nature communications

Article

https://doi.org/10.1038/s41467-024-52733-4

Time-dependent neural arbitration between cue associative and episodic fear memories

Received: 15 September 2023

Accepted: 16 September 2024

Published online: 21 October 2024

Check for updates

Aurelio Cortese[1,8] ✉, Ryu Ohata[2,8], Maria Alemany-González[3,8], Norimichi Kitagawa[4,5], Hiroshi Imamizu[2,6,7] ✉ & Ai Koizumi[3] ✉

After traumatic events, simple cue-threat associative memories strengthen while episodic memories become incoherent. However, how the brain prioritises cue associations over episodic coding of traumatic events remains unclear. Here, we developed an original episodic threat conditioning paradigm in which participants concurrently form two memory representations: cue associations and episodic cue sequence. We discovered that these two distinct memories compete for physiological fear expression, reorganising overnight from an overgeneralised cue-based to a precise sequence-based expression. With multivariate fMRI, we track inter-area communication of the memory representations to reveal that a rebalancing between hippocampal- and prefrontal control of the fear regulatory circuit governs this memory maturation. Critically, this overnight re-organisation is altered with heightened trait anxiety. Together, we show the brain prioritises generalisable associative memories under recent traumatic stress but resorts to selective episodic memories 24 h later. Time-dependent memory competition may provide a unifying account for memory dysfunctions in post-traumatic stress disorders.

Traumatic events impact human memories in multiple ways[1,2]. On one hand, threatening events *strengthen* associative learning to link aversive outcomes with predictive cues in the environment[3,4]. For example, emotionally neutral cues—such as a bicycle—acquire a threatening value after they are experienced in a traumatic vehicle accident. The cues can then evoke fear-like responses in anticipation of the associated threat, which become prolonged and exaggerated in patients suffering from post-traumatic stress disorder (PTSD)[5,6].

Yet, traumatic events may also *weaken* episodic aspects of memories, as reported among some PTSD patients[1,2,5]. One notable deficit is the disruption of sequential memory encoding which cue happened in what order[7,8] (but for debate, see refs. 9,10), a temporal organisation that is fundamental to episodic memory[11,12].

Because there has been little crosstalk between research on each memory domain[13,14], a unifying account of how the brain prioritises cue associations over episodic coding in post-traumatic memory dysfunctions is missing. Given that episodic details of traumatic memories could become initially inaccessible with acute stress[15,16] (but see refs. 9,10), the brain may first rely on generalisable cue associations for immediate defence while switching to episodic coding of threatening memories with time passage. In this study, we developed an episodic threat conditioning paradigm whereby a simulated threatening episode (i.e., a car crash) followed multiple environmental cues presented in specific temporal sequences. Fear memory expression was assessed during conditioning, immediately after, and 24 h later. This design allowed us to study how the brain arbitrates between the concurrently formed cue associative memories and episodic cue sequence memories across days.

Studies in humans and rodents have shown the hippocampus (HPC) and dorsolateral prefrontal cortex (DLPFC) are primary areas

[1]ATR Computational Neuroscience Laboratories, Kyoto, Japan. [2]Department of Psychology, Graduate School of Humanities and Sociology, The University of Tokyo, Tokyo, Japan. [3]Sony Computer Science Laboratories, Inc., Tokyo, Japan. [4]Yoshika Institute of Psychology, Shimane, Japan. [5]BKC Research Organization of Social Sciences, Ritsumeikan University, Shiga, Japan. [6]ATR Cognitive Mechanisms Laboratories, Kyoto, Japan. [7]The Research into Artifacts, Center for Engineering, The University of Tokyo, Tokyo, Japan. [8]These authors contributed equally: Aurelio Cortese, Ryu Ohata, Maria Alemany-González. ✉e-mail: cortese.aurelio@gmail.com; imamizu@l.u-tokyo.ac.jp; bellkoizumi@gmail.com

supporting innocuous sequence learning[17–20]. Moreover, both HPC and DLPFC project to a key centre of the fear circuitry—the ventromedial prefrontal cortex (VMPFC)—with tight functional[21] and anatomical connections[22]. VMPFC, in turn, regulates the expression of cue-threat association memories housed in the Amygdala[4,23]. Whereas the role of the Amygdala alone in human fear memories remains debated[24–26], its role has been supported in a recent large-scale fMRI study[25]. Thus, the Amygdala, together with other brain areas such as VMPFC, may contribute to post-traumatic memory dysfunctions[27–29].

We thus hypothesised that HPC and DLPFC may regulate the transmission of episodic temporal sequences to the VMPFC-Amygdala circuit[25,27], controlling the balance between expression of cue-threat associative memories and episodic memories. To test this hypothesis, we tracked the transmission of episodic sequence representations from both HPC and DLPFC to the VMPFC-Amygdala circuit during, immediately or 24 h after experience of simulated threatening events. Specifically, we used a multivariate analysis of neuroimaging data called information transmission[30–32], which assesses the likelihood that a sequence representation in one brain area (i.e., HPC or DLPFC) is transmitted to a different brain area (i.e., VMPFC-Amygdala circuit).

Since the anticipation of threat is central to pathologies such as PTSD[5,6], we first examined whether physiological defensive responses (i.e., skin conductance reactivity)[3] expressed during the anticipation of threats are governed by cue association memories or episodic sequence memories. We uncovered a phenomenon whereby human participants initially expressed generalised physiological defensive responses based on simple cue associations, which 24 h later transformed into more selective fear responses based on an episodic cue sequence.

Second, with multivariate information transmission analyses, we found that memory maturation was governed by a drastic overnight change in the coupling of HPC and DLPFC with the VMPFC-Amygdala circuit. Specifically, whereas DLPFC consistently communicated episodic temporal sequences during the threat anticipatory epoch across days, HPC only initially communicated sequence representations based on the last cue signalling imminent threat timing but not 24 h later. This time-dependent change in the VMPFC governance was reduced in participants with heightened trait anxiety, a vulnerability factor for post-traumatic psychopathologies[33,34].

Here, we show the brain allows flexible competition between episode-independent overgeneralised fear expression and episode-dependent precise fear expression while prioritising cue-based overgeneralisation of fear under recent emotional stress or heightened trait anxiety. The study may thus provide a unifying account for how two post-traumatic memory consequences (strengthened associative learning and weakened episodic encoding) emerge through a time-dependent competition between HPC and DLPFC to govern the VMPFC-Amygdala circuit.

Results

Our original threat conditioning paradigm assessed both cue-threat associations and episodic sequence memories following experience of simulated threatening events (Fig. 1, Supplementary Movies 1–4). Audiovisual stimuli were created with external image and sound sources including Wikipedia under a Creative Commons licence[35] (see Supplementary text for full descriptions). In a typical threat conditioning paradigm, a specific cue is followed by intrinsically aversive unconditioned stimuli (US)[3]. In the current task, a specific sequence of multiple cues was predictive of the US during a simulated episode of a car accident (Supplementary Movies 1).

Three sound cues (a/b/c) common in traffic scenes (e.g., bicycle bell) were played in one of the three sequences followed by a car crashing accident serving as US. The first sequence (a-b-c) was proportionally followed by US and served as the conditioned sequence, **CS+$_{\mathbf{sequence}}$**. The second sequence (b-a-c) was never followed by US but shared the common last cue with the CS+$_{sequence}$, thus denoted as the element-wise conditioned sequence, **CS+$_{\mathbf{element}}$**. This last element 'c' did have contingency with the immediately following US when the rest of the sequence is not considered (Fig. 1a), which would be treated as a conditioned stimulus in conventional conditioning paradigms[3]. The third sequence (a-c-b) served as **CS−**, as it was never followed by US and had a unique last cue ('b').

Participants could anticipate the US either based on a specific sequence as in the original episode of the accident (CS+$_{sequence}$) or on a conditioned cue ('c' serving the last cue in CS+$_{sequence}$ and CS+$_{element}$). Successful incorporation of sequential information into fear memory would result in a more selective fear expression following only the sequence in the original threatening episodes (CS+$_{sequence}$). Yet, less integrated cue-association memory would result in a more generalised fear expression following both CS+$_{sequence}$ and CS+$_{element}$ based on the conditioned cue ('c').

Participants underwent Acquisition and Immediate test on Day 1, followed by Long-term test on Day 2 (Fig. 1b) while in an MR scanner. We examined whether participants expressed fear-like skin conductance response (SCR) in anticipation of, but without the actual occurrence of, US (see Fig. 1c and *SCR measurements*).

Time-dependent transformation of fear expression

During Acquisition on Day 1, participants expressed defensive physiological responses based on cue association. Their SCR to the two CS+ sequences ending with the last conditioned cue [CS+$_{sequence}$ (a-b-**c**) and CS+$_{element}$ (b-a-**c**) averaged] was significantly larger than SCR to CS− (Fig. 2a, left, Wilcoxon signed-rank test, $z = 2.35$, $P = 0.009$, $r = 0.36$) with no difference between the actual threat-conditioned CS+$_{sequence}$ and the non-conditioned sequence CS+$_{element}$ (Fig. 2b, left, $z = -0.52$, $P_{FDR} = 0.60$, $r = 0.08$).

Similarly, during Immediate test on Day 1, SCR following both CS+$_{sequence}$ and CS+$_{element}$ did not statistically differ from each other (Fig. 2b, centre, $z = -0.34$, $P_{FDR} = 0.78$, $r = 0.05$). The averaged SCR to the two CS+s was no longer significantly higher than SCR to CS− ($z = -0.65$, $P = 0.74$, $r = 0.10$, Fig. 2a, centre), likely due to fear extinction processes initiated by successive presentations of CSs without US reinforcement. Importantly, this did not explain the absence of a differential response between the two CS+s: A supplementary analysis confirmed significant conditioned responses to the two CS+s in the first two trials of Immediate test (relative to CS−; $z = 2.56$, $P = 0.0053$, $r = 0.41$, Supplementary Fig. 1a), but nevertheless no difference in SCR between the two CS+s ($z = -0.67$, $P_{FDR} = 0.50$, $r = 0.11$, Supplementary Fig. 1b).

Strikingly, when participants were tested 24 h later during Long-term test on Day 2, they showed selective anticipatory SCR following the threat-conditioned CS+$_{sequence}$ (Fig. 2b (right), c, Supplementary Fig. 2). That is, only CS+$_{sequence}$ evoked larger SCR than CS− ($z = 1.92$, $P_{FDR} = 0.042$, $r = 0.30$) when CS+$_{element}$ did not ($z = -0.41$, $P_{FDR} = 0.66$, $r = 0.06$). Critically, SCR to CS+$_{sequence}$ was significantly higher than to CS+$_{element}$ ($z = 2.23$, $P_{FDR} = 0.042$, $r = 0.34$). Note that a significant interaction between the two CS+s and three sessions captured the transition from cue-based (Acquisition) to sequence-based fear memory expression (Long-term test) across sessions [linear mixed effect model, interaction CS_{type} x session (Acquisition vs Long-term test): $t_{246} = -2.16$, $P = 0.032$, CI = [−0.088, −0.0041], see Supplementary Table 1 for full model results].

