

Sections of a Stage 1 Registered Report

PTSD therapy with fMRI-decoded neurofeedback bypassing conscious exposure: a randomized, double-blind, placebo-controlled study

Toshinori Chiba^{1,2}, Kentaro Ide^{1,3,4}, Jessica E. Taylor¹, Aurelio Cortese¹, Takatomi Kubo^{1,5}, Keiichiro Nishida⁶, Shuken Boku⁷, Hiroyuki Toda⁸, Miyako Shirakawa⁴, Kaoru Amano⁹, Hakwan Lau¹⁰, Mitsuo Kawato^{1,11}, Ben Seymour^{12, 13}, Ai Koizumi^{1,14*}, and Tetsufumi Kanazawa^{6,15}

¹ The Department of Decoded Neurofeedback, Computational Neuroscience Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan, ² The Department of Psychiatry, Self-Defense Forces Hanshin Hospital, Kawanishi, Japan, ³ Seven-Mental Clinic, Osaka, Japan, ⁴ Flower of Light Clinic for Mind and Body, Tokyo, Japan, ⁵ Graduate School of Science and Technology, Nara Institute of Science and Technology, Nara, Japan, ⁶ The Department of Neuropsychiatry, Osaka Medical and Pharmaceutical University, Osaka, Japan, ⁷ Department of Psychiatry, Niigata University Graduate School of Medical and Dental Sciences, Niigata, Japan, ⁸ The Department of Psychiatry, National Defense Medical College, Tokorozawa, Japan, ⁹ Graduate School of Information Science and Technology, The University of Tokyo, Tokyo, Japan, ¹⁰ RIKEN Center for Brain Science, Wako, Japan, ¹¹ XNef, Kyoto, Japan, ¹² Department of Engineering Science, Institute of Biomedical Engineering, University of Oxford, Old Road Campus Research Building, Oxford, UK, ¹³ The Department of Neural Computation for Decision-making, Cognitive Mechanism Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan, ¹⁴ Sony Computer Science Laboratories, Inc., Tokyo, Japan, ¹⁵ The Department of Neuropsychiatry, Osaka Medical and Pharmaceutical University, Osaka, Japan

* Corresponding authors: Ai Koizumi (bellkoizumi@gmail.com).

Abstract

Exposure-based therapy is effective for alleviating fear among patients with post-traumatic stress disorder (PTSD). Nonetheless, because the therapy itself induces fear, patients sometimes abandon treatment prematurely. One emerging alternative therapy employs real-time fMRI-Decoded Neurofeedback (DecNef). It aims to alleviate excessive physiological responses to threat, while bypassing conscious exposure procedure. With DecNef, neural activation patterns for feared cues are first identified. Then these patterns are subsequently induced through feedback without the patient's awareness of the cues. However, evidence for DecNef's efficacy is so far limited to laboratory settings and small numbers of patients. In the proposed study, we will test the effectiveness of DecNef for PTSD patients in a double-blind, randomized, placebo-controlled study with an improved procedure to decode nonconscious neural activities in response to trauma-related cues. To minimize patient variability, we will employ a crossover design with a six-month interval, considering the compatibility of DecNef to such a design. We will further examine the supposed implicit and stress-free nature of DecNef treatment. The mechanisms of DecNef will be examined with neuroimaging and computational approaches. If successful, this study may offer a less subjectively unpleasant new avenue for PTSD therapy.

Introduction

Post-traumatic stress disorder (PTSD) is a debilitating disease that occurs following life-threatening traumatic events, and affects around 6.1 to 9.2% of the population sometime in life¹⁻⁴. In typical clinical therapeutic approaches for PTSD, patients are repeatedly exposed to trauma-related stimuli. Accordingly, exposure-based therapy is often distressing for PTSD patients. In some cases, this causes them to re-experience the traumatic memories anew⁵, which can exacerbate their symptoms⁶. The reported dropout rate of exposure-based therapy is sometimes as high as 20-30% within a 2-month treatment period^{7,8}, with some studies reporting even higher rates⁹. This rate is relatively higher than that including other psychological therapies for PTSD such as Present Centred Therapy, which have a dropout rate of around 16% (95% Confidence Intervals [CI] = 14-18%)¹⁰.

One recently proposed strategy to reduce PTSD symptoms without distress is to employ Decoded Neurofeedback (DecNef). DecNef estimates the likelihood that a target neural representation is present in a predefined brain region at a particular moment, through real-time decoding of functional magnetic resonance imaging (fMRI) signals¹¹⁻¹⁵. By immediately presenting monetary reward feedback to the participants when the target neural representation is detected, this procedure reinforces participants to spontaneously induce the target neural activation patterns¹⁶. Importantly, participants typically remain unaware of the subjective content of the induced neural activation patterns¹⁷⁻¹⁹. As a result they do not consciously experience the relevant content and show no elevated physiological responses during the procedure^{17,20}.

Previously, we leveraged the nonconscious nature of DecNef to reduce threat responses toward conditioned stimuli or feared animals at both physiological and neural levels^{17,20,21}. In those studies, DecNef was used to reinforce participants to unconsciously induce the neural representations for feared stimuli with monetary reward. However, the evidence so far has been limited to healthy participants²⁰, subclinical participants¹⁷, participants with specific phobias²¹, and a small number (N = 4) of PTSD patients (in a part of our pilot study²²). Here, we propose to directly investigate the applicability of DecNef to alleviate PTSD clinical symptoms.

Specifically, we will test two primary and three secondary hypotheses related to the therapeutic effects of DecNef (Hypotheses **Therapeutics effect, H_{TPE}**). As primary hypotheses, we will test whether DecNef yields greater pre-to-post symptom reduction than control DecNef which induces fear-irrelevant stimuli (**H_{TPE} 1**). Second, we will assess the magnitude of this reduction, and test if DecNef can lead to a change in symptoms that exceeds a clinically meaningful threshold, defined as 10-point reduction on the clinician-administered PTSD scale (CAPS)^{8,23} (**H_{TPE} 2**). Additionally, our secondary hypotheses are that: DecNef's therapeutic effects on PTSD symptoms persist for 2 months relative to the control DecNef (**H_{TPE} 3a**), and the magnitude of this long-term effect passes the clinically meaningful threshold (**H_{TPE} 3b**). We also hypothesize that DecNef effects will lead to alteration of physiological threat responses (including skin conductance response, SCR, and amygdala responses: **H_{TPE} 4a**), which may covary with symptom reductions (**H_{TPE} 4b**). Finally, we hypothesize that these effects will be specific to PTSD symptoms rather than to symptoms related to comorbid conditions (**H_{TPE} 5**).

Besides the therapeutic effects, three challenges need to be met to develop DecNef procedures for PTSD patients. First, induction of trauma-related neural representations needs to remain nonconscious and should not induce excessive distress for PTSD patients, in order to minimize dropout¹⁰. Although healthy participants reported little distress during DecNef^{17,20}, trauma-related neural representations may more readily evoke unpleasant subjective experiences in patients²⁴. We will directly test the following four secondary hypotheses on **Distress during NeuroFeedback (H_{DNF})**: First, we hypothesize that participants will remain unaware of the DecNef-induced trauma-related content (**H_{DNF} 1**). That is, when asked subsequently what is their strategy for earning rewards during the DecNef procedure, fewer than i) 10%—a threshold for very frequent side effects, and ii) 50%—half—of them will report using trauma-related imagery. Also, fewer than i) 10% and ii) 50% of them will correctly guess that the relevant brain representation that leads to reward during the DecNef procedure is in fact trauma-related. Further, we hypothesize that participants will not be able to accurately guess the order of the two neurofeedback conditions (experimental DecNef vs control) in a two-alternative-forced-choice question, resulting at chance level accuracy (50%) in the whole patient

group ($H_{DNF} 2$). We also hypothesize that DecNef training will not lead to subjective distress levels (SUDS) higher than both a tolerable threshold level at 50 points²⁵, and pre-training baseline levels ($H_{DNF} 3$). Finally, we hypothesize that dropout rates will be lower than 14-18%, which is the 95% CI in the pooled dropout rate in the meta-analysis of randomized controlled trial (RCTs) of PTSD psychotherapy in general¹⁰ ($H_{DNF} 4$). Evaluations of those hypotheses will examine DecNef's feasibility as low-distress PTSD treatment.

The second challenge is to minimize conscious exposure to trauma-related stimuli, not only during DecNef training, but throughout the entire DecNef procedure including its preparatory decoder construction sessions. In the standard DecNef procedure^{17,20}, a participant is repeatedly exposed to visual stimuli to obtain neural activity patterns to build a decoder prior to the DecNef training. The decoder then allows the real-time detection of the neural activation patterns representing the target visual stimuli during DecNef. The explicit decoder construction procedure was not problematic with healthy participants who were fear-conditioned to the visual stimuli only after constructing a decoder. However, this approach is less favorable for PTSD patients, because they are already fearful of the trauma-related stimuli before the decoder construction.

Two potential approaches can bypass conscious presentation of feared stimuli during decoder construction: hyperalignment and visual pattern masking. Using hyperalignment, Taschereau-Dumouchel et al. inferred fear-relevant representations for participants with subclinical levels of phobic symptoms from the neural activity patterns of “surrogate” participants^{17,26,27}. By having both sets of participants view a large set of stimuli (3,600 images of 40 different animals and inanimate objects), a decoder for feared stimuli can be aligned from the neural activation patterns of surrogate participants to the participants with phobia¹⁷. Hyperalignment can thus build a decoder of feared stimuli for participants with phobia, even if participants with phobia do not directly view the feared stimuli. This approach may be appealing for certain clinical conditions such as specific phobias of animals¹⁷ where a target representation is readily defined by other fear-irrelevant representations unlike PTSD.

