.1016/j.cub.2020.06.091
11. **Borgomaneri S.**, Battaglia S., Garofalo S., Tortora F., Avenanti A., di Pellegrino G. (2020b) State-Dependent TMS over Prefrontal Cortex Disrupts Fear-Memory Reconsolidation and Prevents the Return of Fear. *Curr Biol* **30**:3672-3679.e3674. https://doi.org/10.1016/j.cub.2020.06.091
12. **Censor N.**, Dimyan M. A., Cohen L. G (2010) Modification of existing human motor memories is enabled by primary cortical processing during memory reactivation. *Curr Biol* **20**:1545-1549 https://doi.org/10.1016/j.cub.2010.07.047
13. **Chalkia A.**, Schroyens N., Leng L., Vanhasbroeck N., Zenses A.-K., Van Oudenhove L., Beckers T. (2020) No persistent attenuation of fear memories in humans: A registered replication of the reactivation-extinction effect. *Cortex* **129**:496-509 https://doi.org/10.1016/j.cortex.2020.04.017
14. **Chalkia A.**, Van Oudenhove L., Beckers T. (2020) Preventing the return of fear in humans using reconsolidation update mechanisms: A verification report of Schiller, et al. (2010). *Cortex* **129**:510-525 https://doi.org/10.1016/j.cortex.2020.03.031

15. **Chen W.**, Li J., Zhang X., Dong Y., Shi P., Luo P., Zheng X (2021) Retrieval-extinction as a reconsolidation-based treatment for emotional disorders:Evidence from an extinction retention test shortly after intervention. *Behav Res Ther* **139**:103831 https://doi.org/10.1016/j.brat.2021.103831

16. **de Voogd L. D.**, Phelps E. A. (2020) A cognitively demanding working-memory intervention enhances extinction Research article. *Scientific report* **10**:7020 https://doi.org/10.1038/s41598-020-63811-0

17. **Debiec J.**, LeDoux J. E., Nader K (2002) Cellular and systems reconsolidation in the hippocampus. *Neuron* **36**:527-538 https://doi.org/10.1016/s0896-6273(02)01001-2

18. **Dudai Y** (2006) Reconsolidation: the advantage of being refocused. *Curr Opin Neurobiol* **16**:174-178 https://doi.org/10.1016/j.conb.2006.03.010

19. **Dunsmoor J. E.**, Campese V. D., Ceceli A. O., LeDoux J. E., Phelps E. A (2015) Novelty-facilitated extinction: providing a novel outcome in place of an expected threat diminishes recovery of defensive responses. *Biological psychiatry* **78**:203-209 https://doi.org/10.1016/j.biopsych.2014.12.008

20. **Dunsmoor J. E.**, Otto A. R., Phelps E. A (2017) Stress promotes generalization of older but not recent threat memories. *Proceedings of the National Academy of Sciences of the United States of America* **114**:9218-9223 https://doi.org/10.1073/pnas.1704428114

21. **Elsey J. W. B.**, Van Ast V. A., Kindt M. (2018) Human memory reconsolidation: A guiding framework and critical review of the evidence. *Psychological bulletin* **144**:797-848 https://doi.org/10.1037/bul0000152

22. **Engen H. G.**, Anderson M. C (2018) Memory Control: A Fundamental Mechanism of Emotion Regulation. *Trends in Cognitive Sciences* **22**:982-995 https://doi.org/10.1016/j.tics.2018.07.015

23. **Faliagkas L.**, Rao-Ruiz P., Kindt M (2018) Emotional memory expression is misleading: delineating transitions between memory processes. *Current Opinion in Behavioral Sciences* **19**:116-122 https://doi.org/10.1016/j.cobeha.2017.12.018

24. **Faul F.**, Erdfelder E., Buchner A., Lang A.-G (2009) Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* **41**:1149-1160 https://doi.org/10.3758/BRM.41.4.1149

25. **Gagnepain P.**, Hulbert J., Anderson M. C (2017) Parallel Regulation of Memory and Emotion Supports the Suppression of Intrusive Memories. *The Journal of neuroscience* **37**:6423-6441 https://doi.org/10.1523/jneurosci.2732-16.2017

26. **Gershman S. J.**, Jones C. E., Norman K. A., Monfils M. H., Niv Y (2013) Gradual extinction prevents the return of fear: implications for the discovery of state. *Front Behav Neurosci* **7**:164 https://doi.org/10.3389/fnbeh.2013.00164

27. **Gershman S. J.**, Monfils M. H., Norman K. A., Niv Y (2017) The computational nature of memory modification. *eLife* **6** https://doi.org/10.7554/eLife.23763

28. **Gershman S. J.**, Niv Y (2012) Exploring a latent cause theory of classical conditioning. *Learn Behav* **40**:255-268 https://doi.org/10.3758/s13420-012-0080-8

29. **Gershman S. J.**, Schapiro A. C., Hupbach A., Norman K. A (2013) Neural context reinstatement predicts memory misattribution. *The Journal of neuroscience* **33**:8590-8595 https://doi.org/10.1523/jneurosci.0096-13.2013

30. **Heo M.**, Leon A. C (2010) Sample sizes required to detect two-way and three-way interactions involving slope differences in mixed-effects linear models. *J Biopharm Stat* **20**:787-802 https://doi.org/10.1080/10543401003618819

31. **Homan P.**, Levy I., Feltham E., Gordon C., Hu J., Li J., Pietrzak R. H., Southwick S., Krystal J. H., Harpaz-Rotem I. *et al.*(2019) Neural computations of threat in the aftermath of combat trauma. *Nature neuroscience* **22**:470-476 https://doi.org/10.1038/s41593-018-0315-x

32. **Hu J.**, Wang W., Homan P., Wang P., Zheng X., Schiller D (2018) Reminder duration determines threat memory modification in humans. *Scientific report* **8**:8848 https://doi.org/10.1038/s41598-018-27252-0

33. **Hulbert J. C.**, Anderson M. C (2018) What doesn't kill you makes you stronger: Psychological trauma and its relationship to enhanced memory control. *Journal of experimental psychology: General* **147**:1931-1949 https://doi.org/10.1037/xge0000461

34. **Hupbach A.**, Gomez R., Hardt O., Nadel L (2007) Reconsolidation of episodic memories: a subtle reminder triggers integration of new information. *Learn Mem* **14**:47-53 https://doi.org/10.1101/lm.365707

35. **Imai H.**, Kim D., Sasaki Y., Watanabe T (2014) Reward eliminates retrieval-induced forgetting. *Proceedings of the National Academy of Sciences of the United States of America* **111**:17326-17329 https://doi.org/10.1073/pnas.1404469111

36. **Kindt M.**, Soeter M (2018) Pharmacologically induced amnesia for learned fear is time and sleep dependent. *Nature communications* **9**:1316 https://doi.org/10.1038/s41467-018-03659-1

37. **Kindt M.**, Soeter M., Vervliet B (2009) Beyond extinction: erasing human fear responses and preventing the return of fear. *Nature neuroscience* **12**:256-258 https://doi.org/10.1038/nn.2271

38. **Küpper C. S.**, Benoit R. G., Dalgleish T., Anderson M. C (2014) Direct suppression as a mechanism for controlling unpleasant memories in daily life. *Journal of experimental psychology: General* **143**:1443-1449 https://doi.org/10.1037/a0036518

39. **Lee J. L** (2008) Memory reconsolidation mediates the strengthening of memories by additional learning. *Nature neuroscience* **11**:1264-1266 https://doi.org/10.1038/nn.2205

40. **Lee J. L. C.**, Nader K., Schiller D (2017) An Update on Memory Reconsolidation Updating. *Trends in Cognitive Sciences* **21**:531-545 https://doi.org/10.1016/j.tics.2017.04.006

41. **Levy B. J.**, Anderson M. C (2002) Inhibitory processes and the control of memory retrieval. *Trends in Cognitive Sciences* **6**:299-305 https://doi.org/10.1016/s1364-6613(02)01923-x

42. **Liu J.**, Zhao L., Xue Y., Shi J., Suo L., Luo Y., Chai B., Yang C., Fang Q., Zhang Y. *et al.*(2014) An unconditioned stimulus retrieval extinction procedure to prevent the return of fear memory. *Biological psychiatry* **76**:895-901 https://doi.org/10.1016/j.biopsych.2014.03.027

43. **MacLeod M. D.**, Macrae C. N (2001) Gone but not forgotten: the transient nature of retrieval-induced forgetting. *Psychological Science* **12**:148-152 https://doi.org/10.1111/1467-9280.00325

44. **Meier B.**, König A., Parak S., Henke K (2011) Suppressed, but not forgotten. *Swiss Journal of Psychology / Schweizerische Zeitschrift für Psychologie / Revue Suisse de Psychologie* **70**:5-11 https://doi.org/10.1024/1421-0185/a000033

45. **Merlo E.**, Milton A. L., Goozée Z. Y., Theobald D. E., Everitt B. J (2014) Reconsolidation and extinction are dissociable and mutually exclusive processes: behavioral and molecular evidence. *The Journal of neuroscience* **34**:2422-2431 https://doi.org/10.1523/jneurosci.4001-13.2014