The SCR results collectively reveal a phenomenon in which a fear memory shows a qualitative turnover after 24 h from an immediate cue association into a long-term, more temporally integrated episodic memory.

Neural mechanisms

Next, we turned to fMRI data to examine the mechanisms underlying the fear memory transformation. We speculated that the overnight fear memory maturation (Fig. 2b) may be governed by changes in the

https://doi.org/10.1038/s41467-024-52733-4

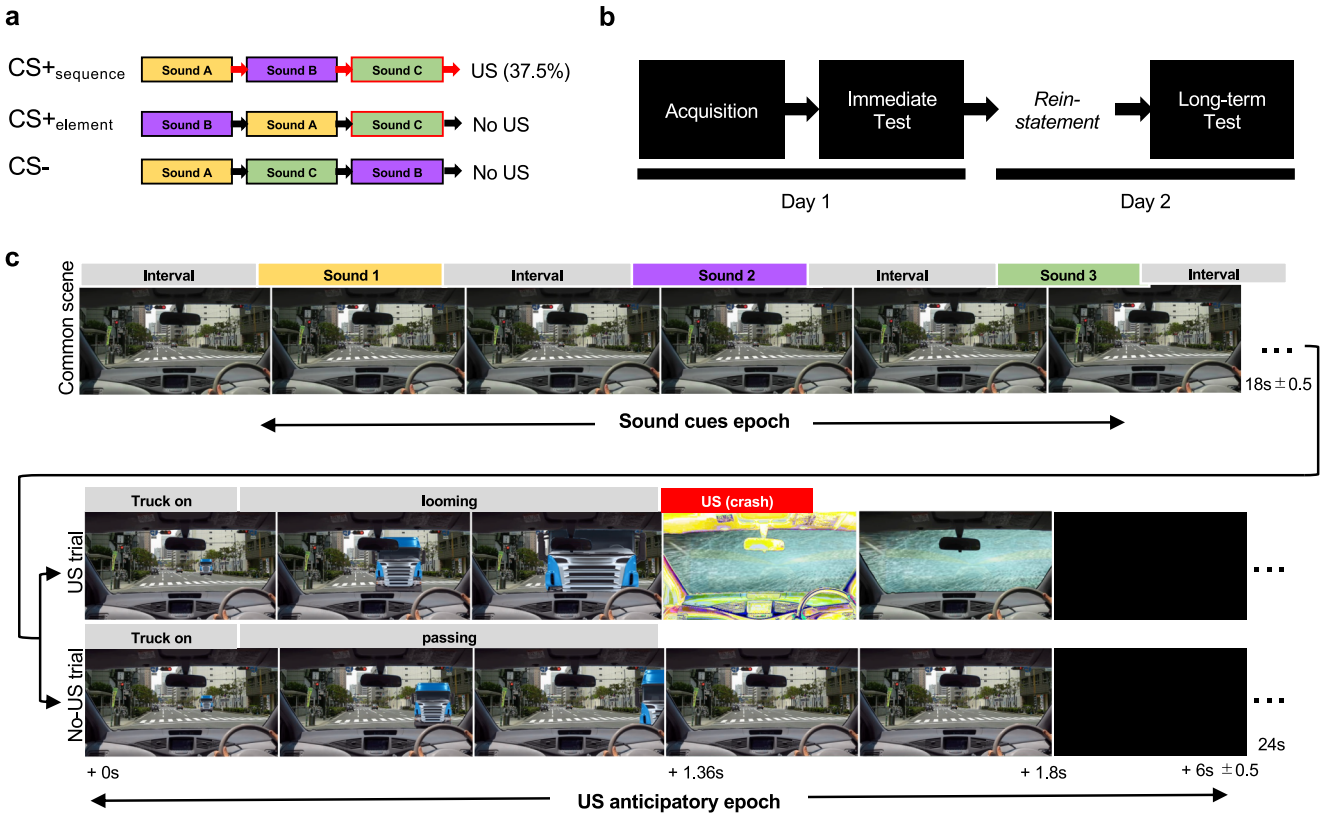

Fig. 1 | Illustration of the task design. a Three sound cues (a/b/c) were presented in three temporal sequences. The first sequence was followed by a car crash (unconditioned stimulus; US), serving as $CS+_{sequence}$. The second sequence was never followed by US but element-wise shared the last conditioned-cue 'c' with $CS+_{sequence}$, hence serving as $CS+_{element}$. The third sequence was never followed by US and did not share the conditioned last cue as the $CS+_{sequence}$, hence serving as CS−. **b** The experimental sessions across 2 days. **c** Schematics of the naturalistic traffic scene played in each trial with and without a car crash (US). The car-truck appears in both types of trials to initiate the US anticipatory epoch. CS conditioned sequence, US unconditioned stimulus. Source images for the car interior and car-truck were purchased from amanaimage and 123RF databases with commercial licences, respectively. The traffic road image was taken from Wikipedia (https://ja.wikipedia.org/wiki/ファイル:四つ橋筋江戸堀1交差点.JPG[35]) with a Creative Commons licence. On the traffic image, a car interior image, a truck car image, and a small yellow dot were superimposed. The sound of a crow call and a standard Japanese traffic light sound effect were downloaded from https://taira-komori.jpn.org/animals01.html and https://necobit.com/necobido/, respectively. We used bicycle bells, background noise, and car crash sounds from the BBC Sound Effects Archives (https://sound-effects.bbcrewind.co.uk/) with commercial licence restrictions. Thus, for demonstrative purposes, in Supplementary Movies 1–4, the sounds of bicycle bells and background noise were replaced with similar commercial licence-free sound files downloaded from https://vsq.co.jp/plus/, the car crash sound was replaced with a similar commercial licence-free file from https://soundeffect-lab.info/. See Supplementary information for the full image and sound credits.

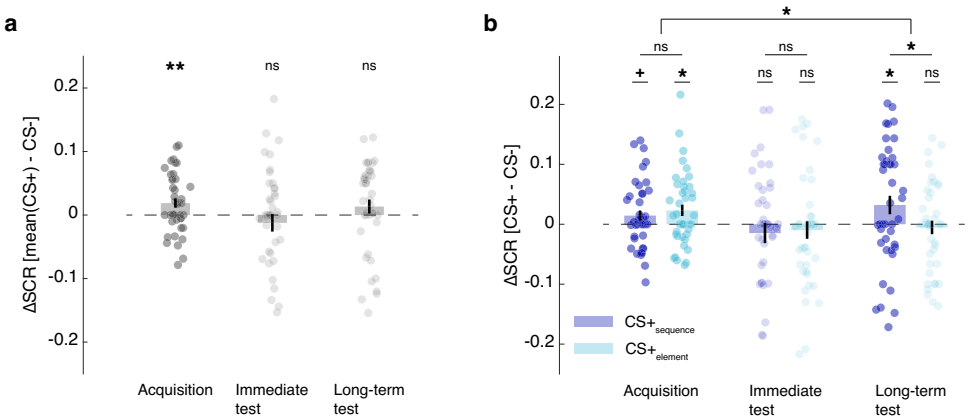

Fig. 2 | Skin conductance response (SCR) to conditioned sequences (CS). a Mean ΔSCR across the two CS+ sequences (sequence and element), after subtracting SCR to CS−. ns: non-significant, $^{**}P = 0.009$ (two-sided Wilcoxon signed-rank test against mean 0). **b** Mean ΔSCR to each CS+ (sequence or element), after subtracting SCR to CS−. Interaction: $^{*}P = 0.032$ (linear mixed effects model); other tests: $^{+}P_{FDR} = 0.094$, $^{*}P_{FDR} < 0.05$ (two-sided Wilcoxon signed-rank test, FDR corrected for multiple comparisons). Note that in both (**a**) and (**b**), ΔSCR larger than 0 at $P < 0.1$ is colour-coded dark for illustrative purposes. Error bars indicate standard errors of means with $N = 42$ individual human participants, in both (**a**) and (**b**). SCR skin conductance response, CS conditioned sequence. Source data provided as a Source Data file.