In our pilot study conducted on patients with PTSD²², we instead used continuous flash suppression (CFS), which is a form of a binocular visual pattern masking, to render trauma-related images less visible²⁸ during decoder construction. Specifically, a trauma-related image was presented to participants' non-dominant eyes while salient flashing Mondrian patches were presented to their dominant eyes. Our previous study achieved clinically significant improvement in PTSD symptom by using CFS to build a decoder from the neural activation patterns in the superior temporal sulcus (STS) for trauma-related face stimuli²². Considering this, we will employ CFS and directly test our hypotheses on distress during decoder construction (Hypotheses on **Distress during Decoder Construction, H_{DDC}**). These hypotheses are that: Participants will remain unaware of trauma-related content during the decoder construction with CFS (detailed in Supplementary), and SUDS ratings for decoder construction sessions will remain below both a tolerable threshold level at 50²⁵ and their baseline level from before the decoder construction procedure (**$H_{DDC} 1$**). Together with the assessments of distress during DecNef training (Hypotheses **H_{DNE}**), evaluations of Hypotheses **H_{DDC}** will determine whether unconscious and low-distress features are maintained throughout the DecNef protocol, ensuring clinical feasibility.

Finally, the third challenge is to resolve the two candidate mechanisms underlying the fear alleviation effects of DecNef. One candidate mechanism is that DecNef recruits a similar neural circuits as exposure-based therapies by repetitively inducing neural representations of feared stimuli, where the ventromedial prefrontal cortex (vmPFC) suppresses the amygdala to extinguish fear^{29,30}. However, our previous study showed that the fear reduction effect was larger among individuals who had less vmPFC involvement during DecNef training²⁰. With a neuroimaging approach, this study suggested an alternative mechanism involving the reward circuits, potentially supporting counter-conditioning of feared stimuli with rewarding feedback³¹. To dissociate the two candidate mechanisms, a preliminary investigation²² adapted a computational approach, applying variants of Rescorla-Wagner models³² to the data obtained from non-clinical populations in previous DecNef studies^{17,20}. While the exposure-based model outperformed the counter-conditioning model, the results remained inconclusive²² and may differ among PTSD patients³³. In this proposed study, we will test

our hypothesis on Neural Mechanisms (Hypotheses H_{NM}) by applying the same neuroimaging²⁰ and computational²² approaches to the data obtained from PTSD patients. Specifically, we will conduct two neuroimaging analyses to test whether DecNef relies on unique neural mechanisms that rely on the reward circuits (caudate and ventral striatum) rather than the fear extinction circuits (vmPFC). First, multivoxel decoding analysis will be used to evaluate transmission of trauma-related neural representations from STS, the target region for DecNef neural induction, to reward versus fear circuits during DecNef (H_{NM} 1a). Second, a functional connectivity analysis will evaluate the connectivity of STS with the reward and fear circuits (H_{NM} 1b). In Hypotheses H_{NM} 1a and H_{NM} 1b, we predict that the engagement of reward circuits during DecNef, compared with fear circuits, will correlate stronger with post-DecNef symptom alleviation, supporting the mechanism of counter-conditioning over exposure/extinction. With computational approaches, we will further examine whether the model based on the counter-conditioning computations will better explain the PTSD symptom reductions compared to the model based on the exposure/extinction-based computations (H_{NM} 2). Such results would provide evidence for reward-circuit-based therapeutic effects of DecNef, suggesting a mechanism distinct from traditional fear-circuit-based therapies.

We will confine participants to female patients who have developed PTSD from male violence. Females generally have twice as great a risk of developing PTSD as males³⁴, and victims of intentional trauma such as male violence show worse prognosis than those of non-intentional trauma, such as natural disasters^{35–37}. Enrolment of female survivors of male violence would enhance homogeneity among participants to facilitate investigations of DecNef effects and mechanisms, as PTSD patients are otherwise typically heterogeneous due to diversity of trauma types^{35,37} and sex differences^{35,38}. In addition, female victims of male violence share a common small set of stimuli associated with traumatic experiences, especially angry male faces. As DecNef typically induces neural representations of a single target stimulus, this patient group with a specific trauma-associated stimulus may especially benefit from DecNef training.

To further minimize across participants variability, we will use a crossover design where participants will undergo experimental and control DecNef sessions with a 6-month interval between.

A previous crossover DecNef experiment has successfully up- and down-regulated subjective confidence within single subjects across two DecNef sessions delivered with only a one week interval between³⁹, which indicates that crossover design is appropriate for DecNef procedures. Our successful pilot PTSD-DecNef study²² also partially employed the crossover design (see Supplementary Table 1). Considering that effects of neurofeedback could last for 2-5 months^{13,40,41}, we will minimize the potential anterograde learning interference effects from the first session to the second session by extending the interval to 6 months. As PTSD symptoms typically persist longer than 6 months⁸, we expect that effects of the second DecNef session can be assessed without a floor effect even after such a long interval.

Lastly, we adopted a co-design approach where participants with PTSD from our pilot study contributed to development and refinement of the study design (<https://www.medsci.ox.ac.uk/research/patient-and-public-involvement>). This pilot data yielded encouraging results, specifically that (a) the proposed decoder could be constructed with minimal distress to participants with PTSD, and (b) PTSD symptoms could be reduced via our DecNef technique. Leveraging the co-design approach, we confirmed that the scanner environment, rather than the presented stimuli, caused distress to participants. To address this, we decided to allow participants to bring a trusted companion and compensate their travel expenses and efforts. In this pre-registered study, the participants from the pilot and another 58 participants will contribute their insight and experience of the study to help us map a subsequent pathway to feasible clinical applications.

Overall, in this proposed study, we will validate the Therapeutic effects (**H_{TPE} 1-5**) of DecNef on PTSD in a randomized, double-blind, placebo-controlled, crossover design with a 6-month interval period with 58 participants with PTSD. We will evaluate low-distress characteristics of DecNef during neurofeedback (**H_{DNF} 1-4**) and its preparatory decoder construction sessions (**H_{DDC} 1**). We will further explore the Neural Mechanisms (**H_{NM} 1, 2**) regarding how DecNef reduces PTSD symptoms in neuroimaging and computational approaches.

Table 1. Design Table

Question	Hypothesis	Sampling plan (power analysis)	Analysis Plan	Interpretation given to different outcomes
Primary Outcomes				
<p>[TPE: Therapeutic effect]</p> <p>Does DecNef provide a clinically meaningful therapeutic effect for PTSD?</p>	<p>(H_{TPE 1}) Participants with PTSD will show greater reduction in CAPS scores in the experimental condition than the control condition.</p>	<p>According to the power analysis, a sample size of 57 would be sufficient to achieve 95% power to detect an effect size of hedge's $g = 0.684$ (see Sample Size section for this choice).</p> <p>We will enroll 58 participants (minimum number for counterbalancing) to complete all sessions.</p> <p>We will conduct interim analyses with predefined significance thresholds: the first analysis at 10% of the data (i.e., 6 participants) with $p < 0.0005$, and the second analysis at 50% of the data (i.e., 29 participants) with $p < 0.015$. The trial will be terminated early if these thresholds are met for both of primary outcomes (i.e., H_{TPE 1} and 2).</p>	<p>Mixed-effect model analyses:</p> <p>$\Delta\text{endpoint} = 1 + \text{condition} + \text{order} + \text{condition}*\text{order} + (1 \text{patient})$</p> <p>Where the $\Delta\text{endpoint}$ is the reduction in the CAPS score from pre-training test (-1 day) to post-training test (+1 week).</p> <p>The condition is coded as 1/0 for the experimental/control condition. The order is coded as 1/0 if the experimental/control condition is delivered first.</p>	<p>Positive/negative and statistically significant effects of the condition will suggest that the experimental/control condition of DecNef provides a therapeutic effect for PTSD.</p> <p>See null hypothesis policy* of interpretation in case the null hypothesis cannot be rejected. The smallest effect size of interest (SESOI) for the equivalent test is defined as $[-0.1368, 0.1368]$, 20% of hedge's $g = 0.684$.</p>

	<p>(H_{TPE 2}) Participants with PTSD will show reduction in CAPS scores larger than 10-points in the experimental condition.</p>	<p>According to the power analysis based on our pilot study, a sample size of five would provide 95% power to detect an effect size of Hedge's $g = 2.35$. However, for consistency and to ensure robust counterbalancing, we will enroll 58 participants, matching the sample size determined for H_{TPE 1}. (see Sample Size section).</p> <p>We will conduct interim analyses with predefined significance thresholds: the first analysis at 10% of the data (i.e., 6 participants) with $p < 0.0005$, and the second analysis at 50% of the data (i.e., 29 participants) with $p < 0.015$. The trial will be terminated early if these thresholds are met for both of primary outcomes (i.e., H_{TPE 1} and 2).</p>	<p>We will apply one sample t-tests to compare reductions in CAPS from pre-training test (-1 day) to post-training test (+1 week) against 10-points.</p>	<p>If reduction in CAPS is statistically larger than 10-points, we will consider DecNef to provide clinically meaningful therapeutic effect for PTSD. If reduction in CAPS is statistically smaller than 10-points, we will consider it to be clinically meaningless for PTSD.</p> <p>See null hypothesis policy*. The SESOI for the equivalent test is defined as [12, 8], considering 20% of 10-points.</p>
Secondary Outcomes				
<p>[TPE: Therapeutic effect]</p> <p>Does the therapeutic effect of DecNef H_{TPE}</p>	<p>(H_{TPE 3a}) Participants with PTSD will show greater reduction in CAPS scores in the experimental condition than the control</p>	<p>We will use the same dataset collected for testing H_{TPE 1}. Interim analyses will be conducted once half the data (N=29) is collected to assess if the planned</p>	<p>Mixed-effect model analyses:</p> $\Delta \text{endpoint} = 1 + \text{condition} + \text{order} + \text{condition} * \text{order} + (1 \text{patient})$ <p>Where the endpoint is the reduction in the CAPS score from pre-training test (-1 day) to post-training test (+2</p>	<p>Positive/negative and statistically significant effects of the condition will suggest that the therapeutic effect of the experimental/control condition of DecNef lasts in the long-term (i.e. 2 months post intervention).</p> <p>See null hypothesis policy*. The SESOI will be</p>