46. **Monfils M. H.**, Cowansage K. K., Klann E., LeDoux J. E (2009) Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. *Science* **324**:951-955 https://doi.org/10.1126/science.1167975

47. **Nader K.**, Schafe G. E., Le Doux J. E. (2000) Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* **406**:722-726 https://doi.org/10.1038/35021052

48. **Nader K.**, Schafe G. E., LeDoux J. E (2000) Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature* **406**:722-726 https://doi.org/10.1038/35021052

49. **Penolazzi B.**, Stramaccia D. F., Braga M., Mondini S., Galfano G (2014) Human memory retrieval and inhibitory control in the brain: beyond correlational evidence. *The Journal of neuroscience* **34**:6606-6610 https://doi.org/10.1523/jneurosci.0349-14.2014

50. **Raio C. M.**, Hartley C. A., Orederu T. A., Li J., Phelps E. A (2017) Stress attenuates the flexible updating of aversive value. *Proceedings of the National Academy of Sciences of the United States of America* **114**:11241-11246 https://doi.org/10.1073/pnas.1702565114

51. **Rossi S.**, Hallett M., Rossini P. M., Pascual-Leone A (2009) Safety, ethical considerations, and application guidelines for the use of transcranial magnetic stimulation in clinical practice and research. *Clin Neurophysiol* **120**:2008-2039 https://doi.org/10.1016/j.clinph.2009.08.016

52. **Rossini P. M.**, Burke D., Chen R., Cohen L. G., Daskalakis Z., Di Iorio R., Di Lazzaro V., Ferreri F., Fitzgerald P. B., George M. S. *et al.*(2015) Non-invasive electrical and magnetic stimulation of the brain, spinal cord, roots and peripheral nerves: Basic principles and procedures for routine clinical and research application. An updated report from an I.F.C.N. Committee. *Clin Neurophysiol* **126**:1071-1107 https://doi.org/10.1016/j.clinph.2015.02.001

53. **Ryan T. J.**, Roy D. S., Pignatelli M., Arons A., Tonegawa S (2015) Memory. Engram cells retain memory under retrograde amnesia. *Science* **348**:1007-1013 https://doi.org/10.1126/science.aaa5542

54. **Saunders J.**, MacLeod M. D (2002) New evidence on the suggestibility of memory: the role of retrieval-induced forgetting in misinformation effects. *J Exp Psychol Appl* **8**:127-142

55. **Schacter D. L.**, Addis D. R., Hassabis D., Martin V. C., Spreng R. N., Szpunar K. K (2012) The future of memory: remembering, imagining, and the brain. *Neuron* **76**:677-694 https://doi.org/10.1016/j.neuron.2012.11.001

56. **Schacter D. L.**, Koutstaal W., Norman K. A (1996) Can cognitive neuroscience illuminate the nature of traumatic childhood memories?. *Curr Opin Neurobiol* **6**:207-214 https://doi.org/10.1016/s0959-4388(96)80074-2

57. **Schiller D.**, Kanen J. W., LeDoux J. E., Monfils M. H., Phelps E. A (2013) Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences of the United States of America* **110**:20040-20045 https://doi.org/10.1073/pnas.1320322110

58. **Schiller D.**, LeDoux J. E., Phelps E. A (2020) Reply to Beckers, McIntosh and Chambers on the verification of 'preventing the return of fear using retrieval-extinction in humans'. *PsyArXiv* https://doi.org/10.31234/osf.io/jn6uw

59. **Schiller D.**, Monfils M. H., Raio C. M., Johnson D. C., Ledoux J. E., Phelps E. A (2010) Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* **463**:49-53 https://doi.org/10.1038/nature08637

60. **Schiller D.**, Raio C. M., Phelps E. A (2012) Extinction training during the reconsolidation window prevents recovery of fear. *J Vis Exp: e* **3893** https://doi.org/10.3791/3893

61. **Schroyens N.**, Beckers T., Luyten L (2023) Appraising reconsolidation theory and its empirical validation. *Psychonomic Bulletin & Review* **30**:450-463 https://doi.org/10.3758/s13423-022-02173-2

62. **Sevenster D.**, Beckers T., Kindt M (2012) Retrieval per se is not sufficient to trigger reconsolidation of human fear memory. *Neurobiol Learn Mem* **97**:338-345 https://doi.org/10.1016/j.nlm.2012.01.009

63. **Sevenster D.**, Beckers T., Kindt M (2013) Prediction error governs pharmacologically induced amnesia for learned fear. *Science* **339**:830-833 https://doi.org/10.1126/science.1231357

64. **Speer M. E.**, Ibrahim S., Schiller D., Delgado M. R (2021) Finding positive meaning in memories of negative events adaptively updates memory. *Nature communications* **12**:6601 https://doi.org/10.1038/s41467-021-26906-4

65. **Storm B.**, Bjork E., Bjork R (2012) On the durability of retrieval-induced forgetting. *Journal of Cognitive Psychology* **24**:617-629 https://doi.org/10.1080/20445911.2012.674030

66. **Stramaccia D. F.**, Meyer A. K., Rischer K. M., Fawcett J. M., Benoit R. G (2021) Memory suppression and its deficiency in psychological disorders: A focused meta-analysis. *Journal of experimental psychology: General* **150**:828-850 https://doi.org/10.1037/xge0000971

67. **Stramaccia D. F.**, Penolazzi B., Altoè G., Galfano G (2017) TDCS over the right inferior frontal gyrus disrupts control of interference in memory: A retrieval-induced forgetting study. *Neurobiol Learn Mem* **144**:114-130 https://doi.org/10.1016/j.nlm.2017.07.005

68. **Su S.**, Deng J., Yuan K., Gong Y., Zhang Y., Li H., Cao K., Huang X., Lin X., Wu P. *et al.*(2022) Continuous theta-burst stimulation over the right dorsolateral prefrontal cortex disrupts fear memory reconsolidation in humans. *iScience* **25**:103614 https://doi.org/10.1016/j.isci.2021.103614

69. **Tonegawa S.**, Liu X., Ramirez S., Redondo R (2015) Memory Engram Cells Have Come of Age. *Neuron* **87**:918-931 https://doi.org/10.1016/j.neuron.2015.08.002

70. **Tonegawa S.**, Pignatelli M., Roy D. S., Ryan T. J (2015) Memory engram storage and retrieval. *Curr Opin Neurobiol* **35**:101-109 https://doi.org/10.1016/j.conb.2015.07.009

71. **Valle T. M.**, Bajo M. T., Gómez-Ariza C. J (2020) Cathodal transcranial direct current stimulation over the right dorsolateral prefrontal cortex cancels out the cost of selective retrieval on subsequent analogical reasoning. *Neuropsychologia* **141**:107431 https://doi.org/10.1016/j.neuropsychologia.2020.107431

72. **Wang Y.**, Zhu Z., Hu J., Schiller D., Li J (2021) Active suppression prevents the return of threat memory in humans. *Commun Biol* **4**:609 https://doi.org/10.1038/s42003-021-02120-2

73. **Wells A.**, Davies M. I (1994) The Thought Control Questionnaire: a measure of individual differences in the control of unwanted thoughts. *Behav Res Ther* **32**:871-878 https://doi.org/10.1016/0005-7967(94)90168-6

74. **Wimber M.**, Alink A., Charest I., Kriegeskorte N., Anderson M. C (2015) Retrieval induces adaptive forgetting of competing memories via cortical pattern suppression. *Nature neuroscience* **18**:582-589 https://doi.org/10.1038/nn.3973

# Peer reviews

## Reviewer #1 (Public review):

Summary:

The novel advance by Wang et al is in the demonstration that, relative to a standard extinction procedure, the retrieval-extinction procedure more effectively suppresses responses to a conditioned threat stimulus when testing occurs just minutes after extinction. The authors provide solid evidence to show that this "short-term" suppression of responding involves engagement of the dorsolateral prefrontal cortex.

Strengths:

Overall, the study is well-designed and the results are valuable. There are, however, a few issues in the way that it is introduced and discussed. It would have been useful if the authors could have more explicitly related the results to a theory - it would help the reader understand why the results should have come out the way that they did. More specific comments are presented below.

Please note: The authors appear to have responded to my original review twice. It is not clear that they observed the public review that I edited after the first round of revisions. As part of these edits, I removed the entire section titled Clarifications, Elaborations and Edits

Theory and Interpretation of Results

(1) It is difficult to appreciate why the first trial of extinction in a standard protocol does NOT produce the retrieval-extinction effect. This applies to the present study as well as others that have purported to show a retrieval-extinction effect. The importance of this point comes through at several places in the paper. E.g., the two groups in study 1 experienced a different interval between the first and second CS extinction trials; and the results varied with this interval: a longer interval (10 min) ultimately resulted in less reinstatement of fear than a shorter interval. Even if the different pattern of results in these two groups was shown/known to imply two different processes, there is nothing in the present study that addresses what those processes might be. That is, while the authors talk about mechanisms of memory updating, there is little in the present study that permits any clear statement about mechanisms of memory. The references to a "short-term memory update" process do not help the reader to understand what is happening in the protocol.