https://doi.org/10.1038/s41467-024-52733-4

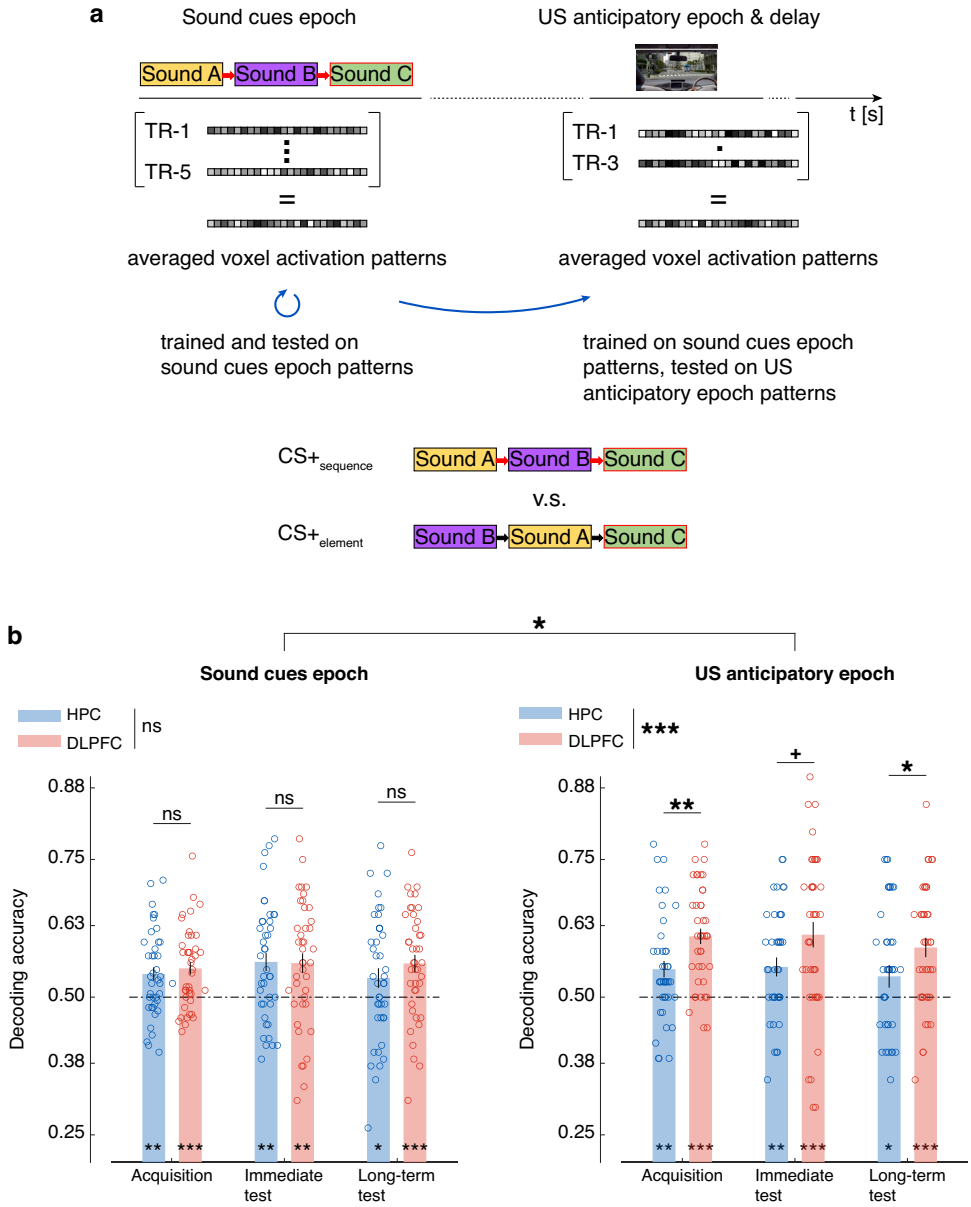

Fig. 3 | Decoding sequence information from multivoxel activity patterns in HPC and DLPFC. a The decoder was trained, within each individual, with the averaged activation patterns during the sound cues epoch of the $CS+_{sequence}$ and $CS+_{element}$ from each session. Then, the decoder was tested on left-out data from the session's sound cues epoch, as well as on data from the US anticipatory epoch. TR (repetition time) denotes each fMRI scan with a 2-s interval. Source images for the car interior and car-truck were purchased from amanaimage and 123RF databases with commercial licences. The traffic road image was taken from Wikipedia (https://ja.wikipedia.org/wiki/ファイル:四つ橋筋江戸堀1交差点.JPG.JPG[35]) with a Creative Commons licence. On the traffic image, a car interior image, a car-truck image, and a small yellow dot were superimposed. See Supplementary information for the full image and sound credits. **b** The accuracy in testing the decoders on the same sound cues epoch (left), or the US anticipatory epoch (right) from the same session. Dashed lines indicate theoretical chance-level decoding accuracy. Error bars indicate standard errors of means with $N = 41$. Open circles indicate individual participants' data points. ns: non-significant; Interaction: $^{*}P = 0.026$ and main effect of ROI: $^{***}P = 0.0005$ (linear mixed effects model); between-ROI comparisons: $^{+}P_{FDR} = 0.05$, $^{*}P_{FDR} = 0.048$, $^{**}P_{FDR} = 0.0017$ (two-sided Wilcoxon signed-rank test, FDR corrected for multiple comparisons); tests against chance: $^{*}P_{FDR} < 0.05$, $^{**}P_{FDR} < 0.01$, $^{***}P_{FDR} < 0.001$ (two-sided Wilcoxon signed-rank test against mean 0.5, FDR corrected for multiple comparisons). CS conditioned sequence, US unconditioned stimulus, HPC hippocampus, DLPFC dorsolateral prefrontal cortex. Source data provided as a Source Data file.

communication of CS sequence representations from HPC and/or DLPFC to the VMPFC-Amygdala circuit. As a prerequisite to test this possibility, we first conducted standard decoding analyses to validate that HPC and DLPFC each harbour sequence representations. We then conducted a dual multivariate pattern analysis (MVPA) called information transmission[30–32] to directly test whether sequence communication from HPC and/or DLPFC to the VMPFC-Amygdala circuit changed over time.

HPC and DLPFC harbour sequence representations

First, to examine whether both HPC and DLPFC harbour sequence representations, we prepared a binary classifier (a "decoder")[36,37], to delineate the CS+ representations ($CS+_{sequence}$ versus $CS+_{element}$) based on the averaged fMRI multivoxel activation patterns during the whole sound cues epoch in each brain area, HPC and DLPFC (Fig. 3a). Note that we did not train a decoder on activation patterns during the US anticipatory epoch, because such decoder would be confounded

by the cue order, favouring the last sound cue presented immediately before the epoch.

Both HPC and DLPFC harboured sequence representations, yielding above-chance decoding accuracy across all sessions (Fig. 3b (left), DLPFC: Acquisition $P_{FDR} = 0.0003$, Immediate test $P_{FDR} = 0.0015$, Long-term test $P_{FDR} = 0.0009$; HPC: Acquisition $P_{FDR} = 0.0018$, Immediate test $P_{FDR} = 0.0013$, Long-term test $P_{FDR} = 0.048$, see Supplementary text for full results, including results from a linear mixed effect model). Since the same constituting cues were used for both CS+s, above-chance decoding accuracy suggests successful encoding of the temporal sequence, rather than a mere summation of the constituting cues. See Supplementary Fig. 3 (left) for additional analyses with binary decoders delineating each CS+ type against CS−, showing similar results.

Critically, both HPC and DLPFC could maintain representations for the two CS+s even during the US anticipatory epoch (Fig. 3b (right), DLPFC: $P_{FDR} < 0.001$ in all sessions; HPC: $P_{FDR} < 0.01$ in Acquisition and Immediate test, $P_{FDR} < 0.05$ in Long-term test, see Supplementary text for full results). See Supplementary Fig. 3 (right) for control analyses with other binary decoders delineating each CS+ type against CS− which could rely on the last cue identity in addition to the cue sequences, and Supplementary Fig. 4 for replicated results with a ternary decoder for three CS types.

Although HPC and DLPFC were equivalent in representing CS+s during the sound cues epoch (Fig. 3b (left), Supplementary Table 2, main effect of ROI: $t_{242} = 0.96$, $P = 0.34$, CI = [−0.012, 0.035]), DLPFC was significantly more robust than HPC in maintaining CS+s representations during the US anticipatory epoch when sounds were no longer played (Fig. 3b (right), Supplementary Table 3, main effect of ROI $t_{242} = 3.51$, $P = 0.0005$, CI = [0.025, 0.089]). This led to a significant ROI x Epoch interaction (Supplementary Table 4, $t_{480} = 2.23$, $P = 0.026$, CI = [0.0058, 0.093]). The DLPFC superiority in the US anticipatory epoch was consistent across sessions (Supplementary Table 3, main effect of session: Acquisition vs Immediate test, $t_{242} = 0.18$, $P = 0.85$, CI = [−0.034, 0.041]; Acquisition vs Long-term test, $t_{242} = -0.95$, $P = 0.34$, CI = [−0.051, 0.018]).

Interestingly, HPC maintained CS sequence representations better during the US anticipatory epoch when it could also rely on their last cue identity. That is, HPC was superior in maintaining separate representations for CS+s versus CS− which differed in their last cue in addition to their temporal sequence, relative to its ability to maintain separate CS+ representations ($CS+_{sequence}$ versus $CS+_{element}$) which differed in their temporal sequence but with a common last cue (Fig. 3b right versus Supplementary Fig. 3a right, Supplementary Table 5, HPC only: main effect of CS_{type} $t_{242} = 2.29$, $P = 0.023$, CI = [0.0038, 0.049]). DLPFC did not benefit from the last cue difference (Supplementary Table 6, DLPFC only: main effect of CS_{type} $t_{240} = -1.69$, $P = 0.093$, CI = [−0.059, 0.0046]), with a significant interaction ROI x CS_{type} (Supplementary Table 7, $t_{480} = -2.55$, $P = 0.011$, CI = [−0.091, −0.012]). This recency effect of HPC to rely on the last cue was again consistent across sessions (Supplementary Fig. 3a, Supplementary Table 5).

In sum, while both HPC and DLPFC maintained representations for episode-specific $CS+_{sequence}$ apart from $CS+_{element}$ during the US anticipatory epoch, DLPFC was consistently superior to HPC across sessions. Meanwhile, only HPC, but not DLPFC, was better at maintaining separate CS representations based on the threat-relevance of the last cue (i.e., CS+s versus CS−) than based only on the threat-relevance of cue sequences (i.e., $CS+_{sequence}$ versus $CS+_{element}$). Nevertheless, the decoding results were consistent across sessions and cannot explain the behavioural transformation from cue-based fear expression on Day 1 to sequence-based fear expression on Day 2 (Fig. 2b). Likewise, whole brain analyses further suggested that accuracies in decoding the neural representations of the episodic-sequences (Fig. 3b) and the last cue (Supplementary Fig. 3a) were statistically similar across sessions in other areas beyond DLPFC and HPC (Supplementary Figs. 5, 6).

Functional coupling of HPC and DLPFC with the fear circuit

As the simple decoding analyses validated that both HPC and DLPFC consistently represented sequence information across days (Fig. 3), we next tested our hypothesis that what may change across days is the degree to which HPC and/or DLPFC regulate fear expression via the VMPFC-Amygdala circuit, based on CS+ sequences representations. Behavioural expression of fear memory may be better explained by interaction between brain areas rather than by representations in single areas[30,38].

To test this hypothesis, we examined the degree to which HPC and DLPFC communicated the CS+s information with the VMPFC-Amygdala circuit. We used multivariate information transmission analysis[30–32] to examine whether the CS+ likelihood ($CS+_{sequence}$ versus $CS+_{element}$) represented in the activation patterns of a seed area (HPC or DLPFC) can be read out from the activation patterns in a target area (combined VMPFC-Amygdala). Whereas the aforementioned decoding results (Fig. 3) would only evaluate neural representations within single areas, this information transmission analysis can assess communications between seed and target areas.

First, we used the sequence decoder (Fig. 3a) to estimate the trial-by-trial CS+ likelihood ($CS+_{sequence}$ versus $CS+_{element}$) in the seed area during the US anticipatory epoch. Then, the activation patterns in the target area (combined VMPFC-Amygdala) were used to predict the decoded CS+ likelihood in the seed areas (see Fig. 4a). Here, higher predictability of CS+ likelihood in the seed areas (HPC or DLPFC) from the activation patterns in the target area would suggest stronger transmission of CS+ representations ($CS+_{sequence}$ versus $CS+_{element}$) from the seed area (e.g., DLPFC) to the target area[30–32].

Information transmission of CS+s representations between HPC/DLPFC and VMPFC-Amygdala revealed a striking resemblance to the SCR results across sessions [Fig. 4b, linear mixed effect model, interaction ROI x Session (Acquisition vs Long-term test), $t_{240} = 2.28$, $P = 0.023$, CI = [0.033, 0.44], non-significant when considering Acquisition vs Immediate test, see Supplementary Table 14]. That is, the transmission of CS+ representations between the HPC and the VMPFC-Amygdala significantly decreased from Day 1 [Acquisition (A) and Immediate test (I)] to Day 2 [Long-term test (L)] (A → L: $z = 3.16$, $P_{FDR} = 0.0024$, $r = 0.49$, I → L: $z = 3.45$, $P_{FDR} = 0.0017$, $r = 0.54$), whereas the transmission between DLPFC and VMPFC-Amygdala remained consistent across sessions (A → L: $z = -0.21$, $P_{FDR} = 0.83$, $r = 0.03$, I → L: $z = -0.41$, $P_{FDR} = 0.83$, $r = 0.06$). This led to a DLPFC advantage over HPC in communicating CS+ information with the VMPFC-Amygdala at Long-term test (Day 2, $z = -3.66$, $P_{FDR} = 0.00038$, $r = 0.57$, Fig. 4b).

Importantly, the emergence of DLPFC advantage over HPC in communicating with the VMPFC-Amygdala circuit could not be explained by mere differences in its ability to maintain the CS+ representations ($CS+_{sequence}$ versus $CS+_{element}$) during the US anticipatory epoch. First, a qualitative comparison indicates the DLPFC superiority over HPC in the information transmission emerged drastically and selectively on Day 2 (Fig. 4b). In contrast, the DLPFC superiority in mere maintenance of sequence information was consistent across days (Fig. 3b). Second, the linear mixed effect models testing main effects of ROI, session, and their interaction, included decoding accuracy as a random effect, and still showed the robust and selective drop of the HPC engagement in the CS+ information transmission relative to DLPFC on Day 2.