<p>3) last in the long-term, H_{TPE 4a} generalize to threat responses, H_{TPE 4b} covary with threat responses, and H_{TPE 5} is specific to PTSD?</p>	<p>condition after the 2 months post intervention.</p>	<p>sample size of 58 provides 95% power to detect the calculated effect size. The hypothesis will only be tested if the required sample size is confirmed to be 58 or fewer.</p>	<p>months).</p>	<p>set at 20% of the calculated effect size determined during the interim analysis.</p>
	<p>(H_{TPE 3b}) Participants with PTSD will show reduction in CAPS scores larger than 10-points in the experimental condition after the 2 months post intervention.</p>	<p>Same as above.</p>	<p>We will apply one sample t-tests to compare reductions in CAPS from pre-training test (-1 day) to post-training test (+2 months) against 10-points.</p>	<p>If reduction in CAPS is statistically larger than 10-points, we will consider clinically meaningful therapeutic effects to last in the long term (i.e. 2 months post intervention) for PTSD. If reduction in CAPS is statistically smaller than 10-points, we will consider it does not last in the long term (i.e. 2 months post intervention).</p> <p>See null hypothesis policy*. The SESOI for the equivalent test is defined as [12, 8], 20% of 10-points.</p>
	<p>(H_{TPE 4a}) Participants with PTSD will show greater pre-to-post intervention changes in i) SCRs, and ii) amygdala reactivity in the experimental condition relative to the control condition.</p>	<p>Same as above.</p>	<p>Mixed-effect model analyses: $\Delta\text{endpoint} = 1 + \text{condition} + \text{order} + \text{condition}*\text{order} + (1 \text{patient})$ Where the $\Delta\text{endpoint}$ is the changes in i) SCRs, and ii) amygdala reactivity from pre-training test (-1 day) to post-training test (+1 day).</p>	<p>Hypothesis will be supported if condition or condition*order is statistically significant.</p> <p>See null hypothesis policy*. The SESOI will be set at 20% of the calculated effect size determined during the interim analysis.</p>
	<p>(H_{TPE 4b}) Variations in i) SCRs and ii) amygdala reactivity to angry faces will be associated with variations in PTSD symptoms.</p>	<p>Same as above.</p>	<p>We will apply a mixed effect model: $\text{Fear} = 1 + \text{Sev} + (\text{Sev} \text{patient})$ Where the Fear is i) SCRs and ii) amygdala reactivity and the Sev is the score of IES-R.</p>	<p>Hypothesis will be supported if Sev is statistically significant.</p> <p>See null hypothesis policy*. The SESOI will be set at 20% of the calculated effect size determined during the interim analysis.</p>

	(H_{TPE} 5) Significant symptom reduction can be seen only for IES-R, but not for BDI nor STAI-Y severity.	Same as above.	Mixed-effect model analyses: $\Delta\text{Sev} = 1 + \text{condition} + \text{order} + \text{condition}*\text{order} + (1 \text{patient})$ Where the Sev is the reduction in i) IES-R, ii) BDI, and iii) STAI-Y from pre-training test (-1 day) to post-training test (+1 week).	Positive/negative and statistically significant effects of the condition in each psychiatric severity will suggest that the experimental/control condition of DecNef is effective in reducing that psychiatric score. See null hypothesis policy*. The SESOI will be set at 20% of the calculated effect size determined during the interim analysis.
[DNF: Distress during Neurofeedback] Does DecNef convey weak distress during the procedure?	(H_{DNF} 1) The ratio of participants who reported their use of conscious strategies related to 1a) visual images of angry male faces and 1b) strategies related to traumatic events in the free answer descriptions and 2) angry face in the four-forced-choice question is lower than i) 10%, and ii) 50%.	Same as above.	Chi-square to test if the ratios of trauma-related strategies are lower than i) 10% and ii) 50% in the 1a) and 1b) free answer descriptions and 2) four-forced-choice question.	Significant results will suggest less/more than i) 10% and ii) 50% of participants are 1) explicitly, and 2) implicitly conscious that the content for the neural representation is related to 1a) and 2) angry male faces and 1b) trauma-related stimuli. See null hypothesis policy*. The SESOI will be set at [8-12] for i) and [40-60] for ii).
	(H_{DNF} 2) The ratio of participants who correctly answered the order of the two conditions in the two-forced-choice question is 0.5, the chance level.	We will use the same dataset collected for testing H_{TPE} 1 . We will calculate the Bayes factor adding participants one by one in the order of the study completion until the Bayes factor is at least 10 times in favour of the random hypothesis over the conscious hypothesis (or vice versa) or until all the participants (N = 58) are	Random hypothesis: $\theta = 0.5$ (participants respond randomly). Alternative conscious hypothesis: $\theta \sim U(0.8, 1)$, which assumes a uniform distribution over [0.8,1]. We will compute the Bayes factor (BF) to compare the random hypothesis against the alternative: $BF = \frac{P(\text{data} \theta = 0.5)}{P(\text{data} \theta \sim U(0.8,1))}$ Where the θ denotes the probability	Random hypothesis is supported with the BF larger than 3, whereas alternative “conscious” hypothesis is supported with the BF smaller than 0.3. Otherwise, the results will be deemed inconclusive regarding the random hypothesis. Bayes factors larger than 1,3, or 10 (or smaller than 1,0.3, or 0.1) provide anecdotal, moderate, and strong evidence for the Random (or Alternative) hypothesis.

		included in the analysis.	of choosing the correct order, and $1 - \theta$ be the probability of choosing the wrong order.	
	(H_{DNF 3}) Induction of angry male faces during the DecNef sessions induce SUDS smaller than i) 50 and ii) baseline.	Same as H_{DNF 1} .	We will apply i) one sample and ii) paired-samples t-tests of SUDS during the induction of angry male faces against i) 50 and ii) baseline.	<p>If SUDS are significantly lower than i) 50 and ii) baseline, we will consider that experimental condition i) carries at most an endurable load and ii) does not cause participants' distress because of the contents of the target neural representation.</p> <p>If SUDS are significantly higher than i) 50 and ii) baseline, we will consider that experimental condition i) carries unendurable load and ii) causes participants' distress because of the contents of the target neural representation.</p> <p>See null hypothesis policy*. The SESOI will be set at 20% of the calculated effect size determined during the interim analysis.</p>
	(H_{DNF 4}) The Dropout definitions 1a, b) and 2) are lower than 16%.	We will use the same dataset collected for testing H_{TPE 1} together with those dropout from the study. We will calculate the Bayes factor adding participants one by one in the order of the study completion until the Bayes factor is at least 10 times in favour of the low-dropout hypothesis over the null hypothesis (or vice versa) or until all the participants (N = 58, and those who dropout from the study) are included in the analysis.	<p>Experimental low-dropout hypothesis:</p> <p>$\theta \sim U(0, 0.14)$, which assumes a uniform distribution of the dropout rate θ over $[0, 0.14]$.</p> <p>Null hypothesis:</p> <p>$\theta = 0.16$, which is consistent with the dropout rate in meta-analysis¹⁰.</p> <p>We will compute the BF to compare the Experimental low-dropout hypothesis against the Null for each dropout in definition 1a, b) and 2).</p> $BF = \frac{P(data \theta \sim U(0,0.14))}{P(data \theta = 0.16)}$	<p>Experimental hypothesis is supported with the BF larger than 3, whereas null hypothesis is supported with the BF smaller than 0.3. Otherwise, the results will be deemed inconclusive regarding the low-dropout hypothesis. Bayes factors larger than 1,3, or 10 (or smaller than 1,0.3, or 0.1) provide anecdotal, moderate, and strong evidence for the Experimental (or null) hypothesis.</p>

<p>[DDC: Distress during Decoder Construction]</p> <p>Does Decoder Construction cause weak distress during the procedure?</p>	<p>(H_{DDC} 1) The presentation of angry male faces during the decoding sessions induce SUDs smaller than a) 50, or b) baseline.</p>	<p>Same as H_{DNF} 1.</p>	<p>We will apply a) one sample and b) paired-samples t-tests of SUDs during the decoding session against a) 50 or b) baseline.</p>	<p>If SUDs are significantly lower than i) 50 and ii) baseline, we will consider that decoding sessions i) carries at most an endurable load and ii) does not cause participants' distress.</p> <p>If SUDs are significantly higher than i) 50 and ii) baseline, we will consider that decoding sessions i) carries unendurable load and ii) causes participants' distress.</p> <p>See null hypothesis policy*. The SESOI will be set at 20% of the calculated effect size determined during the interim analysis.</p>
<p>Possible Neural Mechanisms</p>				
<p>[NM: Neural Mechanisms]</p> <p>What is the mechanism underlying DecNef effects on PTSD symptom reductions? Extinction learning mechanism or counter-conditioning mechanism?</p>	<p>(H_{NM} 1a) CAPS reductions correlate more strongly with information transmission from STS to the caudate and ventral striatum than to the vmPFC.</p>	<p>Same as above.</p>	<p>We will apply a Pearson correlation analysis of changes in PTSD symptoms with i) caudate and ventral striatum, and ii) vmPFC.</p> <p>We will then test the difference between these correlations.</p>	<p>The statistically significant association with i/ii) respectively suggests the involvement of a counter-conditioning/extinction learning mechanism in the DecNef.</p> <p>See null hypothesis policy*. The SESOI will be set at 20% of the calculated effect size determined during the interim analysis.</p>
	<p>(H_{NM} 1b) Caudate and ventral striatum will show stronger functional connectivity with STS as a function of target male angry face likelihood during DecNef, than the functional connectivity between vmPFC with STS.</p>	<p>Same as above.</p>	<p>We will apply a PPI analysis to examine the differential functional connectivity between STS and the i) caudate and ventral striatum and ii) vmPFC, as a function of the angry face representation likelihood in STS.</p> <p>We will then test the difference between these correlations.</p>	<p>The statistically significant association with i/ii) respectively suggests the involvement of a counter-conditioning/extinction learning mechanism in the DecNef.</p> <p>See null hypothesis policy*. The SESOI will be set at 20% of the calculated effect size determined during the interim analysis.</p>