In reply to this point, the authors cite evidence to suggest that "an isolated presentation of the CS+ seems to be important in preventing the return of fear expression." They then note the following: "It has also been suggested that only when the old memory and new experience (through extinction) can be inferred to have been generated from the same underlying latent cause, the old memory can be successfully modified (Gershman et al., 2017). On the other hand, if the new experiences are believed to be generated by a different latent cause, then the old memory is less likely to be subject to modification. Therefore, the way the 1st and 2nd CS are temporally organized (retrieval-extinction or standard extinction) might affect how the latent cause is inferred and lead to different levels of fear expression from a theoretical perspective." This merely begs the question: why might an isolated presentation of the CS+ result in the subsequent extinction experiences being allocated to the same memory state as the initial conditioning experiences?

This is not addressed in the paper. The study was not designed to address this question; and that the question did not need to be addressed for the set of results to be interesting. However, understanding how and why the retrieval-extinction protocol produces the effects that it does in the long-term test of fear expression would greatly inform our understanding of how and why the retrieval-extinction protocol has the effects that it does in the short-term tests of fear expression. To be clear; the results of the present study are very interesting - there is no denying that. I am not asking the authors to change anything in response to this point. It simply stands as a comment on the work that has been done in this paper and the area of research more generally.

(2) The discussion of memory suppression is potentially interesting but raises many questions. That is, memory suppression is invoked to explain a particular pattern of results but I, as the reader, have no sense of why a fear memory would be better suppressed shortly after the retrieval-extinction protocol compared to the standard extinction protocol; and why this suppression is NOT specific to the cue that had been subjected to the retrieval-extinction protocol. I accept that the present study was not intended to examine aspects of memory suppression, and that it is a hypothesis proposed to explain the results collected in this study. I am not asking the authors to change anything in response to this point. Again, it simply stands as a comment on the work that has been done in this paper.

(3) The authors have inserted the following text in the revised manuscript: "It should be noted that while our long-term amnesia results were consistent with the fear memory reconsolidation literatures, there were also studies that failed to observe fear prevention (Chalkia, Schroyens, et al., 2020; Chalkia, Van Oudenhove, et al., 2020; Schroyens et al., 2023). Although the memory reconsolidation framework provides a viable explanation for the long-term amnesia, more evidence is required to validate the presence of reconsolidation, especially at the neurobiological level (Elsey et al., 2018). While it is beyond the scope of the current study to discuss the discrepancies between these studies, one possibility to reconcile these results concerns the procedure for the retrieval-extinction training. It has been shown that the eligibility for old memory to be updated is contingent on whether the old memory and new observations can be inferred to have been generated by the same latent cause (Gershman et al., 2017; Gershman and Niv, 2012). For example, prevention of the return of fear memory can be achieved through gradual extinction paradigm, which is thought to reduce the size of prediction errors to inhibit the formation of new latent causes (Gershman, Jones, et al., 2013). Therefore, the effectiveness of the retrieval-extinction paradigm might depend on the reliability of such paradigm in inferring the same underlying latent cause." ***It is perfectly fine to state that "the effectiveness of the retrieval-extinction paradigm might depend on the reliability of such paradigm in inferring the same underlying latent cause..." This is not uninteresting; but it also isn't saying much. Ideally, the authors would have included some statement about factors that are likely to determine whether one is or isn't likely to see a retrieval-extinction effect, grounded in terms of the latent state theories

that have been invoked here. Presumably, the retrieval-extinction protocol has variable effects because of procedural differences that affect whether subjects infer the same underlying latent cause when shifted into extinction. Surely, the clinical implications of any findings are seriously curtailed unless one understands when a protocol is likely to produce an effect; and why the effect occurs at all? This question is rhetorical. I am not asking the authors to change anything in response to this point. Again, it stands as a comment on the work that has been done in this paper; and remains a comment after insertion of the new text, which is acknowledged and appreciated.

(4) The authors find different patterns of responses to CS1 and CS2 when they were tested 30 min after extinction versus 24 h after extinction. On this basis, they infer distinct memory update mechanisms. However, I still can't quite see why the different patterns of responses at these two time points after extinction need to be taken to infer different memory update mechanisms. That is, the different patterns of responses at the two time points could be indicative of the same "memory update mechanism" in the sense that the retrieval-extinction procedure induces a short-term memory suppression that serves as the basis for the longer-term memory suppression (i.e., the reconsolidation effect). My pushback on this point is based on the notion of what constitutes a memory update mechanism; and is motivated by what I take to be a rather loose use of language/terminology in the reconsolidation literature and this paper specifically (for examples, see the title of the paper and line 2 of the abstract).

To be clear: I accept the authors' reply that "The focus of the current manuscript is to demonstrate that the retrieval-extinction paradigm can also facilitate a short-term fear memory deficit measured by SCR". However, I disagree with the claim that any short-term fear memory deficit must be indicative of "update mechanisms other than reconsolidation", which appears on Line 27 in the abstract and very much indicates the spirit of the paper. To make the point: the present study has examined the effectiveness of a retrieval-extinction procedure in suppressing fear responses 30 min, 6 hours and 24 hours after extinction. There are differences across the time points in terms of the level of suppression, its cue specificity, and its sensitivity to manipulation of activity in the dlPFC. This is perfectly interesting when not loaded with additional baggage re separable mechanisms of memory updating at the short and long time points: there is simply no evidence in this study or anywhere else that the short-term deficit in suppression of fear responses has anything whatsoever to do with memory updating. It can be exactly what is implied by the description: a short-term deficit in the suppression of fear responses. Again, this stands as a comment on the work that has been done; and remains a comment for the revised paper.

(5) It is not clear why thought control ability ought to relate to any aspect of the suppression that was evident in the 30 min tests - that is, I accept the correlation between thought control ability and performance in the 30 min tests but would have liked to know why this was looked at in the first place and what, if anything, it means. The issue at hand is that, as best as I can tell, there is no theory to which the result from the short- and long-term tests can be related. The attempts to fill this gap with reference to phenomena like retrieval-induced forgetting are appreciated but raise more questions than answers. This is especially clear in the discussion, where it is acknowledged/stated: "Inspired by the similarities between our results and suppression-induced declarative memory amnesia (Gagnepain et al., 2017), we speculate that the retrieval-extinction procedure might facilitate a spontaneous memory suppression process and thus yield a short-term amnesia effect. Accordingly, the activated fear memory induced by the retrieval cue would be subjected to an automatic fear memory suppression through the extinction training (Anderson and Floresco, 2022)." There is nothing in the subsequent discussion to say why this should have been the case other than the similarity between results obtained in the present study and those in the literature on retrieval induced forgetting, where the nature of the testing is quite different. Again, this is simply a comment on the work that has been done - no change is required for the revised paper.

https://doi.org/10.7554/eLife.98652.3.sa2

## Reviewer #2 (Public review):

Summary

The study investigated whether memory retrieval followed soon by extinction training results in a short-term memory deficit when tested - with a reinstatement test that results in recovery from extinction - soon after extinction training. Experiment 1 documents this phenomenon using a between-subjects design. Experiment 2 used a within-subject control and sees that the effect is also observed in a control condition. In addition, it also revealed that if testing is conducted 6 hours after extinction, there is not effect of retrieval prior to extinction as there is recovery from extinction independently of retrieval prior to extinction. A third Group also revealed that retrieval followed by extinction attenuates reinstatement when the test is conducted 24 hours later, consistent with previous literature. Finally, Experiment 3 used continuous theta-burst stimulation of the dorsolateral prefrontal cortex and assessed whether inhibition of that region (vs a control region) reversed the short-term effect revealed in Experiments 1 and 2. The results of control groups in Experiment 3 replicated the previous findings (short-term effect), and the experimental group revealed that these can be reversed by inhibition of the dorsolateral prefrontal cortex.

Strengths

The work is performed using standard procedures (fear conditioning and continuous theta-burst stimulation) and there is some justification of the sample sizes. The results replicate previous findings - some of which have been difficult to replicate and this needs to be acknowledged - and suggest that the effect can also be observed in a short-term reinstatement test.

The study establishes links between the memory reconsolidation and retrieval-induced forgetting (or memory suppression) literatures. The explanations that have been developed for these are distinct and the current results integrate these, by revealing that the DLPFC activity involved in retrieval-extinction short-term effect. There is thus some novelty in the present results, but numerous questions remain unaddressed.

Weakness

The fear acquisition data is converted to a differential fear SCR and this is what is analysed (early vs late). However, the figure shows the raw SCR values for CS+ and CS- and therefore it is unclear whether acquisition was successful (despite there being an "early" vs "late" effect - no descriptives are provided).

In Experiment 1 (Test results) it is unclear whether the main conclusion stems from a comparison of the test data relative to the last extinction trial ("we defined the fear recovery index as the SCR difference between the first test trial and the last extinction trial for a specific CS") or the difference relative to the CS- ("differential fear recovery index between CS+ and CS-"). It would help the reader assess the data if Fig 1e presents all the indexes (both CS+ and CS-). In addition, there is one sentence which I could not understand "there is no statistical difference between the differential fear recovery indexes between CS+ in the reminder and no reminder groups (P=0.048)". The p value suggests that there is a difference, yet it is not clear what is being compared here. Critically, any index taken as a difference relative to the CS- can indicate recovery of fear to the CS+ or absence of discrimination relative to the CS-, so ideally the authors would want to directly compare responses to the CS+ in the reminder and no-reminder groups. In the absence of such comparison, little can be concluded, in particular if SCR CS- data is different between groups. The latter issue is

particularly relevant in Experiment 2, in which the CS- seems to vary between groups during the test and this can obscure the interpretation of the result.