Focusing the analysis on the Amygdala (Supplementary Fig. 7a, Supplementary Table 15) and VMPFC (Supplementary Fig. 7b, Supplementary Table 16) alone mirrored the results with the combined target ROI (Fig. 4b).

Several control analyses supported the time-dependent specificity of these results with HPC and VMPFC (Fig. 4b). First, HPC and DLPFC did not show time-dependent alterations of communication with other fear-relevant areas, such as the insula and thalamus that are typically engaged during classical threat conditioning[26] (Supplementary Fig. 8). Second, sensory areas (auditory and fusiform cortices) that

https://doi.org/10.1038/s41467-024-52733-4

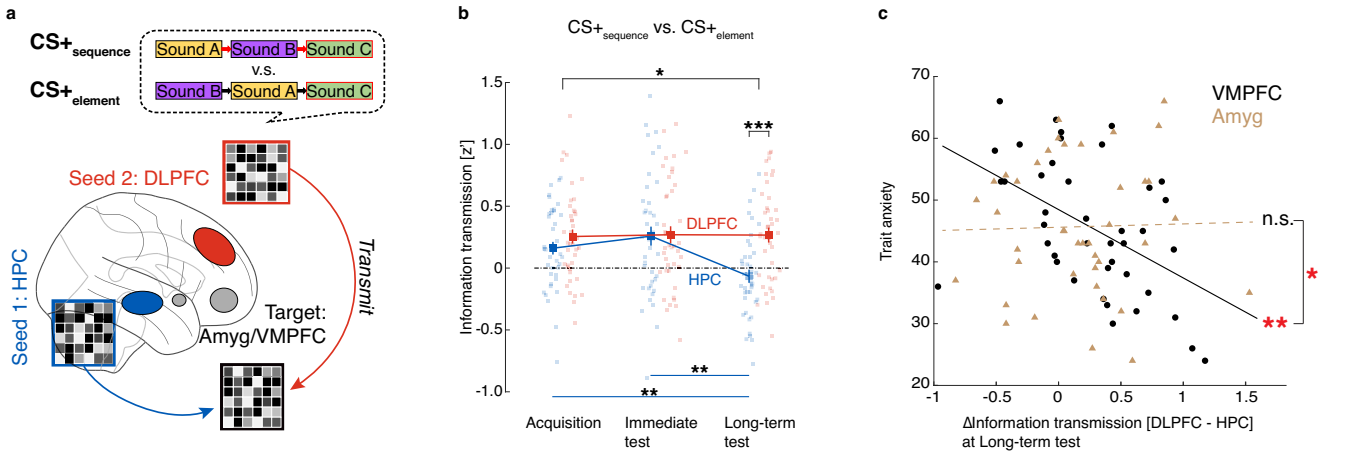

Fig. 4 | Information transmission MVPA to assess communication of sequence representations between brain areas. a Graphic illustration of information transmission analysis. In the illustration, activity patterns in the VMPFC-Amygdala combined ROI (target area) are used to predict the trial-by-trial CS+ sequence likelihood ($CS+_{sequence}$ versus $CS+_{element}$) harboured in the DLPFC or HPC (seed area). **b** The advantage of DLPFC relative to HPC in transmitting CS+ sequence information with the VMPFC-Amygdala circuit emerges selectively at Long-term test on Day 2. Interaction: $^{*}P = 0.023$ (linear mixed effects model), between-session comparisons (HPC): $^{**}P_{FDR} < 0.005$; between-ROI comparison (Long-term test): $^{***}P_{FDR} = 0.00038$ (two-sided Wilcoxon signed-rank test, FDR corrected for multiple comparisons). Error bars indicate standard errors of means with $N = 41$. Larger colour squares represent means where faded colour indicates non-significant (chance level) information transmission. Small coloured squares represent individual participants' data. **c** The correlation between the seed difference (DLPFC versus HPC) in CS+ information transmission ($CS+_{sequence}$ versus $CS+_{element}$) with VMPFC and trait anxiety was significant (Pearson's $r = -0.49$, $P = 0.0012$, robust regression slope $\beta = -12.4$, $P = 0.0003$). The same correlation was absent with the Amygdala (Pearson's $r = 0.025$, $P = 0.88$, robust regression slope $\beta = 0.36$, $P = 0.92$) with a significant difference from the correlation with VMPFC (r-test $z = 2.45$, $P = 0.015$). n.s. non-significant, $^{*}P < 0.05$, $^{**}P < 0.01$ (two-sided). CS conditioned sequence, HPC hippocampus, DLPFC dorsolateral prefrontal cortex, VMPFC ventromedial prefrontal cortex, Amyg Amygdala. Source data provided as a Source Data file.

are less implicated in sequential coding did not show time-dependent communication with the VMPFC-Amygdala circuit (Supplementary Fig. 9).

Heightened anxiety and altered HPC/DLPFC balance in VMPFC communication

Does the temporal re-organisation of the balance between HPC and DLPFC in fear memory governance go awry with heightened trait anxiety, a vulnerable factor for post-traumatic psychopathologies[33,34]. Given that HPC and DLPFC can regulate the Amygdala via VMPFC[23], failure to integrate sequence representations from DLPFC in the fear circuit could emerge selectively at the level of VMPFC. Supporting this idea, the advantage of DLPFC over HPC in communicating CS+ representations with VMPFC on Day 2 robustly correlated with the participants' trait anxiety level (Pearson's $r = -0.49$, $P = 0.0012$, robust regression slope $\beta = -12.4$, $t_{39} = -3.95$, $P = 0.0003$, Fig. 4c). There was no such correlation however between the trait anxiety and the advantage of DLPFC in communicating with the Amygdala (Pearson's $r = 0.025$, $P = 0.88$, robust regression slope $\beta = 0.36$, $t_{39} = 0.10$, $P = 0.92$, Fig. 4c, Amygdala versus VMPFC, r-test: $z = 2.45$, $P = 0.015$). That is, participants with heightened trait anxiety levels showed less DLPFC advantage in communicating with VMPFC, but not with the Amygdala, on Day 2. Although the correlation between SCR and anxiety was non-significant, we found preliminary evidence that episode-specific expression of SCR did not emerge among participants with heightened trait anxiety even on Day 2 (Supplementary Fig. 10). Together, these results suggest that altered turnover of HPC/DLPFC balance in communicating episodic sequence information to VMPFC may contribute to anxiety-related fear memory deficits such as incoherent episodic memories.

Sequence-based fear regulation via DLPFC-VMPFC communication

Was the transformation of HPC/DLPFC balance in their communication with VMPFC driven by $CS+_{sequence}$, $CS+_{element}$, or both representations? One way to resolve these three possibilities is to perform information transmission analyses with the two decoders delineating each CS+ against CS− ([$CS+_{sequence}$ versus CS−] or [$CS+_{element}$ versus CS−], Supplementary Fig. 3). This approach allows estimating the information transmission between a seed ROI (HPC or DLPFC) and VMPFC based on the likelihood of each CS+ (e.g., $CS+_{sequence}$) independently from the other CS+ (e.g., $CS+_{element}$) (Fig. 5a, b).

Surprisingly, DLPFC outcompeted HPC on Day 2 in communicating $CS+_{element}$ representation rather than the $CS+_{sequence}$ representation (Fig. 5a, b, Supplementary text and Supplementary Tables 24, 25). In transmitting $CS+_{element}$ representations (Fig. 5a), HPC-VMPFC communication decayed from Day 1 to Day 2 (A → I: $z = -0.18$, $P_{FDR} = 0.86$, $r = 0.03$, A → L: $z = 1.52$, $P_{FDR} = 0.19$, $r = 0.24$, I → L: $z = 2.25$, $P_{FDR} = 0.074$, $r = 0.35$) to a level indicating very weak coupling (test against 0-chance at Long-term test $z = 0.82$, $P_{FDR} = 0.41$, $r = 0.13$). However, DLPFC-VMPFC communication stayed constant across sessions (A → I: $z = 0.63$, $P_{FDR} = 0.53$, $r = 0.10$, A → L: $z = -1.30$, $P_{FDR} = 0.48$, $r = 0.20$, I → L: $z = -0.99$, $P_{FDR} = 0.48$, $r = 0.15$; all tests against 0-chance $P_{FDR} < 0.01$), resulting in a significant interaction ROI x Sessions (Acquisition vs Long-term test, $t_{240} = 2.38$, $P = 0.018$, CI = [0.043, 0.46], see Supplementary Table 24). Hence, the withdrawal of HPC granted an advantage to DLPFC in communicating $CS+_{element}$ with VMPFC on Day 2 ($z = -3.0$, $P_{FDR} = 0.0041$, $r = 0.47$). Given that $CS+_{element}$ consists of threat-irrelevant sequence and threat-relevant last cue, DLPFC may convey its innocuous sequence information (b-a-c) to VMPFC while HPC may compete to generalise threat prediction based on the last cue ('c') signalling imminent US timing on Day 1.

Supporting this dependence of DLPFC-VMPFC communication on the HPC on-off state, tighter DLPFC-VMPFC communication of $CS+_{element}$ correlated with greater suppression of ΔSCR to $CS+_{element}$ on Day 2 ($CS+_{element}$ - CS−; $r = -0.37$, $P = 0.018$, robust regression slope $\beta = -0.064$, $t_{39} = -2.19$, $P = 0.035$, Fig. 6) when HPC-VMPFC communication was silent (Fig. 5a). On Day 1 when HPC still engaged VMPFC (Fig. 5a), this correlation was not observed ($r = 0.17$, $P = 0.29$, robust regression slope $\beta = 0.022$, $t_{39} = 0.80$, $P = 0.43$, difference between two