	<p>(H_{NM} 2) CAPS reduction will be better predicted by the counter conditioning model ($V_{X(CC)}$), than the exposure/extinction-based model ($V_{X(EB)}$).</p>	<p>Same as above.</p>	<p>We will apply a fixed effect model:</p> <p>Model 1) $V_{X(SEV)} = \beta_{EB} V_{X(EB)}$</p> <p>Model 2) $V_{X(SEV)} = \beta_{CC} V_{X(CC)}$</p> <p>Model 3) $V_{X(SEV)} = \beta_{EB} V_{X(EB)} + \beta_{CC} V_{X(CC)}$</p> <p>For each model, we will calculate Bayesian Information Criterion (BIC) and compare it across models if the difference is larger than 2.</p>	<p>If the BIC is lower for Model 2 compared to Model 1, this will support the hypothesis that counter-conditioning computations better explain CAPS reductions than exposure/extinction-based computations. Conversely, if the BIC is lower for Model 1 compared to Model 2, it would suggest that exposure/extinction mechanisms better explain the reductions. If the BIC is lower for Model 3 compared to Model 1 and 2, this would suggest both mechanisms contribute.</p> <p>See null hypothesis policy*. The SESOI for $V_{X(EB)}$ and $V_{X(CC)}$ will be set at 20% of the calculated effect size determined during the interim analysis.</p>
--	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------	---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

*null hypothesis policy: In case the null hypothesis cannot be rejected, we will conduct an equivalence test using an interval of $[-SESOI, SESOI]$. If the observed 90% confidence interval is entirely within this equivalence range, we will interpret this as evidence supporting a null effect. Conversely, if the confidence interval falls outside this range, the results will be deemed inconclusive regarding the null hypothesis.

Methods

Ethics information

This study will be conducted with approval from the Institutional Review Boards of Osaka Medical and Pharmaceutical University and Advanced Telecommunication Research Institute International. Participants will provide written informed consent prior to each session. This study has been registered with the University Hospital Medical Information Network (UMINID: UMIN000028148).

Co-design and pilot data

A co-design approach was adopted (<https://www.medsci.ox.ac.uk/research/patient-and-public-involvement>) in which patient partners with PTSD played a key role in developing the paradigm, in terms of experimental logistics, study environment, management of adverse events, design of the task interface, optimisation of patient comfort, and task instructions. The resulting paradigm presented here reflects several iterations of this design cycle, with patient experience (as part of an explicit patient co-led risk management strategy) a fundamental part of this process. These patient partners and new patients to be enrolled in the pre-registered study will be actively involved as an oversight panel through the duration of the study, and will jointly consider the outcome of the study when complete.

We conducted a pilot study on a total of seven female participants with PTSD (mean age 39.7, SD = 10.1), three of which participated in both experimental and control conditions. This pilot study included data from four participants with PTSD that were reported elsewhere²² and the procedure used for these four participants was the same as that used for the other (previously unreported) three participants. In brief, neural activation patterns for trauma-related male angry faces and trauma-unrelated happy female faces (see *decoding session* for detail) were induced via DecNef in the experimental and control conditions, respectively. Procedures of the pilot study were mostly identical to those of this pre-registered study with two exceptions. First, while the pilot study was conducted in a single-blinded manner, we will conduct the pre-registered study in a double-blinded,

completely randomized manner. That is, experimental and control conditions in the pilot study were blinded only to participants whereas the pre-registered study will be further blinded to experimenters and analysts. The pre-registered study will thus validate that the observed effect in the pilot study is not due to a placebo or experimenter effect. Second, we mainly used mrVISTA and TurboBrainVoyager instead of Statistical Parametric Mapping (SPM; Wellcome Department of Imaging Neuroscience, London, UK) to analyze the pilot fMRI data, consistent with our previous studies²⁰ (see Supplementary Information for details). In the proposed study, we will use more commonly used software, SPM12, for ease of DecNef applications in clinical settings (see Design section for details).

In our pilot study, we observed that participants in experimental conditions ($N = 6$) showed improvement in symptom severity relative to participants in control conditions ($N = 4$) (Figure 1). Symptom severity was measured using the clinician-administered PTSD scale for DSM-IV, 1-wk Symptom Status Version (CAPS-SX²⁵), which is the gold standard in this field. On average, symptom severity assessed by CAPS²⁵ was reduced by 35.8 points ($SD = 8.8$) in the experimental group from the pre-training test (-1 day) to the post-training test (+1 week). Meanwhile, only 4 points ($SD = 21.1$) were reduced in the control condition group. As 10-point changes in CAPS are generally considered clinically meaningful^{8,23}, we concluded that those experimental conditions are worth pursuing in the pre-registered study. The alleviating effects of DecNef were consistent across all symptom clusters of PTSD defined in the DSM-IV: Re-experiencing symptoms was reduced by 15.3-points ($SD = 6.3$) in the experimental group and by 3-points ($SD = 6.4$) in the control group. Avoidance symptoms were reduced by 12.7-points ($SD = 7.6$) in the experimental group and by 0.75-points ($SD = 11.9$) in the control group. Hypervigilance symptoms were reduced by 7.8-points ($SD = 3.4$) in the experimental group and by 0.25-points ($SD = 6.7$) in the control group. Using pilot data, a two-tailed t-test comparing total CAPS score reduction from the pre-training (-1 day) to post-training (+1 week) yielded significant results ($t(8) = 3.4$, $p = 0.01$, Cohen's $d = 1.97$, 95% CI: 10.0-53.7).

In qualitative interviews after the entire pilot procedure, some participants reported a state of distress unrelated to DecNef itself, i.e., distress in traveling to the institute or distress in entering the

confined space of an fMRI scanner. Further, one out of ten participants used a strategy related to her trauma during manipulation of brain activity in the first experimental session out of six sessions (three experimental sessions and three control sessions), during which she had higher distress. However, she used a different strategy unrelated to trauma in the remaining five sessions, which involved no distress, as revealed during the post-experiment qualitative interview. All participants, including this one participant, reported that the decoding and DecNef procedures were generally endurable. Further, all participants reported that they had no idea as to the content of target neural representations. Beyond laboratory assessments, we observed real-life improvements of patient symptoms. Because of their symptoms, two of six participants were unemployed at the time of enrollment in the experimental group. After participating in the experiment, however, these two were able to start new jobs for the first time in one year and ten years, respectively.

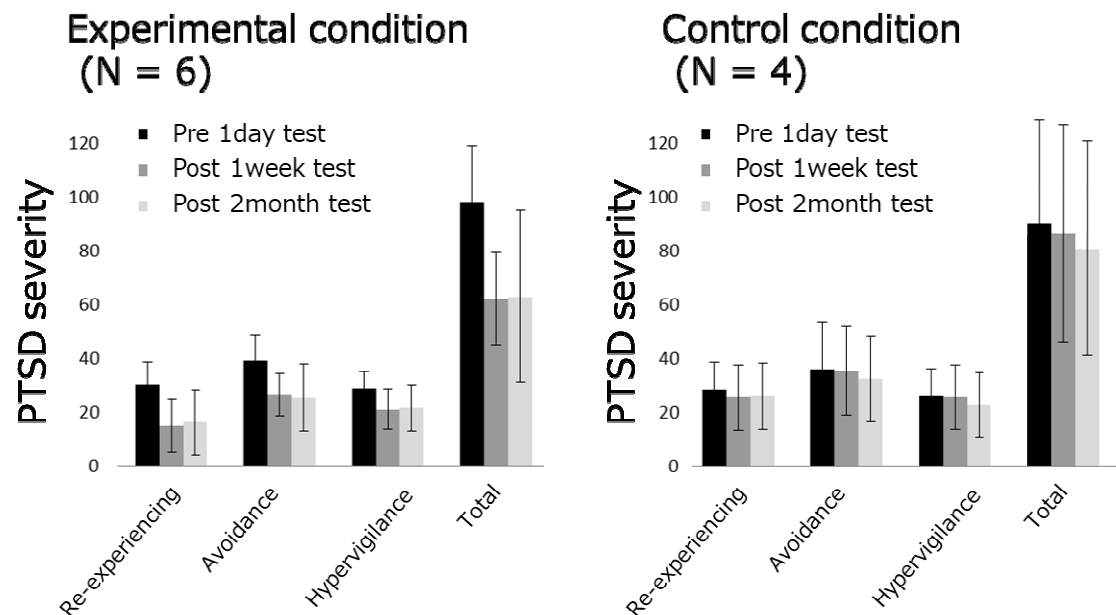


Figure 1. DecNef effects in the pilot study. Mean PTSD severity, as measured with CAPS scores, for three PTSD symptom clusters and their total in each test session. Error bars represent standard deviations.

Intended Endpoint

The primary endpoint of this study is PTSD severity reduction, measured by CAPS-SX²⁵, as in our pilot study. Consistent with previous studies, 10-point changes in CAPS will be considered clinically meaningful^{8,23}. The secondary endpoints are: 1) changes in physiological threat responses to threats (angry faces), as measured by SCRs and amygdala reactivity, and 2) reduction of PTSD, depression, and anxiety severity, as measured with validated self-administered questionnaires: the Impact of Event Scale-Revised (IES-R)⁴², BDI⁴³, and STAI-Y⁴⁴, respectively. These additional psychological assessments will be included to examine whether the effects of DecNef are specific to PTSD symptoms.