In experiment 1, the findings suggest that there is a benefit of retrieval followed by extinction in a short-term reinstatement test. In Experiment 2, the same effect is observed to a cue which did not undergo retrieval before extinction (CS2+), a result that is interpreted as resulting from cue-independence, rather than a failure to replicate in a within-subjects design the observations of Experiment 1 (between-subjects). Although retrieval-induced forgetting is cue-independent (the effect on items that are supressed [Rp-] can be observed with an independent probe), it is not clear that the current findings are similar, and thus that the strong parallels made are not warranted. Here, both cues have been extinguished and therefore been equally exposed during the critical stage.

The findings in Experiment 2 suggest that the amnesia reported in experiment 1 is transient, in that no effect is observed when the test is delayed by 6 hours. The phenomena whereby reactivated memories transition to extinguished memories as a function of the amount of exposure (or number of trials) is completely different from the phenomena observed here. In the former, the manipulation has to do with the number of trials (or total amount of time) that the cues are exposed. In the current Experiment 2, the authors did not manipulate the number of trials but instead the retention interval between extinction and test. The finding reported here is closer to a "Kamin effect", that is the forgetting of learned information which is observed with intervals of intermediate length (Baum, 1968). Because the Kamin effect has been inferred to result from retrieval failure, it is unclear how this can be explained here. There needs to be much more clarity on the explanations to substantiate the conclusions.

There are many results (Ryan et al., 2015) that challenge the framework that the authors base their predictions on (consolidation and reconsolidation theory), therefore these need to be acknowledged. These studies showed that memory can be expressed in the absence of the biological machinery thought to be needed for memory performance. The authors should be careful about statements such as "eliminate fear memores" for which there is little evidence.

The parallels between the current findings and the memory suppression literature are speculated in the general discussion, and there is the conclusion that "the retrieval-extinction procedure might facilitate a spontaneous memory suppression process". Because one of the basic tenets of the memory suppression literature is that it reflects an "active suppression" process, there is no reason to believe that in the current paradigm the same phenomenon is in place, but instead it is "automatic". In other words, the conclusions make strong parallels with the memory suppression (and cognitive control) literature, yet the phenomena that they observed is thought to be passive (or spontaneous/automatic). Ultimately, it is unclear why 10 mins between the reminder and extinction learning will "automatically" supress fear memories. Further down in the discussion it is argued that "For example, in the well-known retrieval-induced forgetting (RIF) phenomenon, the recall of a stored memory can impair the retention of related long-term memory and this forgetting effect emerges as early as 20 minutes after the retrieval procedure, suggesting memory suppression or inhibition can occur in a more spontaneous and automatic manner". I did not follow with the time delay between manipulation and test (20 mins) would speak about whether the process is controlled or automatic. In addition, the links with the "latent cause" theoretical framework are weak if any. There is little reason to believe that one extinction trial, separated by 10 mins from the rest of extinction trials, may lead participants to learn that extinction and acquisition have been generated by the same latent cause.

Among the many conclusions, one is that the current study uncovers the "mechanism" underlying the short-term effects of retrieval-extinction. There is little in the current report that uncovers the mechanism, even in the most psychological sense of the mechanism, so this needs to be clarified. The same applies to the use of "adaptive".

Whilst I could access the data in the OFS site, I could not make sense of the Matlab files as there is no signposting indicating what data is being shown in the files. Thus, as it stands, there is no way of independently replicating the analyses reported.

The supplemental material shows figures with all participants, but only some statistical analyses are provided, and sometimes these are different from those reported in the main manuscript. For example, the test data in Experiment 1 is analysed with a two-way ANOVA with main effects of group (reminder vs no-reminder) and time (last trial of extinction vs first trial of test) in the main report. The analyses with all participants in the sup mat used a mixed two-way ANOVA with group (reminder vs no reminder) and CS (CS+ vs CS-). This makes it difficult to assess the robustness of the results when including all participants. In addition, in the supplementary materials there are no figures and analyses for Experiment 3.

One of the overarching conclusions is that the "mechanisms" underlying reconsolidation (long term) and memory suppression (short term) phenomena are distinct, but memory suppression phenomena can also be observed after a 7-day retention interval (Storm et al., 2012), which then questions the conclusions achieved by the current study.

References:

Baum, M. (1968). Reversal learning of an avoidance response and the Kamin effect. Journal of Comparative and Physiological Psychology, 66(2), 495.

Chalkia, A., Schroyens, N., Leng, L., Vanhasbroeck, N., Zenses, A. K., Van Oudenhove, L., & Beckers, T. (2020). No persistent attenuation of fear memories in humans: A registered replication of the reactivation-extinction effect. Cortex, 129, 496-509.

Ryan, T. J., Roy, D. S., Pignatelli, M., Arons, A., & Tonegawa, S. (2015). Engram cells retain memory under retrograde amnesia. Science, 348(6238), 1007-1013.

Storm, B. C., Bjork, E. L., & Bjork, R. A. (2012). On the durability of retrieval-induced forgetting. Journal of Cognitive Psychology, 24(5), 617-629.

Comments on revisions:

Thanks to the authors for trying to address my concerns.

(1 and 2) My point about evidence for learning relates to the fact that in none of the experiments an increase in SCR to the CSs+ is observed during training (in Experiment 1 CS+/CS- differences are even present from the outset), instead what happens is that participants learn to discriminate between the CS+ and CS- and decrease their SCR responding to the safe CS-. This begs the question as to what is being learned, given that the assumption is that the retrieval-extinction treatment is concerned with the excitatory memory (CS+) rather than the CS+/CS- discrimination. For example, Figures 6A and 6B have short/Long term amnesia in the right axes, but it is unclear from the data what memory is being targeted. In Figure 6C, the right panels depicting Suppression and Reconsolidation mechanisms suggest that it is the CS+ memory that is being targeted. Because the dependent measure (differential SCR) captures how well the discrimination was learned (this point relates to point 2 which the authors now acknowledge that there are differences between groups in responding to the CS-), then I struggle to see how the data supports these CS+ conclusions. The fact that influential papers have used this dependent measure (i.e., differential SCR) does not undermine the point that differences between groups at test are driven by differences in responding to the CS-.

(3, 4 and 5) The authors have qualified some of the statements, yet I fail to see some of these parallels. Much of the discussion is speculative and ultimately left for future research to address.

(6) I can now make more sense of the publicly available data, although the files would benefit from an additional column that distinguishes between participants that were included in the final analyses (passed the multiple criteria = 1) and those who did not (did not pass the criteria = 0). Otherwise, anyone who wants to replicate these analyses needs to decipher the multiple inclusion criteria and apply it to the dataset.

https://doi.org/10.7554/eLife.98652.3.sa1

## Author response:

The following is the authors' response to the previous reviews

> ***Reviewer #1 (Public review):***
>
> *Introduction & Theory*
>
> *(1) It is difficult to appreciate why the first trial of extinction in a standard protocol does NOT produce the retrieval-extinction effect. This applies to the present study as well as others that have purported to show a retrieval-extinction effect. The importance of this point comes through at several places in the paper. E.g., the two groups in Study 1 experienced a different interval between the first and second CS extinction trials; and the results varied with this interval: a longer interval (10 min) ultimately resulted in less reinstatement of fear than a shorter interval. Even if the different pattern of results in these two groups was shown/known to imply two different processes, there is nothing in the present study that addresses what those processes might be. That is, while the authors talk about mechanisms of memory updating, there is little in the present study that permits any clear statement about mechanisms of memory. The references to a "short-term memory update" process do not help the reader to understand what is happening in the protocol.*

We agree with the reviewer that whether and how the retrieval-extinction paradigm works is still under debate. Our results provide another line of evidence that such a paradigm is effective in producing long term fear amnesia. The focus of the current manuscript is to demonstrate that the retrieval-extinction paradigm can also facilitate a short-term fear memory deficit measured by SCR. Our TMS study provided some preliminary evidence in terms of the brain mechanisms involved in the causal relationship between the dorsolateral prefrontal cortex (dlPFC) activity and the short-term fear amnesia and showed that both the retrieval interval and the intact dlPFC activity were necessary for the short-term fear memory deficit and accordingly were referred to as the "mechanism" for memory update. We acknowledge that the term "mechanism" might have different connotations for different researchers. We now more explicitly clarify what we mean by "mechanisms" in the manuscript (line 99) as follows:

"In theory, different cognitive mechanisms underlying specific fear memory deficits, therefore, can be inferred based on the difference between memory deficits."