https://doi.org/10.1038/s41467-024-52733-4

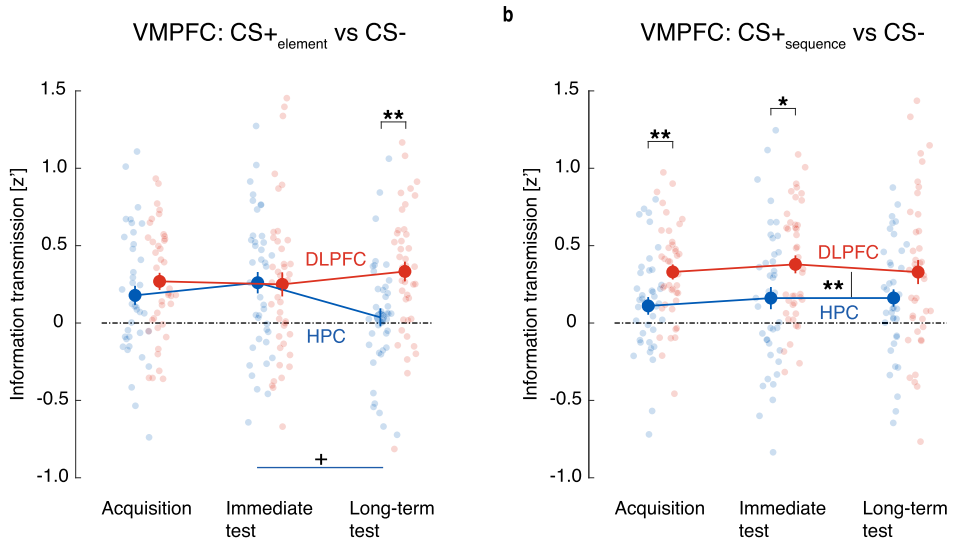

Fig. 5 | The role of HPC and DLPFC in transmitting CS+ versus CS− sequence information via VMPFC. a The analysis used activity patterns in VMPFC to predict the trial-by-trial $CS+_{element}$ likelihood (versus CS−) harboured in the DLPFC or HPC seed. The network was constantly governed by DLPFC across days, while HPC withdrew on Day 2. Error bars indicate standard errors of means with $N = 41$. Between-session comparison: $^{+}P_{FDR} = 0.074$; between-ROI comparison: $^{**}P_{FDR} = 0.0041$ (both two-sided Wilcoxon signed-rank test, FDR corrected for multiple comparisons). **b** Same as (**a**), except that it assessed information transmission of $CS+_{sequence}$ versus CS− representations. The advantage of DLPFC relative to HPC in transmitting sequence information with VMPFC was constant. Larger colour circles represent means where faded colour indicates non-significant (chance level) information transmission. Small coloured circles represent individual participants' data. Error bars indicate standard errors of means with $N = 41$. Main effect: $^{**}P = 0.0017$ (linear mixed effects model); between-ROI comparisons: $^{*}P_{FDR} = 0.024$, $^{**}P_{FDR} = 0.005$ (two-sided Wilcoxon signed-rank test, FDR corrected for multiple comparisons). CS conditioned sequence, HPC hippocampus, DLPFC dorsolateral prefrontal cortex, VMPFC ventromedial prefrontal cortex. Source data provided as a Source Data file.

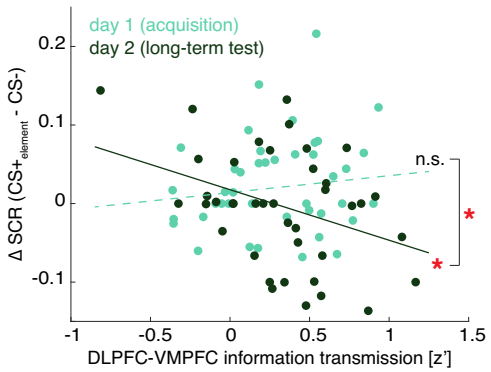

Fig. 6 | DLPFC-VMPFC communication of $CS+_{element}$ representation negatively correlates with SCR towards $CS+_{element}$ on Day 2. The correlation between the DLPFC-VMPFC information transmission of $CS+_{element}$ (versus CS−) with ΔSCR to $CS+_{element}$ (versus CS−) was absent in Acquisition on Day 1 (Pearson $r = 0.17$, $P = 0.29$, robust regression slope $\beta = 0.022$, $P = 0.43$) but emerged in Long-term test on Day 2 (Pearson $r = -0.37$, $P = 0.018$, robust regression slope $\beta = -0.064$, $P = 0.035$). The difference in the correlations between Day 1 (Acquisition) versus Day 2 (Long-term test) was significant (r-test $z = 2.42$, $P = 0.015$). n.s. non-significant, $^{*}P < 0.05$ (two-sided). SCR skin conductance response, CS conditioned sequence, DLPFC dorsolateral prefrontal cortex, VMPFC ventromedial prefrontal cortex. Source data provided as a Source Data file.

days, r-test: $z = 2.42$, $P = 0.015$, Fig. 6, see Supplementary Fig. 13 for similar results with the contrast between Immediate test and Long-term test). Because DLPFC stably communicated CS+ sequences with VMPFC across days (Figs. 4b, 5a, b), the correlational results suggest withdrawal of HPC allowed DLPFC to dampen fear expression to $CS+_{element}$ based on its innocuous sequence, achieving episode-selective fear expression with $CS+_{sequence}$ (Fig. 6).

In contrast, with $CS+_{sequence}$, both HPC and DLPFC displayed constant communication with VMPFC. However, the effect was consistently higher in DLPFC compared with HPC (Fig. 5b, linear mixed effect model, no significant main effects of sessions A-I $t_{240} = 0.84$, $P = 0.40$, CI = [−0.11, 0.28], A-L $t_{240} = 0.61$, $P = 0.54$, CI = [−0.13, 0.24], main effect of ROI $t_{240} = 3.25$, $P = 0.0013$, CI = [0.091, 0.37]; see Supplementary Table 30 for full results, as well as Supplementary text for statistical test results of individual ROIs as well as session-by-session comparisons). Given that $CS+_{sequence}$ consists of the cue sequence (a-b-c) and the last cue ('c') that are both accurately predictive of threat, DLPFC may start communicating the threat-relevant $CS+_{sequence}$ representation to VMPFC already when its earlier cues (a-b-…) are revealed. This removes HPC's need to redundantly communicate the $CS+_{sequence}$ representation to VMPFC based on its threat-relevant last cue ('c'). How and whether the fear circuit 'stops' receiving information after accumulating sufficient threat-relevant evidence remains an open question.

The time-dependent VMPFC-DLPFC communication did not emerge in the control analyses inverting the seed and target ROIs to assess transmission of sequence representations from VMPFC to HPC/DLPFC (Supplementary Fig. 11). Although validating the directionality of communication is generally difficult with fMRI data, these results at least suggest it is the signalling of CS representations in DLPFC to VMPFC, rather than vice versa, that transforms across days.

Finally, DLPFC and HPC similarly communicated each CS+ identity to the Amygdala without time-dependence (Supplementary Fig. 12a, b). While DLPFC has no direct connection with the Amygdala[39], HPC does[40]. Hence, in parallel with the HPC-VMPFC route for fear regulation, HPC may directly communicate the potential imminent threat based on the last cue identity to the Amygdala to prepare for more rapid fear expression across days. However, such fear expression was suppressed when HPC withdrew from its communication with VMPFC on Day 2. This difference between the Amygdala and VMPFC in their time-dependence of CS+ information communication was not observed (Supplementary Fig. 7) unless the communication of specific CS+ (versus CS−) was investigated separately (See also Supplementary Fig. 13 for additional analyses of the Amygdala).

Discussion

Previously, no unifying account could accommodate the parallel post-traumatic memory dysfunctions in cue associations and episodic temporal coding. We report a previously unknown phenomenon whereby human participants initially express generalised physiological fear responses based on a single cue-threat association, which 24 h later transform into selective responses reflecting the temporal sequences of multiple cues (Fig. 2). With information transmission MVPA[30–32] to track inter-area communication of neural representations, we reveal that this memory transformation emerged through the time-dependent rebalancing between HPC and DLPFC in communicating sequence representations to the VMPFC-Amygdala fear circuit (Fig. 4).

In regulating the fear circuit, DLPFC communicated the threat-relevance of temporal cue sequences while HPC prioritised communication of the last cue signalling imminent threats. By recruiting HPC only on the day of the fearful experience of simulated accidents, this dual mechanism arbitrating between DLPFC and HPC may allow flexible gating of fear memory generalisability. On the day of the fearful experience, anticipating threats based on cue sequences alone would potentially lead to an over-conservative prediction of the next threat occurrences. Cue-based fear expression may be prioritised here, given high expectations for similar threats within a shared event boundary[41,42]. However, cue associations may be dismissed with time because maintaining overgeneralised threat anticipation is costly in terms of false alarms.

Critically, such a time-dependent mechanism was altered among participants with higher trait anxiety. Instead, these participants showed an absence of rebalancing between DLPFC and HPC in communicating with the defensive circuit even on Day 2. This result may provide a unifying account for both exaggerated cue association memories[5,6] and incoherent episodic memory[7,8] among PTSD patients, given trait anxiety serves as a vulnerable factor toward post-traumatic psychopathologies[33,34]. Although speculative, heightened trait anxiety may prolong stress responses after threatening events, interfering with the overnight rebalancing between HPC and DLPFC in the regulation of fear memory. This time-dependent mechanism will be an exciting avenue of research as the field has only begun to elucidate the interaction between these types of memory[13,14], potentially opening up new treatment strategies.

Overnight memory consolidation processes can contribute to the qualitative turnover of fear memory after 24 h. In a previous study, memory for cues that were temporally proximal to reward was enhanced overnight, an effect ascribed to consolidatory processes during sleep[43]. In offline states such as rest and sleep, temporal sequences of experiences are replayed to enhance learning[44]. Future studies may examine whether sequential replaying of cues in traumatic episodes during offline states, potentially involving DLPFC, contributes to the enhanced sequence-dependent fear expression after 24 h. This highlights a time-limited opportunity for potential intervention between Day 1 and Day 2, where the post-traumatic memory transformation may be rescued by DLPFC enhancement.

Fear memory is a unique type of memory residing in multiple neural representations (or 'engrams')[45]. One well-known example is the dual representations of original fear and extinction memories, which compete for expression[46]. Despite the altered inter-area communication of sequence information across days (Fig. 4), we found that neural representations of sequence information per se remained consistently available in both HPC and DLPFC with slight overall superiority of DLPFC across days (Fig. 3). Our study thus elucidates a case of multiple memory traces representing generalised-cue-association versus integrated-temporal-episode, each competing for expression over time.

Both HPC and DLPFC roles have been widely demonstrated in innocuous sequence learning[17–20]. However, their functional contributions to the temporal integration of fear memories have remained unclear. Our study reveals their distinct roles in regulating human fear memory expression. DLPFC encodes and communicates temporal sequences made of multiple cues preceding threats, in line with its role in maintaining and manipulating information in working memory[47]. Meanwhile, HPC appears poorer than DLPFC at maintaining the temporal orders of multiple cues and instead communicates the identity of a cue based on its temporal proximity to threat timing.

Though speculative, this bias of HPC to favour a single cue-threat association may be partly driven by projections from the Amygdala[48], especially under immediate traumatic stress or heightened trait anxiety. The transient contribution of HPC in initial memory encoding, but not in long-term memory, is in line with recent findings with innocuous memories[49]. Our results extend this by providing specific evidence on how the overnight withdrawal of HPC leads to fear memory maturation by granting regulatory advantage to DLPFC.

Our original paradigm (Fig. 1) differs from conventional conditioning paradigms involving multiple cues, such as higher-order and contextual threat-conditioning paradigms. First, conventional higher-order conditioning paradigms induce indirect threat-conditioning by pairing two cues, only one of which is directly paired with US before or after the indirect conditioning between the cues[50]. Because two conditioned cues are typically not co-experienced with the US, the paradigm is unlikely to induce a unified episodic memory, including the multiple cues and the US. Second, threat-conditioning paradigms to pair the US with a context or multiple cues typically lack temporal order among the cues because they are often presented simultaneously[51]. Given that temporal configurations such as cue order or gap between CS and US (or lack thereof) modulate threat conditioning[52,53], our paradigm may be more optimal for studying the encoding of sequential order during fearful experiences. Nevertheless, while operational definitions may clearly distinguish among experimental paradigms, a real-life complex emotional experience will likely require more than a single paradigm to describe.