Objective threat responses will be assessed to determine their association with changes in CAPS scores that rely on subjective reports. Our objective measurements will include SCRs and amygdala reactivity, which are gold standards to index physiological threat responses. Alterations in the levels of both measures are pronounced during PTSD development^{45,46}. However, while elevated physiological threat responses (including SCRs and amygdala reactivity) are usually associated with higher overall PTSD severity, lower responses are also known to accompany greater dissociative symptoms of PTSD^{47–51}. Given that both elevated and reduced threat responses signify the severity of PTSD, we do not make any predictions regarding directionality of changes in physiological threat responses following DecNef. As an exploratory analysis, we will minutely examine associations between physiological measures, subjective PTSD symptoms, and PTSD subtype (see Supplementary Information for details).

Design

Design overview

We will test the effectiveness of DecNef for patients with PTSD in a double-blind, randomized, placebo-controlled crossover study. First, all participants will undergo one preparatory decoding session where neural representations for angry faces and happy faces will be identified for each patient. Second, they will undergo session blocks for the *experimental* and *control* conditions of

DecNef, in a counterbalanced order (Figure 2). Each session block consists of three consecutive days of DecNef sessions. In the experimental condition of the DecNef session, neural representations of trauma-related male angry faces will be induced as target activation patterns. In the control condition, neural representations of trauma-unrelated happy female faces will be the target of induction. In each session, target neural representations (angry male or happy female faces) will be reinforced with monetary rewards on three consecutive days (Figure 2). The effect of each condition will be evaluated in one pre-training test (-1 day from DecNef) and three post-training tests (+1 day, +1 week, and +2 months). We will allow a maximum of 2 days of jitter for post-training test (+1 week) and 2 weeks of jitter for post-training test (+2 months), depending upon participant schedules or requests, or MRI malfunctions. Decoding and the first DecNef session block will be separated by a few days to a few weeks. The second DecNef session block will be conducted more than 6 months after the first DecNef session block (i.e. more than 4 months after the final post-training test of the first DecNef session block).

A double-blind, randomized, placebo-controlled crossover design

In addition to a typical double-blind design where conditions are blinded to the participants and experimenters, evaluators and analysts will also remain blind. To achieve this, before starting the experiment, a correspondence table of IDs and conditions will be created by a program and kept in a state that only the program can reference. The experimenter will then assign IDs to the participants without accessing the assigned conditions. During DecNef sessions, the execution program will refer to the correspondence table to apply the conditions corresponding to the IDs. Likewise, the analyst will refer only to the IDs, and the analysis execution program will automatically reference the correspondence table internally to distinguish between the two conditions. The correspondence table will be maintained and verified only by a non-author monitoring personnel, and researchers will access the table only after completing all the analyses.

Half the participants will undergo the experimental condition first and another half will undergo the control condition first (see Supplementary for the details). A computer program will semi-randomly assign the condition order for a given participant. To ensure similar levels of PTSD

severity between the condition order groups prior to the enrollment, the program will assign participants to approximately equate the proportion of moderate PTSD ($45 \leq \text{CAPS} \leq 60$; M-PTSD) and severe PTSD ($\text{CAPS} > 60$) cases.

We will use trauma-unrelated neural representations—those for happy female faces—as a placebo in the control condition instead of sham or yoked feedback, which typically relies on signals regardless of the participants’ performance, for four reasons. First, it is less ethical to require patients to engage in seemingly ineffective procedures albeit their significant effort and time. Second, if both the experimental and control conditions reduce distinct PTSD symptom clusters as hypothesized (see “Analysis of DecNef effects on each PTSD symptoms cluster” in the Supplementary Information), we can evaluate distinct effects across conditions within participants using a crossover design. Third, we observed that the control condition is less effective than experimental conditions in our pilot²². Fourth, this approach prevents participants from guessing the assigned condition based on the controllability of brain activity.

The crossover design can cancel across participants variability since each participant will undergo both conditions sequentially. To minimize a potential carryover effect from the first to second condition, we will insert a 6-month interval between them. Carryover effects, if any, should last at most for 6 months as a previous ROI-based neurofeedback study reported symptom reduction that lasted several months maximum⁴⁰. Similarly, effects of DecNef or functional connectivity based neurofeedback may also last several months^{13,41}. While anterograde effects were reported in previous DecNef experiments³⁹, Cortese et al. successfully induced bi-directional behavioral changes by reinforcing two opposing neural representations in two session blocks with just one week interval³⁹. Thus, an interval of 6 months is likely to minimize the anterograde effect from the first training.

One may worry that most participants might no longer exhibit PTSD symptoms when they enter the second DecNef session block after the 6-month interval. However, this is unlikely since PTSD patients typically remain symptomatic (total CAPS score > 20 ⁵²) even after several years of

spontaneous observation or following post-trauma-focused interventions^{53,54}. Women with PTSD, the target population in this proposed study, have been shown to remain symptomatic (CAPS > 45 at the lower boundary of the mean) 6 months after trauma-focused psychotherapy, as demonstrated in a large randomized controlled trial (N = 284)⁸.

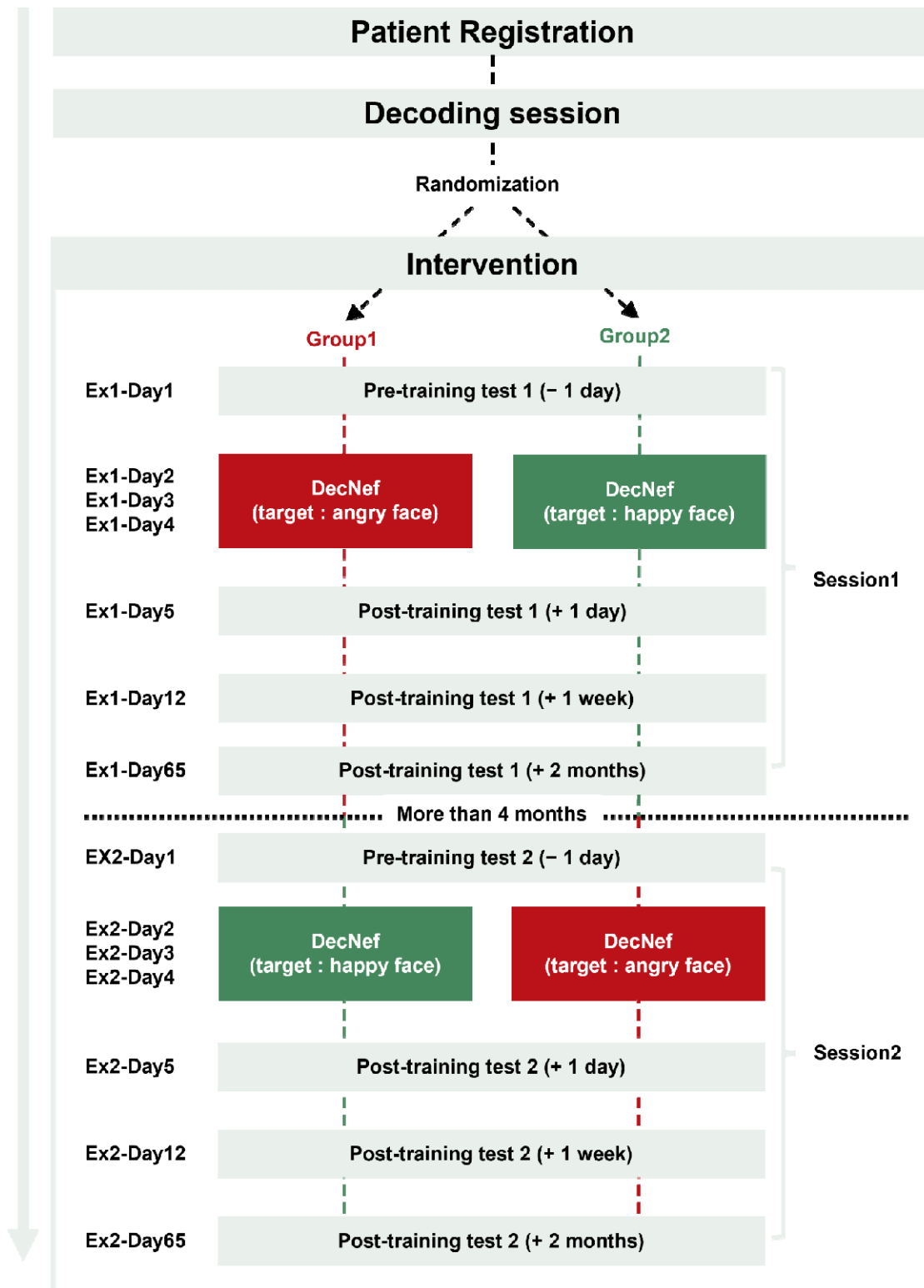


Figure 2. Overview of experimental design. Fifty-eight participants with PTSD will perform DecNef training in both the experimental and control conditions. The effects of DecNef in

each condition will be assessed with pre- and post-training tests. The order of conditions will be double-blinded and semi-randomized for each patient.