> *In reply to this point, the authors cite evidence to suggest that "an isolated presentation of the CS+ seems to be important in preventing the return of fear expression." They then note the following: "It has also been suggested that only when the old memory and new experience (through extinction) can be inferred to have been generated from the same underlying latent cause, the old memory can be successfully modified (Gershman et al., 2017). On the other hand, if the new experiences are believed to be generated by a different latent cause, then the old memory is less likely to be subject to modification. Therefore, the way the 1stand 2ndCS are temporally organized (retrieval-extinction or standard extinction) might affect how the latent cause is inferred and lead to different*

> *levels of fear expression from a theoretical perspective." This merely begs the question: why might an isolated presentation of the CS+ result in the subsequent extinction experiences being allocated to the same memory state as the initial conditioning experiences? This is not yet addressed in any way.*

As in our previous response, this manuscript is not about investigating the cognitive mechanism why and how an isolated presentation of the CS+ would suppress fear expression in the long term. As the reviewer is aware, and as we have addressed in our previous response letters, both the positive and negative evidence abounds as to whether the retrieval-extinction paradigm can successfully suppress the long-term fear expression. Previous research depicted mechanisms instigated by the single CS+ retrieval at the molecular, cellular, and systems levels, as well as through cognitive processes in humans. In the current manuscript, we simply set out to test that in addition to the long-term fear amnesia, whether the retrieval-extinction paradigm can also affect subjects' short-term fear memory.

> *(2) The discussion of memory suppression is potentially interesting but, in its present form, raises more questions than it answers. That is, memory suppression is invoked to explain a particular pattern of results but I, as the reader, have no sense of why a fear memory would be better suppressed shortly after the retrieval-extinction protocol compared to the standard extinction protocol; and why this suppression is NOT specific to the cue that had been subjected to the retrieval-extinction protocol.*

Memory suppression is the hypothesis we proposed that might be able to explain the results we obtained in the experiments. We discussed the possibility of memory suppression and listed the reasons why such a mechanism might be at work. As we mentioned in the manuscript, our findings are consistent with the memory suppression mechanism on at least two aspects: 1) cue-independence and 2) thought-control ability dependence. We agree that the questions raised by the reviewer are interesting but to answer these questions would require a series of further experiments to disentangle all the various variables and conceptual questions about the purpose of a phenomenon, which we are afraid is out of the scope of the current manuscript. We refer the reviewer to the discussion section where memory suppression might be the potential mechanism for the short-term amnesia we observed (lines 562-569) as follows:

"Previous studies indicate that a suppression mechanism can be characterized by three distinct features: first, the memory suppression effect tends to emerge early, usually 10-30 mins after memory suppression practice and can be transient (MacLeod and Macrae, 2001; Saunders and MacLeod, 2002); second, the memory suppression practice seems to directly act upon the unwanted memory itself (Levy and Anderson, 2002), such that the presentation of other cues originally associated with the unwanted memory also fails in memory recall (cue-independence); third, the magnitude of memory suppression effects is associated with individual difference in control abilities over intrusive thoughts (Küpper et al., 2014)."

> *(3) Relatedly, how does the retrieval-induced forgetting (which is referred to at various points throughout the paper) relate to the retrieval-extinction effect? The appeal to retrieval-induced forgetting as an apparent justification for aspects of the present study reinforces points 2 and 3 above. It is not uninteresting but lacks clarification/elaboration and, therefore, its relevance appears superficial at best.*

We brought the topic of retrieval-induced forgetting (RIF) to stress the point that memory suppression can be unconscious. In a standard RIF paradigm, unlike the think/no-think paradigm, subjects are not explicitly told to suppress the non-target memories. However, to successfully retrieve the target memory, the cognitive system actively inhibits the non-target memories, effectively implementing a memory suppression mechanism (though unconsciously). Therefore, it is possible our results might be explained by the memory suppression framework. We elaborated this point in the discussion section (lines 578-584):

"In our experiments, subjects were not explicitly instructed to suppress their fear expression, yet the retrieval-extinction training significantly decreased short-term fear expression. These results are consistent with the short-term amnesia induced with the more explicit suppression intervention (Anderson et al., 1994; Kindt and Soeter, 2018; Speer et al., 2021; Wang et al., 2021; Wells and Davies, 1994). It is worth noting that although consciously repelling unwanted memory is a standard approach in memory suppression paradigm, it is possible that the engagement of the suppression mechanism can be unconscious."

> *(4) I am glad that the authors have acknowledged the papers by Chalkia, van Oudenhove & Beckers (2020) and Chalkia et al (2020), which failed to replicate the effects of retrieval-extinction reported by Schiller et al in Reference 6. The authors have inserted the following text in the revised manuscript: "It should be noted that while our long-term amnesia results were consistent with the fear memory reconsolidation literature, there were also studies that failed to observe fear prevention (Chalkia, Schroyens, et al., 2020; Chalkia, Van Oudenhove, et al., 2020; Schroyens et al., 2023). Although the memory reconsolidation framework provides a viable explanation for the long-term amnesia, more evidence is required to validate the presence of reconsolidation, especially at the neurobiological level (Elsey et al., 2018). While it is beyond the scope of the current study to discuss the discrepancies between these studies, one possibility to reconcile these results concerns the procedure for the retrieval-extinction training. It has been shown that the eligibility for old memory to be updated is contingent on whether the old memory and new observations can be inferred to have been generated by the same latent cause (Gershman et al., 2017; Gershman and Niv, 2012). For example, prevention of the return of fear memory can be achieved through gradual extinction paradigm, which is thought to reduce the size of prediction errors to inhibit the formation of new latent causes (Gershman, Jones, et al., 2013). Therefore, the effectiveness of the retrieval-extinction paradigm might depend on the reliability of such paradigm in inferring the same underlying latent cause." Firstly, if it is beyond the scope of the present study to discuss the discrepancies between the present and past results, it is surely beyond the scope of the study to make any sort of reference to clinical implications!!!*

As we have clearly stated in our manuscript that this paper was not about discussing why some literature was or was not able to replicate the retrieval-extinction results originally reported by Schiller et al. 2010. Instead, we aimed to report a novel short-term fear amnesia through the retrieval-extinction paradigm, above and beyond the long-term amnesia reported before. Speculating about clinical implications of these finding is unrelated to the long-term, amnesia debate in the reconsolidation world. We now refer the reader to several perspectives and reviews that have proposed ways to resolve these discrepancies as follows (lines 642-673).

> *Secondly, it is perfectly fine to state that "the effectiveness of the retrieval-extinction paradigm might depend on the reliability of such paradigm in inferring the same underlying latent cause..." This is not uninteresting, but it also isn't saying much. Minimally, I would expect some statement about factors that are likely to determine whether one is or isn't likely to see a retrieval-extinction effect, grounded in terms of this theory.*

Again, as we have responded many times, we simply do not know why some studies were able to suppress the fear expression using the retrieval-extinction paradigm and other studies weren't. This is still an unresolved issue that the field is actively engaging with, and we now refer the reader to several papers dealing with this issue. However, this is NOT the focus of our manuscript. Having a healthy debate does not mean that every study using the retrieval-extinction paradigm must address the long-standing question of why the retrieval-extinction paradigm is effective (at least in some studies).

> *Clarifications, Elaborations, Edits*
>
> *(5) Some parts of the paper are not easy to follow. Here are a few examples (though there are others):*
>
> *(a) In the abstract, the authors ask "whether memory retrieval facilitates update mechanisms other than memory reconsolidation"... but it is never made clear how memory retrieval could or should "facilitate" a memory update mechanism.*

We meant to state that the retrieval-extinction paradigm might have effects on fear memory, above and beyond the purported memory reconsolidation effect. Sentence modified (lines 25-26) as follows:

“Memory reactivation renders consolidated memory fragile and thereby opens the window for memory updates, such as memory reconsolidation.”

> *(b) The authors state the following: "Furthermore, memory reactivation also triggers fear memory reconsolidation and produces cue specific amnesia at a longer and separable timescale (Study 2, N = 79 adults)." Importantly, in study 2, the retrieval-extinction protocol produced a cue-specific disruption in responding when testing occurred 24 hours after the end of extinction. This result is interesting but cannot be easily inferred from the statement that begins "Furthermore..." That is, the results should be described in terms of the combined effects of retrieval and extinction, not in terms of memory reactivation alone; and the statement about memory reconsolidation is unnecessary. One can simply state that the retrieval-extinction protocol produced a cue-specific disruption in responding when testing occurred 24 hours after the end of extinction.*

The sentence the reviewer referred to was in our original manuscript submission but had since been modified based on the reviewer’s comments from last round of revision. Please see the abstract (lines 30-35) of our revised manuscript from last round of revision:

“Furthermore, across different timescales, the memory retrieval-extinction paradigm triggers distinct types of fear amnesia in terms of cue-specificity and cognitive control dependence, suggesting that the short-term fear amnesia might be caused by different mechanisms from the cue-specific amnesia at a longer and separable timescale (Study 2, $N$ = 79 adults).”

> *(c) The authors also state that: "The temporal scale and cue-specificity results of the short-term fear amnesia are clearly dissociable from the amnesia related to memory reconsolidation, and suggest that memory retrieval and extinction training trigger distinct underlying memory update mechanisms." ***The pattern of results when testing occurred just minutes after the retrieval-extinction protocol was different to that obtained when testing occurred 24 hours after the protocol. Describing this in terms of temporal scale is unnecessary; and suggesting that memory retrieval and extinction trigger different memory update mechanisms is not obviously warranted. The results of interest are due to the combined effects of retrieval+extinction and there is no sense in which different memory update mechanisms should be identified with the different pattern of results obtained when testing occurred either 30 min or 24 hours after the retrieval-extinction protocol (at least, not the specific pattern of results obtained here).*

Again, we are afraid that the reviewer referred to the abstract in the original manuscript submission, instead of the revised abstract we submitted in the last round. Please see lines 37-39 of the revised abstract where the sentence was already modified (or the abstract from last round of revision).