An inspiring future question is whether the current results generalise to the spatial domain. In real life, many cues are bound to specific locations where cue sequences are an intrinsic aspect of experience during spatial navigation. Given the roles of HPC in both spatial and temporal mappings[54], the time-dependent interplay between the HPC and DLPFC may be even more pronounced in regulating fear memory expression conditional on spatiotemporal information.

This study further highlights prospective research questions. First, addressing the directionality of inter-region communications is generally limited with slow fMRI signals. Our information transmission analyses suggested the alternations of communications from HPC and DLPFC to VMPFC rather than vice versa (Supplementary Fig. 11). However, intracranial methods with primates or human patients may more directly validate this point. Work in primates or rodents, especially with viral tracing and optogenetic methods, could shed light on the exact circuit and direction of information flow[55].

Second, although preliminary, we found that most participants expressed sequence-based fear despite their lack of awareness of the contingency between sequence and US (Supplementary Fig. 15). While this is surprising given that awareness is critical for establishing classical threat conditioning when CS and US are temporally apart[56], our results suggest this may not be the case when the sparsely presented CSs and US are intrinsic to ecologically valid threatening scenarios. Our results add to the continued debate on the relationship between physiological and subjective fear[57] and propose a question of how learning sequences reach conscious awareness. Nevertheless, although our US reinforcement rate (37.5%) was similar to previous literature[58,59], it remains to be tested whether higher reinforcement rates may facilitate contingency awareness, especially with stimulus complexity as used in our paradigm.

Together, we reveal a phenomenon whereby human fear memories remodel from simple associative learning into a more complex

episode-dependent fear memory through an overnight transition from HPC to DLPFC governance of VMPFC. This opens a novel time window to potentially target offline temporal re-organisation processes to rescue incoherent episodic memories in PTSD.

Methods

Participants

Forty-four participants (29 males, mean age = 22.0 ± 1.6 years old) were enrolled. They were recruited through a social network service account. They received 10,000 JPY for their 2.5 h participation across two days. Two participants (one male) failed to participate on Day 2, and thus their data were discarded from the analyses. Our first participant was enrolled with several MRI sequence differences from the rest of the participants, and thus their MRI data were also discarded from the analyses while their SCR data were included. Thus, a total of $N = 41$ participants were included for fMRI data analyses and a total of $N = 42$ for SCR analyses. The protocol was approved by the ethical committee of Sony Japan. All participants provided written consent forms before beginning the experiments.

Sample size estimation

We estimated our sample size with a power analysis using G*power 3.1.9.6[60] based on our behavioural pilot study with $N = 17$. To detect the difference in SCR between $CS+_{sequence}$ and $CS+_{element}$ in Long-term test with a two-tailed Wilcoxon signed-rank test yielding an effect size of dz = 0.57 with 95% power, the required sample size was estimated to be $N = 38$. To ensure counterbalancing of sound stimuli as well as the final sample size with some expected drop-outs with a two-day experiment design, we liberally recruited 44 participants, leading to a total of $N = 41$ for MRI and $N = 42$ for behavioural data (see above "Participants" section).

Procedures

Participants underwent three sessions; Acquisition, Immediate test, and Long-term test (Fig. 1b). On any given trial during each session, participants viewed a semi-animated video clip depicting a traffic scene of a cross-road from the perspective of a car driver who is waiting for a red traffic light to turn green (Fig. 1c and Supplementary Movies 1–4). This waiting period was followed by the sudden appearance of a truck that either crashed into the front glass of the driver's car (in some trials of Acquisition session) or simply passed by (in other trials of Acquisition session and in all trials of other sessions). Here, the sight of a truck looming to crash along with a noxious crashing sound (~85 dB) served as a multisensory unconditioned stimulus (US), and the passing of a truck meant an omission of a US on that trial. Aversiveness of the US was ensured with a loud presentation of naturally aversive sound[61] and fear-evoking looming[62]. On each trial, the waiting period included three sound elements; bicycle bell, traffic light melody, and crow calls. The sound elements were played in one of three sequences of a triplet (1) a-b-c, (2) b-a-c, and (3) a-c-b, where each letter (a, b, or c) corresponds to one of the three sound elements (e.g., bicycle bell) in a counterbalanced manner across participants. See "Stimuli" for more detailed descriptions of the video clips.

In Acquisition session, a total of 72 trials were presented across four runs. Each CS sequences trial type was played for 20 times without subsequent car crashing (i.e., 60 no-US trials). Trials with the first sequence a-b-c was played for an additional 12 times, followed by a car crashing (i.e., US trials). Thus, the sequence a-b-c served as a threat-conditioned sequence ($CS+_{sequence}$) (Fig. 1a). The second sequence, b-a-c, was never followed by the US but shared the sound 'c' as the last element, which was immediately followed by the US in the $CS+_{sequence}$, thus serving as the element-wise conditioned sequence $CS+_{element}$. The third sequence, a-c-b, was never followed by the US and did not share the last element 'c', thus serving as unreinforced CS−.

On a small portion of the trials, a video clip was paused after the sound elements were played and participants were prompted with an instruction on the screen to rate the likelihood of getting crashed by a truck on that trial with a four-point scale using a keypad. Specifically, the ratings were prompted on two trials with each of the three CS sequences without the US. The ratings were prompted on two additional $CS+_{sequence}$ trials with the US so that participants would not erroneously associate the presence of likelihood rating with the absence of US. After 6 s elapsed since the instruction, the rest of the video clip was played. Before those US-likelihood rating trials began, there appeared a preparatory instruction on a screen to guide the participants to anticipate the upcoming rating task.

Immediate and Long-term test sessions were similar to Acquisition session, except that no US-trial was presented. In both test sessions, 12 trials were presented with each of the three CS sequences. On two of the trials with each sequence, the US-likelihood ratings were prompted. Acquisition and Immediate test sessions were conducted on Day 1, with a 10-min interval during which another resting-state scan (8.3 min) was taken. Long-term test session was conducted on Day 2. To reinstate fear, Long-term test session was preceded by three unsignaled US. Specifically, a cropped video clip (133.3 s) selectively showing the car crashing scene was played three times without any preceding sound elements. After a subsequent 10-min break during which a resting-state scan (8.3 min) was taken, Long-term test session began.

The trial order was semi-randomised in each session so that a given CS sequence would not be repeated in three or more consecutive trials and that it would not be absent on six or more consecutive trials. In Acquisition session, the US trials were equally distributed across four runs (3 trials each) so that the CS-US contingency would remain stable across the runs. The US-likelihood ratings were equally distributed across the runs (2 trials each) to equate the total duration between the runs.

Participants provided their trait anxiety level with Spielberger State-Trait Anxiety Inventory[63] after the Long-term test on Day 2.

Before enrolment, participants were only informed of the potential occurrence of car crashing scenes to avoid unexpected surprises during the task. They did not receive any instructions on the potential contingency between the cue(s) and/or their sequences with the US (car crash).

Stimuli

A custom-made video clip played as follows: Each video started with a scene of a traffic road depicted from a car driver's perspective, who remained still in front of a red traffic light (Fig. 1c and Supplementary videos). During this period, the backlight of another car idling far ahead kept blinking at 2 Hz. During this waiting period, three sound elements, bicycle bell, traffic light melody, and crow calls, were played in a given sequence, a-b-c ($CS+_{sequence}$), b-a-c ($CS+_{element}$), or a-c-b (CS −). We chose auditory cues rather than visual cues to allow naturalistic manipulation of cue orders and timings as well as decoding of cue information without being embedded in other visual information such as background scenery.

In each trial, after an inter-stimulus interval (ISI) of 2.5 s with a jitter of ±0.5 s, a first sound element was played within an epoch of 2.5 s. After 2.5 s from the onset of the first sound element, a second ISI and sound followed. After the third ISI and sound element, the image of a truck abruptly appeared on a screen at 18 ± 0.5 s from the onset of a trial. On the no-US trials, the truck simply passed by, and the video clip faded into a black screen within a 1.8 s period. On the US trials, the truck gradually approached and crashed into the driver's car front at 1.36 s from the onset of the truck appearance, and the video clip faded for another 0.4 s. In the US trials, the crashing sound was played from the onset of the car crashing. The crashing sound was 2 s in duration. After the blank screen remained for 4.2 s (i.e., inter-trial interval), the

next trial began. Behind the entire video, a background noise depicting naturalistic outdoor space was constantly played. The duration of each video was 24 s without the rating and 30 s with the rating.

The duration of the three sound elements, bicycle bell, traffic light melody, and crow calls, were 0.61 s, 2.5 s, and 2.06 s, respectively. We maintained the naturalistic variability in the duration of each sound element as would be commonly heard in real-life traffic loads. Note that sound elements were counterbalanced across participants by assigning each sound element to the item a, b, or c within sequences (see "Procedures").

The video clips were custom-made from the image and sound sources downloaded from the Internet (see Supplementary videos and text).

Equipment

Sound stimuli were presented with noise-cancelling headphones OptoACTIVE™ (OptoAcoustics, Ltd). Video clips were presented with a Full HD MR projector (Resonance Technology, Inc.) onto a screen placed in a boa of the scanner with a viewing distance of 67 cm. Matlab on a 13' MacBookPro controlled the timings of the sound and the clip presentation. Skin conductance response (SCR) was measured with BIOPAC (BIOPAC Systems, Inc.) disposable electrodes attached to the index and middle fingers of a participant's left hand.

fMRI measurements

Participants were scanned in a Siemens Prisma 3 Tesla MRI scanner (Erlangen, Germany) with a 20-channel head coil installed in the Center for Evolutionary Cognitive Sciences, the University of Tokyo (Tokyo, Japan). For functional echo-planar image (EPI) acquisition; TR = 2000 ms, TE = 27 ms, flip angle = 70°, FOV = 200 mm, Slices = 75, voxel size: 2 × 2 × 2 mm (without gap), phase encoding direction = anterior to posterior, multiband factor = 3. An anatomical T1-weighted image was acquired with MPRAGE sequence; TR = 2300 ms, TE = 2.98 ms, flip angle = 9°, FOV = 256 mm, Slices = 208, voxel size: 1 × 1 × 1 mm (no gap).

SCR measurements

The raw time course of SCR was acquired at 2000 Hz and was low pass filtered (with a cutoff frequency of 1 Hz) offline to eliminate high-frequency noise. For each trial, phasic SCR to the onset of the US (i.e., the truck) appearance on the screen was quantified as a base-to-peak response rise within a 0.9 to 4 s time window considering a slow rise in SCR[64]. The phasic SCR below 0.02 microsiemens was scored as zero[65]. The phasic SCR was transformed as log(SCR + 1) and then scaled to a maximum SCR to the non-US trials within each session[64]. We then extracted SCR for each CS type. Only the trials without US were analysed to selectively examine the conditioned responses to each CS, reflecting anticipatory responses rather than a direct reaction to US presentations. We report the difference (ΔSCR) between CS+ ($CS+_{sequence}$ or $CS+_{element}$) and CS−, with positive values indicating that SCR to CS+ is greater than SCR to CS− (successful conditioning).