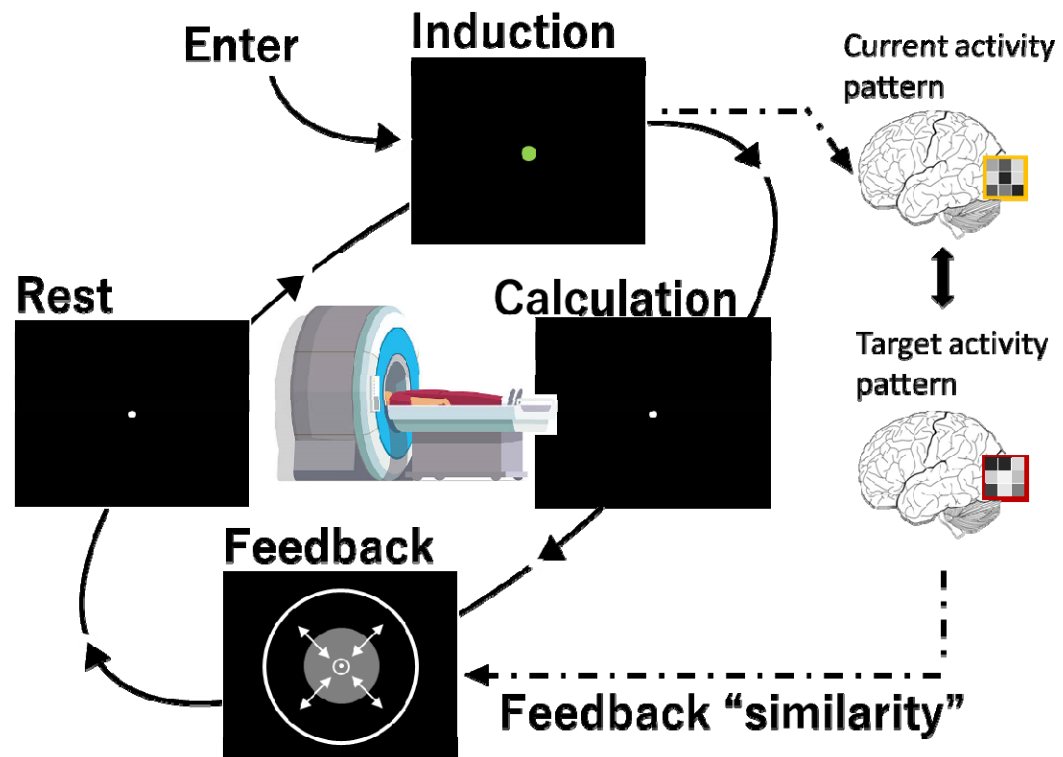


Figure 3. Overview of DecNef session. Each trial during DecNef sessions will begin with an induction phase, in which participants will manipulate their brain activity patterns to approach a predetermined target activity pattern. In the calculation phase, similarity between current brain activity patterns and the target activity patterns (angry or happy faces) will be calculated as likelihoods of the target representation using a pre-trained, multivoxel decoder. Calculated similarity will be fed back to the participant in the feedback phase. After the rest phase, participants will enter the induction phase for the next trial. By repeating trials, participants will be reinforced to induce the target brain activity patterns through trial and error.

MRI parameters and equipment

Experiment will be conducted using a 3.0T scanner (Prisma; Siemens, Erlangen, Germany) with a 32-channel head coil at the ATR Brain Activation Imaging Center, using parameters identical to those used in Koizumi et al. and Taschereau-Dumouchel et al^{17,20}. Specifically, we will scan 33 interleaved axial slices, 3.5-mm thick, without a gap, parallel to the anterior and posterior commissure line, with a T2*-weighted gradient, echo-planar imaging (EPI) sequence [repetition time (TR) = 2000 ms, echo time (TE) = 30 ms, flip angle (FA) = 80°, field of view (FOV) = 192 × 192 mm², voxel size = 3 × 3 × 3.5 mm³]. During the scan, we will record SCR from distal phalanges of the index and middle fingers of the left hand using BrainAmp Ag/AgCl sintered MR electrodes (Brain Products, inc.). For anatomical reference, we will acquire high-resolution T1-weighted images of the whole brain from all participants using a magnetization-prepared, rapid-acquisition gradient echo sequence (MPRAGE) [TR = 2250 ms, TE = 3.06 ms, FA = 9°, FOV = 256 × 256 mm², thickness = 1 mm, 0 mm slice gap, voxel size = 1 × 1 × 1 mm³].

Decoding session

The aim of the decoding session is to obtain fMRI data for constructing a multivariate pattern analysis (MVPA) decoder to classify activation patterns evoked by trauma-related angry male faces versus unrelated happy female faces. Based on previous studies reporting that the superior temporal sulcus (STS) contains information about the category of perceived emotions in faces^{55,56}, the STS was selected as a region of interest (ROI) for our MVPA (see Decoder construction and Online fMRI analyses during DecNef). The decoder will be used in subsequent DecNef sessions to evaluate the trial-by-trial likelihood that current activation patterns in participants represent trauma-related angry male or unrelated happy female faces in experimental or control conditions, respectively. Here, we will refrain from using specific faces of actual trauma perpetrators for a given participant since such an image is not always available in future clinical settings. Instead, we will use grayscale pictures of four angry male faces and four happy female faces from the ATR Facial Expression Image database (DB99) where recognizability of intended emotions is well-validated (ATR-Promotions, Kyoto, Japan, <http://www.atr-p.com/face-db.html>).

A modified CFS method will be applied to render face presentations less conscious and distressful. Because decoding brain activity patterns for invisible visual stimuli are likely more difficult^{57,58}, we will customize the CFS method to achieve an optimal balance between high decoding accuracy and conscious awareness of stimuli by gradually reducing mask contrast.

In each trial, a face stimulus (angry male or happy female) will be gradually faded into the non-dominant eye by linearly increasing its contrast over 1 s and then remaining constant for 6 s, while CFS masks containing salient Mondrian patches will be flashed to the other eye for 7 s at 10Hz. The mask will be presented at the same contrast as the face images for the first 1 s and then gradually faded-out linearly over 6 s. These presentations will be followed by presentation of a fixation disc to both eyes (for 7 s). We will extract preprocessed fMRI signals from the 6 s of each trial where a face is presented with a constant contrast. Because the TR = 2 s, this will provide three data points per trial, which will be averaged for each voxel from STS subregions and then used to construct a decoder to classify activation patterns for trauma-related angry male versus unrelated happy female faces. The session is subdivided into 11 runs of 17 trials (5 min) each. Sixteen exemplars (eight angry male faces and eight happy female faces) will be shown once per run in a randomized order. Each run starts with a dummy trial where a randomly selected happy female face is presented to capture irrelevant physiological threat responses due to the orienting effect⁵⁹. This dummy trial will be discarded from subsequent analyses. Thus, the session comprises a total of 187 trials (88 trials per face category and 11 dummy).

To assess the level of conscious awareness of masked face images, participants will be told to press a button when they see any image other than Mondrian images during MRI scan. Trials where the button is pressed/unpressed will be defined as conscious/nonconscious trials, respectively. We will not impose further questions to confine the content of images reaching consciousness, as such questions themselves could alter prior expectations for presented images, which would in turn modulate actual neural activations⁶⁰ to interfere with decoding. Before the beginning of the decoder session, eye dominance will be examined using the hole-in-a-card test⁶¹. T1-weighted images will be acquired in this session and used as anatomical references for functional images from all sessions.

To assess the level of distress during decoding, participants will be asked to rate the

subjective units of discomfort (SUDS²⁵) that they experienced during the decoding session, as well as before the first fMRI of the decoding session, as a baseline. The SUDS scale is a continuum from 0 (no distress) to 100 (maximum load), and 50 represents the strongest load that is still considered endurable. SUDS will be assessed at the end of each fMRI run, roughly every 5 min. This is comparable to a previous study that assessed SUDS during prolonged exposure therapy⁶². To avoid imposing task pressure on participants to report extra discomfort due to frequent assessments, we will add a dummy question on sleepiness level (Stanford sleepiness scale) following each SUDS assessment.

DecNef session

The aim of DecNef sessions is for participants to repetitively induce target STS activation patterns, specifically those for trauma-related angry male faces in the experimental condition and those for the trauma-unrelated happy female faces in the control condition. To achieve nonconscious induction of target activation patterns, we will neither give explicit instructions on how to manipulate their brain activity nor disclose the identity of the target patterns. We will reinforce participants with larger monetary rewards for inducing brain activity patterns more similar to the target patterns. In each session block, DecNef sessions will be conducted on three consecutive days. On each day, participants will engage in up to 12 fMRI runs (one run = 350 s), which are separated by brief break periods upon participant request. Each run consists of an initial 30 s fixation period followed by 16 trials (one trial = 20 s). To avoid unsaturated T1 effects, we will discard fMRI data for the initial 10 s of the fixation period. We will instruct participants to fixate their eyes on a dot presented at the center of the display throughout a run (Figure 3). Each trial involves the following sequence: an induction period (6 s), a fixation period (7 s), a feedback period (1 s), and an inter-trial interval (6 s) (Figure 3). During the induction period with a green dot presented centrally on-screen, participants will be instructed to maximize the size of the gray disc (surrounding the green dot) that serves as feedback by somehow manipulating their brain activity. Subsequent to the induction period, the fixation target will be presented as a central white dot on-screen, and the STS activation pattern during the induction period will be analyzed online to estimate the likelihood of target activation patterns (see Online fMRI

analyses during DecNef). Then, feedback will be presented as a gray disc, with a radius proportional to the average likelihood (ranging from 0 to 100%) of target face representation in the STS during the induction period of the current trial. Participants, importantly, will not be informed of the association between the induced neural representations, e.g., an angry face, and the size of the disk itself.

In addition, at the end of each DecNef session block (Day 4 for the first DecNef session block; Day 184 for the second DecNef session block), participants will write the strategy they used during the induction period when they were manipulating their brain activity to maximize the size of the gray disc. Participants will then be asked to respond to a four-forced-choice question: “Do you think you were guided to induce in your brain a representation of 1) mechanical tools, 2) animals, 3) a smiling woman, or 4) an angry man.” In addition, a two-forced-choice question will be presented at the end of all procedures, i.e., the last test following the second session block, as follows: “You participated in two conditions in randomized order. Experimenters tried to induce representations in your brain of angry male faces and happy female faces. Which condition do you think you received first? (choose either angry male faces or happy female faces condition).” The participants will not be provided with any feedback, which minimizes the risk that these procedures interfere with the therapeutic effects.