The facts that the 30min, 6hr and 24hr test results are different in terms of their cue-specificity and thought-control ability dependence are, to us, an important discovery in terms

of delineating different cognitive processes at work following the retrieval-extinction paradigm. We want to emphasize that the fear memories after going through the retrieval-extinction paradigm showed interesting temporal dynamics in terms of their magnitudes, cue-specificity and thought-control ability dependence.

> *(d) The authors state that: "We hypothesize that the labile state triggered by the memory retrieval may facilitate different memory update mechanisms following extinction training, and these mechanisms can be further disentangled through the lens of temporal dynamics and cue-specificities." *** The first part of the sentence is confusing around usage of the term "facilitate"; and the second part of the sentence that references a "lens of temporal dynamics and cue-specificities" is mysterious. Indeed, as all rats received the same retrieval-extinction exposures in Study 2, it is not clear how or why any differences between the groups are attributed to "different memory update mechanisms following extinction"*

The term “facilitate” was used to highlight the fact that the short-term fear amnesia effect is also memory retrieval dependent, as study 1 demonstrated. The novelty of the short-term fear memory deficit can be distinguished from the long-term memory effect via cue-specificity and thought-control ability dependence. Sentence has been modified (lines 97-101) as follows:

“We hypothesize that the labile state triggered by the memory retrieval may facilitate different memory deficits following extinction training, and these deficits can be further disentangled through the lens of temporal dynamics and cue-specificities. In theory, different cognitive mechanisms underlying specific fear memory deficits, therefore, can be inferred based on the difference between memory deficits.”

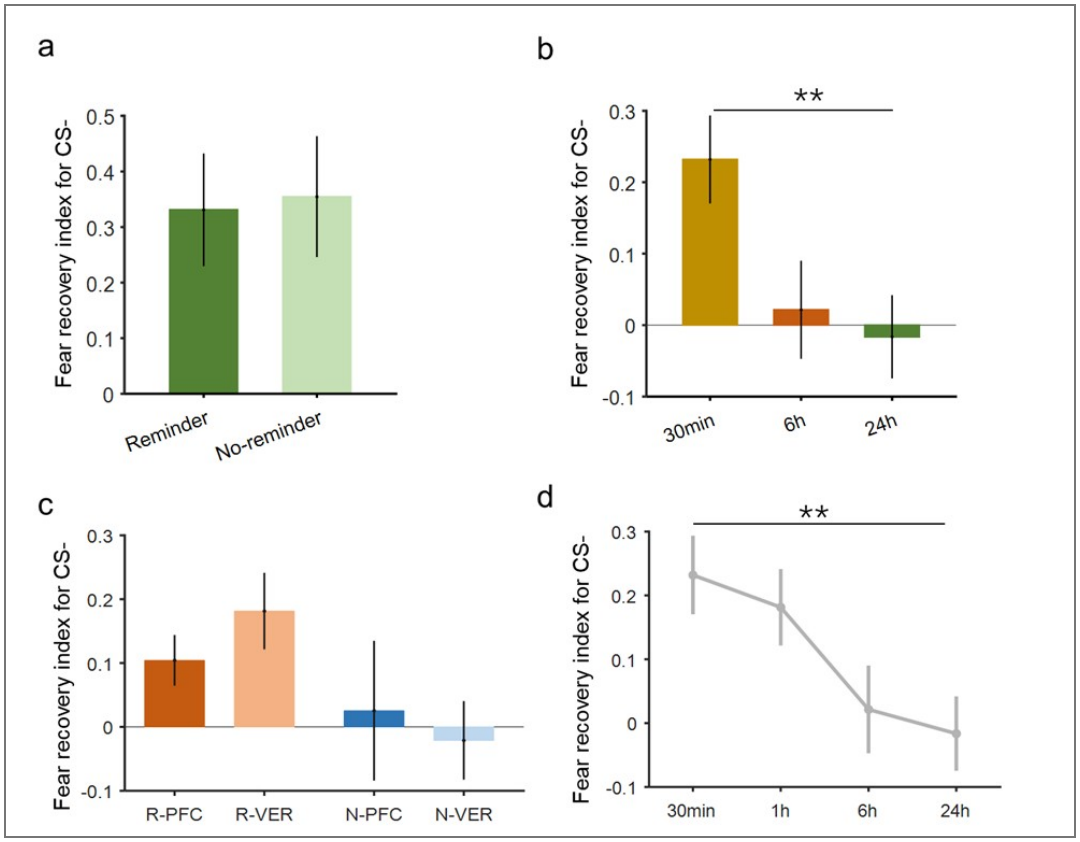
> *Data*
>
> *(6A) The eight participants who were discontinued after Day 1 in Study 1 were all from the no reminder group. The authors should clarify how participants were allocated to the two groups in this experiment so that the reader can better understand why the distribution of non-responders was non-random (as it appears to be).*
>
> *(6B) Similarly, in study 2, of the 37 participants that were discontinued after Day 2, 19 were from Group 30 min and 5 were from Group 6 hours. The authors should comment on how likely these numbers are to have been by chance alone. I presume that they reflect something about the way that participants were allocated to groups: e.g., the different groups of participants in studies 1 and 2 could have been run at quite different times (as opposed to concurrently). If this was done, why was it done? I can't see why the study should have been conducted in this fashion - this is for myriad reasons, including the authors' concerns re SCRs and their seasonal variations.*

As we responded in the previous response letters (as well as in the revised the manuscript), subjects were excluded because their SCR did not reach the threshold of 0.02 S when electric shock was applied. Subjects were assigned to different treatments daily (eg. Day 1 for the reminder group and Day 2 for no-reminder group) to avoid potential confusion in switching protocols to different subjects within the same day. We suspect that the non-responders might be related to the body thermal conditions caused by the lack of central heating for specific dates. Please note that the discontinued subjects (non-responders) were let go immediately after the failure to detect their SCR (< 0.02 S) on Day 1 and never invited back on Day 2, so it’s possible that the discontinued subjects were all from certain dates on which the body thermal conditions were not ideal for SCR collection. Despite the number of excluded subjects, we verified the short-term fear amnesia effect in three separate studies, which to us should serve as strong evidence in terms of the validity of the effect.

> *(6C) In study 2, why is responding to the CS- so high on the first test trial in Group 30 min? Is the change in responding to the CS- from the last extinction trial to the first test trial different across the three groups in this study? Inspection of the figure suggests that it is higher in Group 30 min relative to Groups 6 hours and 24 hours. If this is confirmed by the analysis, it has implications for the fear recovery index which is partly based on responses to the CS-. If not for differences in the CS- responses, Groups 30 min and 6 hours are otherwise identical. That is, the claim of differential recovery to the CS1 and CS2 across time may simply an artefact of the way that the recovery index was calculated. This is unfortunate but also an important feature of the data given the way in which the fear recovery index was calculated.*

We have provided detailed analysis to this question in our previous response letter, and we are posting our previous response there:

Following the reviewer's comments, we went back and calculated the mean SCR difference of CS- between the first test trial and the last extinction trial for all three studies (see Author response image 1 below). In study 1, there was no difference in the mean CS- SCR (between the first test trial and last extinction trial) between the reminder and no-reminder groups (Kruskal-Wallis test $\chi_1^2 = 0.230, P = 0.631$, though both groups showed significant fear recovery even in the CS- condition (Wilcoxon signed rank test, reminder: $P$ = 0.0043, no-reminder: $P$ = 0.0037). Next, we examined the mean SCR for CS- for the 30min, 6h and 24h groups in study 2 and found that there was indeed a group difference (one-way ANOVA,$F_{2.76}$ = 5.3462, $P$ = 0.0067, panel b), suggesting that the CS- related SCR was influenced by the test time (30min, 6h or 24h). We also tested the CS- related SCR for the 4 groups in study 3 (where test was conducted 1 hour after the retrieval-extinction training) and found that across TMS stimulation types (PFC vs. VER) and reminder types (reminder vs. no-reminder) the ANOVA analysis did not yield main effect of TMS stimulation type ($F_{1.71}$ = 0.322, $P$ = 0.572) nor main effect of reminder type ($F_{1.71}$ = 0.0499, $P$ = 0.824, panel c). We added the R-VER group results in study 3 (see panel c) to panel b and plotted the CS- SCR difference across 4 different test time points and found that CS- SCR decreased as the test-extinction delay increased (Jonckheere-Terpstra test, $P$ = 0.00028). These results suggest a natural "forgetting" tendency for CS- related SCR and highlight the importance of having the CS- as a control condition to which the CS+ related SCR was compared with.