Statistical analyses

All statistical analyses were performed with MATLAB Version 9.7 (R2019b) (MathWorks), with built-in functions, other functions available in the MathWorks File Exchange, as well as custom-made code.

Statistical tests between pairs or against chance levels. For all paired tests between conditions or simple one-sample tests, we used Wilcoxon signed-rank test. *P*-values are reported as one-sided or two-sided, depending on the underlying hypothesis as follows. For testing the conditioned SCR, we report one-tailed tests given the expected directionality (i.e., CS+ > CS−). For testing other differences, such as between $CS+_{sequence}$ or $CS+_{element}$ with no directionality assumption, we used two-tailed tests unless otherwise specified. *P*-values were then corrected for multiple comparisons using the Benjamini & Hochberg (1995) procedure for controlling the false discovery rate (FDR), implemented in Matlab using the function *fdr_bh*, available on Mathworks File Exchange. The correction was done within-session unless the test was explicitly between-sessions. For example, in Fig. 2b, the correction was applied to *P*-values from the tests $CS+_{sequence}$ versus 0, $CS+_{element}$ versus 0, and their comparison test $CS+_{sequence}$ versus $CS+_{element}$ in Acquisition, Immediate test and Long-term test separately.

Statistical tests across conditions. We used linear mixed effects (LME) models to analyse relationships between conditions and groups (and other continuous or grouping variables, where needed). All models included random intercepts and random slopes (i.e., the parameters were estimated for each subject). Variables 'session', 'CS_{type}', 'ROI', or 'decoding epoch' were treated as categorical. This meant that, e.g., for the 'session' variable that had 3 levels (Acquisition, Immediate test, and Long-term test), the effects were estimated for Immediate and Long-term tests separately, with Acquisition taken as reference. Models related to information transmission analyses additionally included a second random effect (on the intercept) of decoding accuracy, to account for potential effects that may have been driven by differences in decodability. All models were fitted in Matlab with the native function *fitglme*. Models used for each analysis are described in the relevant supplementary materials related to each figure/analysis.

Region-of-interest (ROI) definitions

Hippocampus (HPC): atlas-based definition (Automated anatomical labelling: AAL)[66], comprising 1475 ± 26 voxels (mean ± SEM). Amygdala: atlas-based definition (AAL), comprising 360 ± 6 voxels (mean ± SEM). Ventromedial prefrontal cortex (VMPFC): definition based on the MNI coordinates [0, 40, −12] with 15 mm radius[30,67] considering previous representative findings on fear memory regulation[68–70], comprising 1248 ± 26 voxels (mean ± SEM). Dorsolateral prefrontal cortex (DLPFC): atlas-based definition (AAL) Brodmann areas 9 and 46, comprising 2417 ± 64 voxels (mean ± SEM).

fMRI scans preprocessing for decoding

The Blood-Oxygen-Level-Dependent (BOLD) fMRI signals in native space were preprocessed in MATLAB Version 9.7 (R2019b) (MathWorks) with SPM12 [http://www.fil.ion.ucl.ac.uk/spm/]. All functional images underwent 3D motion correction. No spatial or temporal smoothing was applied. Rigid-body transformations were performed to align the functional images to the structural image for each subject. Time courses of BOLD signal intensities were extracted from each voxel in each ROI and shifted by 6 s to account for the hemodynamic delay. Time courses were detrended (linear trend), and further z-score normalised for each voxel in each task block to remove baseline differences across blocks. The data samples for computing the binary and multinomial CS_{type} decoders were created by averaging the BOLD signal intensities of each voxel for the entire period of the sound cues epochs (approximately 12 s), in each trial. The same averaging was performed for the data samples from the US anticipatory epoch (which included the ITI as well, in total approximately 6 s) used to test the decoders (see next section below for details on the two types of decoder testing).

Decoding: multivoxel pattern analysis (MVPA)

Two types of MVPA were performed on BOLD activity patterns in HPC and DLPFC. In both cases, decoders were trained on fMRI data corresponding to the sound cues epoch but were then tested on (1) the same epoch from left-out trials and (2) the US anticipatory epoch. Decoders were constructed for each participant separately. The procedure was repeated for each session separately (Acquisition, Immediate test, Long-term test). Since the procedure was very similar to our previous work, the following text is partly quoted and partly adapted from

ref. 71. As with the SCR analyses, only the trials without US were analysed.

We used sparse logistic regression (SLR)[36,37], an algorithm designed to automatically select the most relevant voxels for classification to construct binary or ternary decoders (all binary combinations of CS_{type}, and a ternary decoder classifying each CS vs others). K-fold cross-validation was used for each MVPA by repeatedly subdividing the dataset into a 'training set' and a 'test set' in order to evaluate the predictive power of the trained (fitted) model. The number of folds was fixed and set to 20. In each fold, a random 80% of the data was assigned to the 'training set', while the remaining 20% was allocated to the 'test set'. SLR classification was optimised by using an iterative approach: in each fold of the cross-validation, the feature-selection process was repeated 10 times[36,37]. On each iteration, the selected features (voxels) were removed from the pattern vectors, and only features with unassigned weights were used for the next iteration. At the end of the k-fold cross-validation, the test accuracies were averaged for each iteration across folds, in order to evaluate the accuracy at each iteration. The decoder (number of iterations) with the best accuracy was selected for reporting results of analysis (1) on the same sound cues epoch and to train a new decoder, with all trials from the sound cues epoch, to be then tested (2) on the US anticipatory epoch. The average (±SD) number of voxels selected by the HPC decoders tested on the sound cues epoch was 114 ± 11 (Acquisition), 81 ± 11 (Immediate test), 82 ± 11 (Long-term test). The average (±SD) number of voxels selected by the HPC decoders tested on the US anticipatory epoch was 57 ± 24 (Acquisition), 35 ± 17 (Immediate test), 36 ± 20 (Long-term test). The average (±SD) number of voxels selected by the DLPFC decoders tested on the sound cues epoch was 111 ± 12 (Acquisition), 83 ± 11 (Immediate test), 86 ± 15 (Long-term test). The average (±SD) number of voxels selected by the DLPFC decoders tested on the US anticipatory epoch was 58 ± 28 (Acquisition), 38 ± 18 (Immediate test), 40 ± 22 (Long-term test). Note that the difference in the number of selected voxels between sound cues epoch and US anticipatory epoch decoders simply reflects the fact that the former was trained 20 times (once for each cross-validation run) while the latter was trained once on all the sound cues epoch data before being tested on US anticipatory epoch data.

Information transmission analysis

Here, we used the same data prepared for the decoding analyses. In short, we used the decoders trained on the sound cues epoch to infer the CS representations likelihood during the US anticipatory epoch and used the activity patterns from the US anticipatory epoch to predict the CS likelihood. Information transmission analysis was based on a sparse linear regression to predict the CS_{type} representation likelihood in a seed area (e.g., HPC or DLPFC), from the activity patterns in a target area (e.g., VMPFC-Amygdala). A predicted value was obtained as the linearly weighted sum of the voxel activities in each area. Prediction accuracy was defined as the correlation coefficient between predicted CS likelihood and actual likelihood and was evaluated by leave-one-trial-out cross-validation procedure. That is, for each left-out trial, we obtained a predicted likelihood. Then, taking all trials, we correlated original likelihoods and predicted likelihoods. Correlation coefficients were then Fisher z-transformed before averaging across participants. Coefficients higher than 0 mean information (of the CS representations) are shared between seed and target areas.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data generated in this study and reported in the paper's figures and supplementary information are provided in the Supplementary Information/Source Data file and at http://github.com/AureCor/Cortese_et_al_2024. Source data are provided with this paper.

Code availability

The code to perform decoding analysis is freely available at https://bicr.atr.jp/~oyamashi/SLR_WEB.html. The code to generate main results and figures is available at http://github.com/AureCor/Cortese_et_al_2024.

References

1. Brewin, C. R. Episodic memory, perceptual memory, and their interaction: foundations for a theory of posttraumatic stress disorder. *Psychol. Bull.* **140**, 69–97 (2014).
2. van der Kolk, B. A. & Fisler, R. Dissociation and the fragmentary nature of traumatic memories: overview and exploratory study. *J. Trauma. Stress* **8**, 505–525 (1995).
3. Lonsdorf, T. B. et al. Don't fear 'fear conditioning': methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. *Neurosci. Biobehav. Rev.* **77**, 247–285 (2017).
4. Phelps, E. A. & LeDoux, J. E. Contributions of the amygdala to emotion processing: from animal models to human behavior. *Neuron* **48**, 175–187 (2005).
5. Ehlers, A. & Clark, D. M. A cognitive model of posttraumatic stress disorder. *Behav. Res. Ther.* **38**, 319–345 (2000).
6. Peri, T., Ben-Shakhar, G., Orr, S. P. & Shalev, A. Y. Psychophysiologic assessment of aversive conditioning in posttraumatic stress disorder. *Biol. Psychiatry* **47**, 512–519 (2000).
7. Ehlers, A., Hackmann, A. & Michael, T. Intrusive re-experiencing in post-traumatic stress disorder: phenomenology, theory, and therapy. *Memory* **12**, 403–415 (2004).
8. Bedard-Gilligan, M. & Zoellner, L. A. Dissociation and memory fragmentation in post-traumatic stress disorder: an evaluation of the dissociative encoding hypothesis. *Memory* **20**, 277–299 (2012).
9. Rubin, D. C. et al. Participant, rater, and computer measures of coherence in posttraumatic stress disorder. *J. Abnorm. Psychol.* **125**, 11–25 (2016).
10. Brewin, C. R. Coherence, disorganization, and fragmentation in traumatic memory reconsidered: a response to Rubin et al. (2016). *J. Abnorm. Psychol.* **125**, 1011–1017 (2016).
11. Tulving, E. Episodic and semantic memory. *Organization of memory* (Academic Press, 1972).
12. Clayton, N. S. & Dickinson, A. Episodic-like memory during cache recovery by scrub jays. *Nature* **395**, 272–274 (1998).
13. Dunsmoor, J. E. & Kroes, M. C. W. Episodic memory and Pavlovian conditioning: ships passing in the night. *Curr. Opin. Behav. Sci.* **26**, 32–39 (2019).
14. Starita, F., Kroes, M. C. W., Davachi, L., Phelps, E. A. & Dunsmoor, J. E. Threat learning promotes generalization of episodic memory. *J. Exp. Psychol. Gen.* **148**, 1426–1434 (2019).
15. Gagnon, S. A. & Wagner, A. D. Acute stress and episodic memory retrieval: neurobiological mechanisms and behavioral consequences. *Ann. N. Y. Acad. Sci.* **1369**, 55–75 (2016).
16. Brewin, C. R. & Ehlers, A. Posttraumatic stress disorder in *Handbook of Human Memory: Foundations and Applications* (ed. Wagner, M. K.) (Oxford Univ. Press, 2023).
17. Davachi, L. & DuBrow, S. How the hippocampus preserves order: the role of prediction and context. *Trends Cogn. Sci.* **19**, 92–99 (2015).
18. Naya, Y., Chen, H., Yang, C. & Suzuki, W. A. Contributions of primate prefrontal cortex and medial temporal lobe to temporal-order memory. *Proc. Natl Acad. Sci. USA* **114**, 13555–13560 (2017).
19. Long, N. M. & Kahana, M. J. Hippocampal contributions to serial-order memory. *Hippocampus* **29**, 252–259 (2019).