To assess the level of distress during DecNef, the SUDS will be assessed at the end of each fMRI run, roughly every 5 min, as well as before the first fMRI run of each DecNef session as a baseline. A dummy question on sleepiness level (Stanford sleepiness scale) following the SUDS assessment will also be added, as in the decoding session. In addition, a two-forced-choice question will be delivered at the end of all procedures, i.e., the last test following the second session block, as follows: “Which DecNef session block was more discomforting: the first or second DecNef session block?”.

Pre/Post test sessions

We will assess participant PTSD severity, depression severity, anxiety severity, and the strength of their physiological threat responses at the following times for each DecNef session block:

the day before the first day of DecNef (pre-training test (-1 day)), and one day after (post-training test (+1 day)), one week after (post-training test (+1 week)), and 2 months after (post-training test (+2 months)) the last day of DecNef (Figure 1). The procedure for all of these tests will be mostly identical. Each test will generally assess subjective PTSD severity as measured using CAPS, physiological threat responses to threat (angry faces) as measured by SCRs and amygdala reactivity, and subjective severity of other psychiatric symptoms as measured by questionnaire-based psychological assessments (see section *Endpoint*). One exception is that the CAPS will not be assessed during the post-training test (+1 day). This is because CAPS-1-week assesses PTSD severity over the past week, which means that at least one week should lapse after the pre-training test to avoid an overlap in periods. Therefore, changes in PTSD severity from before until after DecNef will be assessed using the CAPS at the pre-training test and the post-training test (+1 week). Further, long-term effects of DecNef will be assessed using the CAPS at the post-training test (+2 months).

In each test, physiological threat responses to angry male faces and happy female faces will be measured simultaneously via SCRs and amygdala reactivity inside an MRI scanner. The procedure of the MRI experiment for pre/post-training test sessions is consistent with that of the decoding session, except for the following two points. First, CFS masks that render the target face less visible will be presented at the same contrast as the target face without gradually decreasing their contrast, while the target face is presented to the other non-dominant eye. While the CFS masks may somewhat suppress SCRs and amygdala reactivity, we prioritize minimization of subjective distress during the test. Second, the whole experiment will comprise two runs, instead of 11 runs as in the decoding session.

Sampling plan

Sample size

We will conduct this study at Advanced Telecommunications Research Institute International (ATR), Japan, with a total of 58 participants with PTSD between 20 and 55 years old.

H_{TPE} 1 (Primary Outcomes): We compared two sample size estimates to select the larger sample size, avoiding potential underestimation due to sole dependence on a single pilot study or on previous

literature^{63,64}. In the first sample size estimation, we calculated the effect size using the SIMR⁶⁵ in R based on the pilot data. The SIMR can estimate the sample size of the mixed-effects model planned in this study, based on the simulated or estimated coefficients and variability. Here, we used coefficients and variability estimated from our pilot study²² to test the efficacy of DecNef on PTSD symptoms (also see **Table 1. Design Table**), which revealed that a sample size of 20 for each condition is sufficient to achieve 95% power. In the second estimation, we conducted a power analysis based on the latest meta-analyses on the PTSD psychotherapy (including prolonged exposure)⁶⁶. Here, since the coefficients and variability in the mixed-effects model are unavailable and highly uncertain, we used G*Power 3.1.9.7⁶⁷ to calculate the sample size to detect the minimally clinical important differences as smallest effect sizes with a t-test. Based on the meta-analyses on the psychotherapy that the effect size of Hedge's $g = 1.248$ (95% CI: 0.684-1.813) in comparison with treatment as usual (or routine care) or effect size of hedge's $g = 1.524$ (95% CI: 1.235-1.814) in comparison with wait list⁶⁶, we assume 0.684 (the lowest bound) as the smallest effect size of interest (SESOI). With a two-tailed t-test ($\alpha = .05$, $\beta = .95$), this calculation estimated that we need 57 participants to detect a significant effect. We will therefore enroll 58 participants (minimum number for counterbalancing) to complete all sessions. Experiments will be continued until 58 participants complete all procedures. For the efficacy, safety, futility, and time efficiency we will conduct interim analyses based on the O'Brien-Fleming stopping boundaries⁶⁸, with predefined significance thresholds: the first analysis at 10% of the data (i.e., 6 participants) with $p < 0.0005$, and the second analysis at 50% of the data (i.e., 29 participants) with $p < 0.015$. The trial will be terminated early if these thresholds are met for both of primary outcomes (i.e., **H_{TPE} 1 and 2**). Considering the long interval period of more than half a year, we will continue enrolling participants toward the full study target (i.e., $N = 58$) until the stopping rules are met. If the stopping rules are met, we will halt further enrollment while allowing already enrolled participants to complete the full study procedures. All analyses will be conducted using data from participants who have completed the full procedures.

H_{TPE} 2 (Primary Outcomes): H_{TPE} 2 evaluates the DecNef effect using a clinically meaningful SESOI, defined as a 10-point reduction in CAPS (equivalent to hedge's $g = 2.35$ from our pilot study).

A two-tailed one-sample t-test ($\alpha = .05$, $\beta = .95$) estimates that five participants are required to detect significance. Considering power analysis for **H_{TPE 1}**, the total sample size was set at 58 participants.

Other Hypotheses: For secondary analyses on therapeutic effects (**H_{TPE 3-5}**), distress during DecNef (**H_{DNF}**) and during preparatory decoder construction (**H_{DDC}**), and neural mechanisms (**H_{NM}**) where effect sizes are uncertain, the dataset collected for primary outcomes (i.e. **H_{TPE 1, 2}**) will be repurposed. Interim analyses will be conducted at the halfway point (N=29) to assess whether a sample size of 58 provides 95% power for each hypothesis. Hypotheses will be tested only if the required sample size remains within this threshold. The SESOI for the equivalent test is defined as 20% of the interim-calculated effect size to ensure meaningful and robust hypothesis testing⁶⁹. For the hypothesis where the Bayesian approach is adopted (**H_{DNF 2}** and **H_{DNF 4}**), we will calculate the Bayes factor adding participants one by one in the order of the study completion until the Bayes factor is at least 10 times in favour of the experimental hypothesis over the random hypothesis (or vice versa) or until all the participants (**H_{DNF 2}**: N = 58 that defined for **H_{TPE 1}**; **H_{DNF 4}**: N=58, plus all participants who dropped out of the study) are included in the analysis.

Inclusion criteria

PTSD patients will be recruited from Osaka Medical and Pharmaceutical University, the Seven-Mental Clinic (Osaka, Japan), and the Flower of Light Clinic for Mind and Body (Tokyo, Japan). Patients will be registered in this study by medical doctors, based on inclusion and exclusion criteria approved by Osaka Medical and Pharmaceutical University. Inclusion criteria require participants to be female, 20 to 55 years of age, meeting DSM-IV criteria for current PTSD diagnosis, having a score of ≥ 45 on the clinician-administered PTSD scale (CAPS²⁵), and having strong fear when passively viewing pictures of angry male faces, with a score of >60 on the SUDS. Participants will be also required to agree not to receive other psychotherapy for PTSD during the study treatment. Additionally, if participants are being treated with psychoactive medication, they must have been on a stable regimen (with no changes in drugs or doses) for at least 2 months prior to the DecNef session. Psychoactive medication can be modified after the post-training test (+2 months) of the first DecNef

session block. However, the pre-training test of the second DecNef session block will be postponed until 2 months have passed on a stable regimen.

Exclusion criteria

Exclusion criteria are: active suicidality, a current diagnosis of substance dependence or psychosis, a history of moderate or severe head injury, and/or general contraindication to MRI. Participants with excessive head-motion during fMRI scanning (above 1.5 voxels on average), as well as those who provide fMRI data with lower classification accuracy (<56%) will be also excluded from analyses. We will use 56% as the cutoff, since pilot participants showed decoding accuracy higher than 56%, which led to clinically significant symptom improvement through DecNef in all participants.

fMRI Analysis

Decoder construction

During decoder construction, for each trial, functional data will first be realigned to match coordinates of the first EPI image acquired in the Decoding session. Data will then undergo head-movement correction (using the realignment function of SPM12). We will use a gray-matter mask to extract fMRI data only from gray-matter voxels. Coordinates of the AAL atlas definition of the STS (Temporal_Sup_L, Temporal_Sup_R) will be extracted from this gray-matter using WFU_Pickatlas in MNI space and will then be retransformed to native space. Using the resulting coordinates in native space, we will extract the time-courses of BOLD signal intensities from the STS. Voxels with exceptional value will be removed. These will include voxels with exceptionally low BOLD signal intensities (mean <100) or those with exceptionally large variances ($SD > 8$ or $SD < -8$). We will remove a linear trend from the time-course. To further minimize baseline differences across runs, the time-course will be z-score normalized for each voxel in each run. We will extract pre-processed fMRI signals from the 6 s time-period of each trial of the Decoding session in which the face stimulus is constantly present on-screen, with a delay of 6 s to account for the hemodynamic delay. STS signals

from this 6 s period, i.e., three TRs, will be averaged for each voxel. These will then be used to construct a decoder to classify activation patterns for angry male versus happy female faces. We will use Sparse logistic regression (SLR)⁷⁰ to automatically select relevant voxels for classification. To estimate the validity of the constructed decoder, leave-one-run-out cross-validation will be performed. To be used in the DecNef session, we will train the decoder, without cross-validation, using data from all fMRI runs, i.e., 176 data points from 176 trials. Details on fMRI analysis tools are described in Supplementary Method.