**Author response image 1.**

> *(6D) The 6 hour group was clearly tested at a different time of day compared to the 30 min and 24 hour groups. This could have influenced the SCRs in this group and, thereby, contributed to the pattern of results obtained.*

Again, we answered this question in our previous response. Please see the following for our previous response:

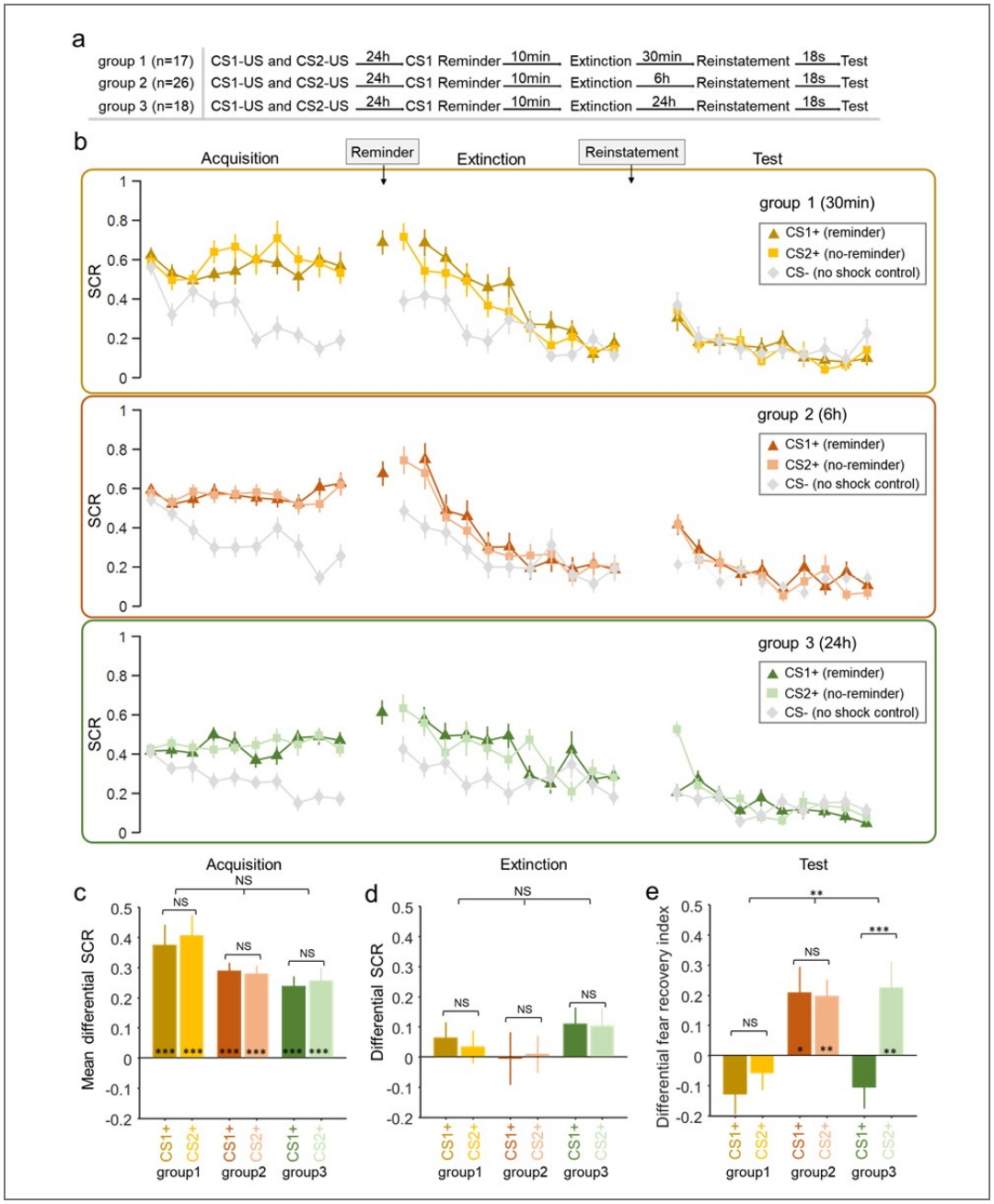
For the 30min and 24h groups, the test phase can be arranged in the morning, in the afternoon or at night. However, for the 6h group, the test phase was inevitably in the afternoon or at night since we wanted to exclude the potential influence of night sleep on the expression of fear memory (see Author response table 1 below). If we restricted the test time in the afternoon or at night for all three groups, then the timing of their extinction training was not matched.

| Test time/Group | Morning | Afternoon | Night | Total# Subjects |
|---|---|---|---|---|
| 30min | 10 | 9 | 8 | 27 |
| 6h | 0 | 12 | 14 | 26 |
| 24h | 8 | 9 | 9 | 26 |

**Author response table 1.**

Nevertheless, we also went back and examined the data for the subjects only tested in the afternoon or at nights in the 30min and 24h groups to match with the 6h group where all the subjects were tested either in the afternoon or at night. According to the table above, we have

17 subjects for the 30min group (9+8),18 subjects for the 24h group (9 + 9) and 26 subjects for the 6h group (12 + 14). As Author response image 2 shows, the SCR patterns in the fear acquisition, extinction and test phases were similar to the results presented in the original figure.




**Author response image 2.**

> *(6E) The authors find different patterns of responses to CS1 and CS2 when they were tested 30 min after extinction versus 24 h after extinction. On this basis, they infer distinct memory update mechanisms. However, I still can't quite see why the different patterns of responses at these two time points after extinction need to be taken to infer different memory update mechanisms. That is, the different patterns of responses at the two time points could be indicative of the same "memory update mechanism" in the sense that the retrieval-extinction procedure induces a short-term memory suppression*

> *that serves as the basis for the longer-term memory suppression (i.e., the reconsolidation effect). My pushback on this point is based on the notion of what constitutes a memory update mechanism; and is motivated by what I take to be a rather loose use of language/terminology in the reconsolidation literature and this paper specifically (for examples, see the title of the paper and line 2 of the abstract).*

As we mentioned previously, the term "mechanism" might have different connotations for different researchers. We aim to report a novel memory deficit following the retrieval-extinction paradigm, which differed significantly from the purported reconsolidation related long-term fear amnesia in terms of its timescale, cue-specificity and thought-control ability. Further TMS study confirmed that the intact dlPFC function is necessary for the short-term memory deficit. It's based on these results we proposed that the short-term fear amnesia might be related to a different cognitive "mechanism". As mentioned above, we now clarify what we mean by "mechanism" in the abstract and introduction (lines 31-34, 97-101).

> ***Reviewer #2 (Public review):***
>
> *The fear acquisition data is converted to a differential fear SCR and this is what is analysed (early vs late). However, the figure shows the raw SCR values for CS+ and CS- and therefore it is unclear whether acquisition was successful (despite there being an "early" vs "late" effect - no descriptives are provided).*
>
> *(1) There are still no descriptive statistics to substantiate learning in Experiment 1.*

We answered this question in our previous response letter. We are sorry that the definition of "early" and "late" trials was scattered in the manuscript. For example, we wrote "the late phase of acquisition (last 5 trials)" (Line 375-376) in the results section. Since there were 10 trials in total for the acquisition stage, we define the first 5 trials and the last 5 trials as "early" and "late" phases of the acquisition stage and explicitly added them into the first occasion "early" and "late" terms appeared (lines 316-318).

In the results section, we did test whether the acquisition was successful in our previous manuscript (Line 316-325):

"To assess fear acquisition across groups (Figure 1B and C), we conducted a mixed two-way ANOVA of group (reminder vs. no-reminder) x time (early vs. late part of the acquisition; first 5 and last 5 trials, correspondingly) on the differential fear SCR. Our results showed a significant main effect of time (early vs. late; $F_{1,55}$ = 6.545, $P$ = 0.013, $\eta^2$ = 0.106), suggesting successful fear acquisition in both groups. There was no main effect of group (reminder vs. no-reminder) or the group x time interaction (group: $F_{1,55}$ = 0.057, $P$ = 0.813, $\eta^2$ = 0.001; interaction: $F_{1,55}$ = 0.066, $P$ = 0.798, $\eta^2$ = 0.001), indicating similar levels of fear acquisition between two groups. Post-hoc *t*-tests confirmed that the fear responses to the CS+ were significantly higher than that of CS- during the late part of acquisition phase in both groups (reminder group: $t_{29}$ = 6.642, $P$ < 0.001; no-reminder group: $t_{26}$ = 8.522, $P$ < 0.001; Figure 1C). Importantly, the levels of acquisition were equivalent in both groups (early acquisition: $t_{55}$ = -0.063, $P$ = 0.950; late acquisition: $t_{55}$ = -0.318, $P$ = 0.751; Figure 1C)."