20. Fortin, N. J., Agster, K. L. & Eichenbaum, H. B. Critical role of the hippocampus in memory for sequences of events. *Nat. Neurosci.* **5**, 458–462 (2002).
21. Neubert, F.-X., Mars, R. B., Sallet, J. & Rushworth, M. F. S. Connectivity reveals relationship of brain areas for reward-guided learning and decision making in human and monkey frontal cortex. *Proc. Natl Acad. Sci. USA* **112**, E2695–E2704 (2015).
22. Euston, D. R., Gruber, A. J. & McNaughton, B. L. The role of medial prefrontal cortex in memory and decision making. *Neuron* **76**, 1057–1070 (2012).
23. Hartley, C. A. & Phelps, E. A. Changing fear: the neurocircuitry of emotion regulation. *Neuropsychopharmacology* **35**, 136–146 (2010).
24. Visser, R. M., Bathelt, J., Scholte, H. S. & Kindt, M. Robust BOLD responses to faces but not to conditioned threat: challenging the amygdala's reputation in human fear and extinction learning. *J. Neurosci.* **41**, 10278–10292 (2021).
25. Wen, Z. et al. Temporally and anatomically specific contributions of the human amygdala to threat and safety learning. *Proc. Natl Acad. Sci. USA* **119**, e2204066119 (2022).
26. Fullana, M. A. et al. Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. *Mol. Psychiatry* **21**, 500–508 (2016).
27. Koenigs, M. & Grafman, J. Posttraumatic stress disorder: the role of medial prefrontal cortex and amygdala. *Neuroscientist* **15**, 540–548 (2009).
28. Chiba, T. et al. A reciprocal inhibition model of alternations between under-/overemotional modulatory states in patients with PTSD. *Mol. Psychiatry* **26**, 5023–5039 (2021).
29. Yehuda, R. et al. Post-traumatic stress disorder. *Nat. Rev. Dis. Prim.* **1**, 15057 (2015).
30. Koizumi, A. et al. Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nat. Hum. Behav.* **1**, 0006 (2016).
31. Cortese, A., Amano, K., Koizumi, A., Kawato, M. & Lau, H. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat. Commun.* **7**, 13669 (2016).
32. Shibata, K., Watanabe, T., Sasaki, Y. & Kawato, M. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* **334**, 1413–1415 (2011).
33. Weger, M. & Sandi, C. High anxiety trait: a vulnerable phenotype for stress-induced depression. *Neurosci. Biobehav. Rev.* **87**, 27–37 (2018).
34. Kok, L. et al. Trait anxiety mediates the effect of stress exposure on post-traumatic stress disorder and depression risk in cardiac surgery patients. *J. Affect. Disord.* **206**, 216–223 (2016).
35. Kansai Explorer. Yotsubashisuji—May 2007. *Wikimedia Foundation*. https://ja.wikipedia.org/wiki/ファイル:四つ橋筋江戸堀1交差点.JPG (2007).
36. Yamashita, O., Sato, M.-A., Yoshioka, T., Tong, F. & Kamitani, Y. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *Neuroimage* **42**, 1414–1429 (2008).
37. Hirose, S., Nambu, I. & Naito, E. Iterative Sparse Logistic Regression (iSLR): a new ensemble pattern classification method for fMRI decoding. *Neurosci. Res.* **71**, e97 (2011).
38. Roy, D. S. et al. Brain-wide mapping reveals that engrams for a single memory are distributed across multiple brain regions. *Nat. Commun.* **13**, 1799 (2022).
39. Barbas, H. Connections underlying the synthesis of cognition, memory, and emotion in primate prefrontal cortices. *Brain Res. Bull.* **52**, 319–330 (2000).
40. Kishi, T., Tsumori, T., Yokota, S. & Yasui, Y. Topographical projection from the hippocampal formation to the amygdala: a combined anterograde and retrograde tracing study in the rat. *J. Comp. Neurol.* **496**, 349–368 (2006).
41. Dunsmoor, J. E. et al. Event segmentation protects emotional memories from competing experiences encoded close in time. *Nat. Hum. Behav.* **2**, 291–299 (2018).
42. Wang, J., Tambini, A. & Lapate, R. C. The tie that binds: temporal coding and adaptive emotion. *Trends Cogn. Sci.* **26**, 1103–1118 (2022).
43. Braun, E. K., Wimmer, G. E. & Shohamy, D. Retroactive and graded prioritization of memory by reward. *Nat. Commun.* **9**, 4886 (2018).
44. Foster, D. J. & Wilson, M. A. Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature* **440**, 680–683 (2006).
45. Tonegawa, S., Liu, X., Ramirez, S. & Redondo, R. Memory engram cells have come of age. *Neuron* **87**, 918–931 (2015).
46. Rescorla, R. A. & Heth, C. D. Reinstatement of fear to an extinguished conditioned stimulus. *J. Exp. Psychol. Anim. Behav. Process.* **1**, 88–96 (1975).
47. Barbey, A. K., Koenigs, M. & Grafman, J. Dorsolateral prefrontal contributions to human working memory. *Cortex* **49**, 1195–1205 (2013).
48. Wang, J. & Barbas, H. Specificity of primate amygdalar pathways to hippocampus. *J. Neurosci.* **38**, 10019–10041 (2018).
49. Barry, D. N. & Maguire, E. A. Remote memory and the hippocampus: a constructive critique. *Trends Cogn. Sci.* **23**, 128–142 (2019).
50. Gewirtz, J. C. & Davis, M. Using Pavlovian higher-order conditioning paradigms to investigate the neural substrates of emotional learning and memory. *Learn. Mem.* **7**, 257–266 (2000).
51. Alvarez, R. P., Biggs, A., Chen, G., Pine, D. S. & Grillon, C. Contextual fear conditioning in humans: cortical-hippocampal and amygdala contributions. *J. Neurosci.* **28**, 6211–6219 (2008).
52. Knight, D. C., Cheng, D. T., Smith, C. N., Stein, E. A. & Helmstetter, F. J. Neural substrates mediating human delay and trace fear conditioning. *J. Neurosci.* **24**, 218–228 (2004).
53. Seymour, B. et al. Temporal difference models describe higher-order learning in humans. *Nature* **429**, 664–667 (2004).
54. Deuker, L., Bellmund, J. L., Navarro Schröder, T. & Doeller, C. F. An event map of memory space in the hippocampus. *Elife* **5**, e16534 (2016).
55. Deisseroth, K. Optogenetics: 10 years of microbial opsins in neuroscience. *Nat. Neurosci.* **18**, 1213–1225 (2015).
56. Clark, R. E. & Squire, L. R. Classical conditioning and brain systems: the role of awareness. *Science* **280**, 77–81 (1998).
57. Taschereau-Dumouchel, V., Michel, M., Lau, H., Hofmann, S. G. & LeDoux, J. E. Putting the 'mental' back in 'mental disorders': a perspective from research on fear and anxiety. *Mol. Psychiatry* **27**, 1322–1330 (2022).
58. Hartley, C. A., Fischl, B. & Phelps, E. A. Brain structure correlates of individual differences in the acquisition and inhibition of conditioned fear. *Cereb. Cortex* **21**, 1954–1962 (2011).
59. Huff, N. C., Hernandez, J. A., Blanding, N. Q. & LaBar, K. S. Delayed extinction attenuates conditioned fear renewal and spontaneous recovery in humans. *Behav. Neurosci.* **123**, 834–843 (2009).
60. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G. Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav. Res. Methods* **39**, 175–191 (2007).
61. Lau, J. Y. F. et al. Fear conditioning in adolescents with anxiety disorders: results from a novel experimental paradigm. *J. Am. Acad. Child Adolesc. Psychiatry* **47**, 94–102 (2008).
62. Mobbs, D. et al. When fear is near: threat imminence elicits prefrontal-periaqueductal gray shifts in humans. *Science* **317**, 1079–1083 (2007).
63. Hidano, T., Fukuhara, M., Iwawaki, M., Soga, S., & Spielberger, C. D. *State Trait Anxiety Inventory (Form JYZ) Test Manual (Japanese Adaptation of STAI)* (Jitsumu Kyouiku, 2000).

64. Gerlicher, A. M. V., Tüscher, O. & Kalisch, R. Dopamine-dependent prefrontal reactivations explain long-term benefit of fear extinction. *Nat. Commun.* **9**, 4294 (2018).
65. Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M.-H. & Phelps, E. A. Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proc. Natl Acad. Sci. USA* **110**, 20040–20045 (2013).
66. Tzourio-Mazoyer, N. et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* **15**, 273–289 (2002).
67. Lonsdorf, T. B., Haaker, J. & Kalisch, R. Long-term expression of human contextual fear and extinction memories involves amygdala, hippocampus and ventromedial prefrontal cortex: a reinstatement study in two independent samples. *Soc. Cogn. Affect. Neurosci.* **9**, 1973–1983 (2014).
68. Kalisch, R. et al. Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *J. Neurosci.* **26**, 9503–9511 (2006).
69. Phelps, E. A., Delgado, M. R., Nearing, K. I. & LeDoux, J. E. Extinction learning in humans: role of the amygdala and vmPFC. *Neuron* **43**, 897–905 (2004).
70. Milad, M. R. et al. Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biol. Psychiatry* **62**, 446–454 (2007).
71. Cortese, A., Lau, H. & Kawato, M. Unconscious reinforcement learning of hidden brain states supported by confidence. *Nat. Commun.* **11**, 4429 (2020).

Acknowledgements

We thank Ben Seymour, Vincent Taschereau-Dumouchel, and Mitsuo Kawato for their helpful comments on earlier versions of the manuscript. We thank Seyrin Lee and Wen Wen for their support in conducting experiments and Shuntaro Sasai for his technical advice on the sound presentations. We appreciate the support of the Center for Evolutionary Cognitive Sciences at the University of Tokyo. The study was supported by the following grants: Moonshot (20343198) and Presto (18068712) from the Japan Science and Technology Agency (JST) (A.K.). Grant-in-Aid for Transformative Research Area (A) (23H04833) and Scientific Research (B) programme (18H02714 and 22H01111) from the Japan Society for the Promotion of Science (JSPS) (A.K.). Ikegaya Brain-AI Hybrid ERATO (JPMJER1801) from the Japan Science and Technology Agency (A.C.). Grant-in-Aid for Transformative Research Area (A) (JP22H05156) from JSPS (A.C.). The Innovative Science and Technology Initiative for Security (JPJ004596), ATLA, Japan (A.C.).

Author contributions

A.K., A.C., H.I. and N.K. designed the study; R.O., A.K. and M.A.G. collected the data; A.C. and A.K. analysed the data; A.C. created the visualisations; A.K. and A.C. drafted the manuscript, and all authors revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52733-4.

Correspondence and requests for materials should be addressed to Aurelio Cortese, Hiroshi Imamizu or Ai Koizumi.

Peer review information *Nature Communications* thanks Renée Visser and the other anonymous reviewer(s) for their contribution to the peer review of this work. A peer review file is available.