> *In Experiment 1 (Test results) it is unclear whether the main conclusion stems from a comparison of the test data relative to the last extinction trial ("we defined the fear recovery index as the SCR difference between the first test trial and the last extinction trial for a specific CS") or the difference relative to the CS- ("differential fear recovery index between CS+ and CS-"). It would help the reader assess the data if Fig 1e presents all the indexes (both CS+ and CS-). In addition, there is one sentence which I could not understand "there is no statistical difference between the differential fear recovery indexes between CS+ in the reminder and no reminder groups (P=0.048)". The p value suggests that there is a difference, yet it is not clear what is being compared here. Critically, any index taken as a difference relative to the CS- can indicate recovery of fear*

> *to the CS+ or absence of discrimination relative to the CS-, so ideally the authors would want to directly compare responses to the CS+ in the reminder and no-reminder groups. In the absence of such comparison, little can be concluded, in particular if SCR CS- data is different between groups. The latter issue is particularly relevant in Experiment 2, in which the CS- seems to vary between groups during the test and this can obscure the interpretation of the result.*
>
> *(2) In the revised analyses, the authors now show that CS- changes in different groups (for example, Experiment 2) so this means that there is little to conclude from the differential scores because these depend on CS-. It is unclear whether the effects arise from CS+ performance or the differential which is subject to CS- variations.*

There was a typo in the "P = 0.048" sentence and we have corrected it in our last response letter. Also in the previous response letter, we specifically addressed how the fear recovery index was defined (also in the revised manuscript).

In most of the fear conditioning studies, CS- trials were included as the baseline control. In turn, most of the analyses conducted also involved comparisons between different groups. Directly comparing CS+ trials across groups (or conditions) is rare. In our study 2, we showed that the CS- response decreased as a function of testing delays (30min, 1hr, 6hr and 24hr). Ideally, it would be nice to show that the CS- across groups/conditions did not change. However, even in those circumstances, comparisons are still based on the differential CS response (CS+ minus CS-), that is, the difference of difference. It is also important to note that difference score is important as CS+ alone or across conditions is difficult to interpret, especially in humans, due to noise, signal fluctuations, and irrelevant stimulus features; therefore trials-wise reference is essential to assess the CS+ in the context of a reference stimulus in each trial (after all, the baselines are different). We are listing a few influential papers in the field that the CS- responses were not particularly equivalent across groups/conditions and argue that this is a routine procedure (Kindt & Soeter 2018 Figs. 2-3; Sevenster et al., 2013 Fig. 3; Liu et al., 2014 Fig. 1; Raio et al., 2017 Fig. 2).

> *In experiment 1, the findings suggest that there is a benefit of retrieval followed by extinction in a short-term reinstatement test. In Experiment 2, the same effect is observed to a cue which did not undergo retrieval before extinction (CS2+), a result that is interpreted as resulting from cue-independence, rather than a failure to replicate in a within-subjects design the observations of Experiment 1 (between-subjects). Although retrieval-induced forgetting is cue-independent (the effect on items that are suppressed [Rp-] can be observed with an independent probe), it is not clear that the current findings are similar, and thus that the strong parallels made are not warranted. Here, both cues have been extinguished and therefore been equally exposed during the critical stage.*
>
> *(3) The notion that suppression is automatic is speculative at best*

We have responded the same question in our previous revision. Please note that our results from study 1 (the comparison between reminder and no-reminder groups) was not set up to test the cue-independence hypothesis for the short-term amnesia with only one CS+. Results from both study 2 (30min condition) and study 3 confirmed the cue-independence hypothesis and therefore we believe interpreting results from study 2 as "a failure to replicate in a within-subject design of the observations of Experiment 1" is not the case.

We agree that the proposal of automatic or unconscious memory suppression is speculative and that's why we mentioned it in the discussion. The timescale, cue-specificity and the thought-control ability dependence of the short-term fear amnesia identified in our studies was reminiscent of the memory suppression effects reported in the previous literature. However, memory suppression typically adopted a conscious "suppression" treatment (such as the think/no-think paradigm), which was absent in the current study. However, the

retrieval-induced forgetting (RIF), which is also considered a memory suppression paradigm via inhibitory control, does not require conscious effort to suppress any particular thought. Based on these results and extant literature, we raised the possibility of memory suppression as a potential mechanism. We make clear in the discussion that the suppression hypothesis and connections with RIF will require further evidence (lines 615-616):

“future research will be needed to investigate whether the short-term effect we observed is specifically related to associative memory or the spontaneous nature of suppression as in RIF (Figure 6C).”

> *(4) It still struggle with the parallels between these findings and the "limbo" literature. Here you manipulated the retention interval, whereas in the cited studies the number of extinction (exposure) was varied. These are two completely different phenomena.*

We borrowed the “limbo” term to stress the transitioning from short-term to long-term memory deficits (the 6hr test group). Merlo et al. (2014) found that memory reconsolidation and extinction were dissociable processes depending on the extent of memory retrieval. They argued that there was a “limbo” transitional state, where neither the reconsolidation nor the extinction process was engaged. Our results suggest that at the test delay of 6hr, neither the short-term nor the long-term effect was present, signaling a “transitional” state after which the short-term memory deficit wanes and the long-term deficit starts to take over. We make this idea more explicit as follows (lines 622-626):

“These works identified important “boundary conditions” of memory retrieval in affecting the retention of the maladaptive emotional memories. In our study, however, we showed that even within a boundary condition previously thought to elicit memory reconsolidation, mnemonic processes other than reconsolidation could also be at work, and these processes jointly shape the persistence of fear memory.”

> *(5) My point about the data problematic for the reconsolidation (and consolidation) frameworks is that they observed memory in the absence of the brain substrates that are needed for memory to be observed. The answer did not address this. I do not understand how the latent cause model can explain this, if the only difference is the first ITI. Wouldn't participants fail to integrate extinction with acquisition with a longer ITI?*

We take the sentence “they observed memory in the absence of the brain substrates that are needed for memory to be observed” as referring to the long-term memory deficit in our study. As we responded before, the aim of this manuscript was not about investigating the brain substrates involved in memory reconsolidation (or consolidation). Using a memory retrieval-extinction paradigm, we discovered a novel short-term memory effect, which differed from the purported reconsolidation effect in terms of timescale, cue-specificity and thought-control ability dependence. We further showed that both memory retrieval and intact dlPFC functions were necessary to observe the short-term memory deficit effect. Therefore, we conclude that the brain mechanism involved in such an effect should be different from the one related to the purported reconsolidation effect. We make this idea more explicit as follows (lines 546-547):

“Therefore, findings of the short-term fear amnesia suggest that the reconsolidation framework falls short to accommodate this more immediate effect (Figure 6A and B).”

> *Whilst I could access the data in the OFS site, I could not make sense of the Matlab files as there is no signposting indicating what data is being shown in the files. Thus, as it stands, there is no way of independently replicating the analyses reported.*
>
> *(6) The materials in the OSF site are the same as before, they haven't been updated.*

Last time we thought the main issue was the OSF site not being publicly accessible and thus made it open to all visitors. We have added descriptive file to explain the variables to help visitors to replicate the analyses we took.

> *(7) Concerning supplementary materials, the robustness tests are intended to prove that you 1) can get the same results by varying the statistical models or 2) you can get the same results when you include all participants. Here authors have done both so this does not help. Also, in the rebuttal letter, they stated "Please note we did not include non-learners in these analyses " which contradicts what is stated in the figure captions " (learners + non learners)"*

In the supplementary materials, we did the analyses of varying the statistical models and including both learners and non-learners separately, instead of both. In fact, in the supplementary material Figs. 1 & 2, we included all the participants and performed similar analysis as in the main text and found similar results (learners + non-learners). Also, in the text of the supplementary material, we used a different statistical analysis method to only learners (analyzing subjects reported in the main text using a different method) and achieved similar results. We believe this is exactly what the reviewer suggested us to do. Also there seems to be a misunderstanding for the "Please note we did not include non-learners in these analyses" sentence in the rebuttal letter. As the reviewer can see, the full sentence read “Please note we did not include non-learners in these analyses (the texts of the supplementary materials)”. We meant to express that the Figures and texts in the supplementary material reflect two approaches: 1) Figures depicting re-analysis with all the included subjects (learners + non learners); 2) Text describing different analysis with learners. We added clarifications to emphasize these approaches in the supplementary materials.

> *(8) Finally, the literature suggesting that reconsolidation interference "eliminates" a memory is not substantiated by data nor in line with current theorising, so I invite a revision of these strong claims.*

We agree and have toned down the strong claims.

> *Overall, I conclude that the revised manuscript did not address my main concerns.*

In both rounds of responses, we tried our best to address the reviewer’s concerns. We hope that the clarifications in this letter and revisions in the text address the remaining concerns. Thank you for your feedback.

Reference:

Kindt, M. and Soeter, M. 2018. Pharmacologically induced amnesia for learned fear is time and sleep dependent. Nat Commun, 9, 1316.

Liu, J., Zhao, L., Xue, Y., Shi, J., Suo, L., Luo, Y., Chai, B., Yang, C., Fang, Q., Zhang, Y., Bao, Y., Pickens, C. L. and Lu, L. 2014. An unconditioned stimulus retrieval extinction procedure to prevent the return of fear memory. Biol Psychiatry, 76, 895-901.

Raio, C. M., Hartley, C. A., Orederu, T. A., Li, J. and Phelps, E. A. 2017. Stress attenuates the flexible updating of aversive value. Proc Natl Acad Sci U S A, 114, 11241-11246.

Sevenster, D., Beckers, T., & Kindt, M. 2013. Prediction error governs pharmacologically induced amnesia for learned fear. Science (New York, N.Y.), 339(6121), 830–833.

https://doi.org/10.7554/eLife.98652.3.sa0