Online fMRI analyses during DecNef

For online fMRI analyses during DecNef sessions, measured functional images will first undergo three-dimensional motion correction in real time so that they match coordinates of the first EPI image acquired in the Decoding session (using the realignment function of SPM12). Second, we will extract time-courses of BOLD signal intensities from each STS voxel that is relevant for decoding. Third, we will remove a linear trend from the time course. The BOLD signal time course will be z-score-normalized for each voxel using BOLD signal intensities measured for the initial 20 s fixation period of each run, after removing the initial 10 s from the whole 30 s fixation period. (Also see *DecNef session*). Fourth, we will create a data sample to calculate feedback by averaging BOLD signal intensities of each voxel for three TRs in the induction period, allowing for 6 s (three TRs) of hemodynamic delay. Before calculating feedback, the correlation between current and decoder activation patterns will be calculated to remove “error trials.” If the correlation is below 0.9, feedback will be presented as an error (a capital letter “E” will be displayed on the display). These trials will be defined as error trials. If the correlation is at or above 0.9, we will calculate the likelihood that neural activity from the induction period represents the target stimulus by applying the to-be-constructed decoder for the to-be-acquired data sample. The radius of the feedback disc will be proportional to the likelihood of target facial characteristics (angry male/happy female) assigned to each condition (experimental/control) on a given DecNef session block by the STS.

All of these procedures are consonant with previous DecNef studies^{13,14,20}, except that we will

use a pipeline based on SPM software (<https://bicr.atr.jp/decnefpro/software/>). This is in contrast to previous DecNef studies, as well as our pilot study, which relied on TurboBrainvoyager for head-motion correction.

Offline fMRI analyses for test sessions

Amygdala reactivity in response to angry male or happy female faces from the pre/post-training test session will be assessed for each test (pre-training, post-training (+1 day), post-training (+1 week), and post-training (+2 months), for each condition) for each participant based on subsequent offline analyses. An exemplar analysis workflow with FSL tools will be executed. First level analysis will utilize FILM (FMRIB's Improved Linear Model) to set up a standard generalized linear model (GLM). Specifically, three types of GLM analyses will be performed for each voxel for each participant to extract contrast statistics between: 1) angry male versus happy female faces, 2) angry faces versus fixation, and 3) happy female faces versus fixation. By averaging the aforementioned contrast statistics (normalized coefficient, beta) calculated from voxels in the anatomically defined amygdala, we will define amygdala reactivity in response to 1) angry male versus happy female faces, 2) angry faces, and 3) happy female faces, respectively.

Amygdala reactivity in response to angry male or happy female faces from the decoding session will be assessed for each participant in a similar manner.

Skin conductance responses Analysis

Consistent with a previous study^{17,20}, stimulus-elicited SCRs for each test to be analyzed, i.e., a pre-training test or a post-training test, consisting of two fMRI runs, will be defined as averaged SCRs for happy female faces subtracted from those for angry male faces. We will use the following methods of SCR extraction: 1) Before analyses, a band-pass filter (transmission range, 0.05-1 Hz) will be applied to each fMRI run to remove noise from the data; 2) The SCR for each stimulus presentation will be calculated as the maximum SCR that occurs during stimulus presentation (from 0.5 s after onset of the presentation until the end of the presentation) as compared to baseline, which

will be defined as the averaged skin conductance from 2 s before stimulus presentation; 3) SCRs lower than 0.02 μ Siemens will be scored as 0; 4) SCRs will be square root transformed to correct for the skewness of their distributions¹⁷; 5) SCRs from each test will then be averaged separately for angry male faces and for happy female faces, after exclusion of the initial dummy trial of each fMRI run to remove irrelevant orienting effects (see pre/post-training test session).

Analytical Plan

Analyses will be performed only for data of completers. In cases where the null hypothesis cannot be rejected, we will perform an equivalence test to determine if the observed 90% confidence interval falls entirely within the equivalence range. The equivalence range will be defined as $\pm 20\%$ of the effect size used in the power analysis for each hypothesis⁶⁹.

To address potential biases from randomization and dropouts, supplementary analyses will include baseline comparisons (**SH_{BC} 1 & 2**) and intent-to-treat analyses (**SH_{TPE} 1-3**). Baseline comparisons will evaluate equivalence of demographic and clinical characteristics between completers and dropouts as well as across randomized groups. Intent-to-treat analyses will replicate the main analyses, incorporating data from all participants, including those who dropped out.

Neyman-Pearson inference will test the alternative hypothesis (predicting a difference), while Bayesian hypothesis testing will assess the random hypothesis (predicting no difference: **H_{DNF} 2**) and compare the two distributions (**H_{DNF} 4**).

Analysis of PTSD severity (H_{TPE} 1, 2, and 3)

H_{TPE} 1: We will analyze the main effect of DecNef conditions, either experimental or control, on potential reductions in CAPS scores (Δ Sev) using the following mixed-effect model:

$$\Delta\text{Sev} = 1 + \text{condition} + \text{order} + \text{condition}*\text{order} + (1|\text{patient})$$

where ΔSev will be defined for the post-training test (+1 week) as a reduction from the pre-training test (-1 day). The condition denotes the type of condition (experimental vs control condition) while the order denotes the order of the condition.

H_{TPE 2}: We will analyze if the DecNef effect exceeds a clinically defined meaningful SESOI, defined as a 10-point reduction. Specifically, we will apply one sample t-tests of reduction in CAPS from pre-training test (-1 day) to post-training test (+1 week) against 10-point reduction.

H_{TPE 3a & b}: We will conduct a similar analysis for **H_{TPE 1 & 2}** using a post-training test (+2 months) instead of post-training test (+1 week).

Supplementary Hypotheses (see Supplementary *Design Table*): We will conduct a similar analysis for **H_{TPE 1-3}** in an intent-to-treat analyses instead of complete analyses (**SH_{TPE 1-3}**).

Analysis of physiological threat responses (H_{TPE 4})

H_{TPE 4a}: The mean SCRs/amygdala reactivity for angry male faces minus those for happy female faces will be analyzed using the mixed-effects model for **H_{TPE 1}**, the CAPS analyses. Baseline correction using the reactivity to fear-irrelevant stimuli—specifically, happy female faces—enables the evaluation of fear-specific reactivity^{17,20,71}. In contrast to CAPS analyses, we will use data from post-training tests (+1 day) instead of post-training tests (+1 week).

H_{TPE 4b}: Whether variations in PTSD severity as measured using IES-R (Sev) are accompanied by variations in SCRs/amygdala reactivity (Fear) will be analyzed using the following mixed effect model:

$$\text{Fear} = 1 + \text{Sev} + (\text{Sev}|\text{patient})$$

Here, Fear and Sev will be derived from all assessment time points: pre-training test (-1 day) and post-training tests (+1 day, +1 week, +2 months).

Supplementary Hypotheses (see Supplementary *Design Table*): Extracted SCRs/amygdala reactivity for each face type (angry male and happy female) will also be analyzed in a similar vein to **H_{TPE} 4a**. The analyses explore the possibility that reactivity for a specific face type, including happy female faces, may be affected by DecNef, considering that happy female faces are reinforced in the control DecNef condition (**SH_{TPE} 4a**).

Analysis of questionnaire-based psychological assessments (H_{TPE} 5)

H_{TPE} 5: The reduction of IES-R, BDI, and STAI-Y scores will be analyzed separately, but in a similar manner to CAPS analyses (see analysis for **H_{TPE} 1**). We will use the IES-R, a self-report questionnaire instead of the CAPS, a structured interview, to enable fair comparison with other self-report measures such as the BDI and the STAI-Y.

Supplementary Hypotheses (see Supplementary *Design Table*): If **H_{TPE} 1** demonstrates the efficacy of DecNef, supplementary analyses will investigate its effects on specific PTSD symptom clusters (**SH_{TPE} 5**), examine how PTSD subtypes influence therapeutic outcomes (**SH_{TPE} 6**), and explore detailed associations between physiological threat responses and symptom characteristics (**SH_{TPE} 7**).

Evaluating consciousness, distress, and dropout with DecNef (H_{DNF})

H_{DNF} 1: Conscious awareness of neural representations of trauma-related stimuli, i.e., angry male faces, during DecNef will be assessed as follows: We will first evaluate ratios of participants who reported their use of conscious strategies related to 1) visual images of angry male faces and 2) strategies related to traumatic events, including visual images of angry male faces, based on their open-ended responses, at the end of each DecNef session (see DecNef session). Categorization of strategy will be performed independently by two judges, including the doctor-in-charge. The corresponding author will categorize the strategy in case of conflict. We will use a chi-square test to examine whether ratios of conscious strategies are lower than i) 10% and ii) 50%. If they are lower than these thresholds, we will consider that explicit consciousness, a form of side effect, i) are not very frequent⁷², and ii) do not emerge in more than half of the patients. Further, chi-square tests will

be performed on the four-forced-choice question asking participants to report the content of induced neural representations during DecNef. Specifically, we will examine if the ratios of participants who selected “an angry man” are lower than ii) 10% and ii) 50%. If they are lower than these thresholds, we will consider that implicit consciousness, a form of side effect, i) are not very frequent⁷², and ii) does not emerge in more than half of the patients. Chi-square tests will be replaced with Fisher's exact test if expected frequencies in cells are below 5.

H_{DNF} 2: To assess differential conscious experiences beyond H_{DNF} 1, we will test whether participants can identify the condition order after the experiment, knowing that the experimental condition involves trauma-related content. Accurate reporting would indicate conscious awareness during DecNef, undermining its clinical benefits. Since we predict no such awareness, we will compare the Random and Alternative Conscious hypotheses using Bayesian testing.

The Random hypothesis assumes random guessing ($\theta = 0.5$). Where the θ denotes the probability of choosing the correct order, and $1 - \theta$ be the probability of choosing the wrong order. The Alternative Conscious hypothesis posits awareness sufficient for accurate reporting ($\theta \sim U(0.8, 1)$), modeled as a uniform distribution over $[0.8, 1]$, allowing for 20% variation as in SESOI estimation. Here, we did not consider awareness that is insufficient for accurate reporting but still above chance (e.g. $\theta = 0.6$), as such vague awareness would not necessarily undermine DecNef's clinical benefits.

H_{DNF} 3: Distress during DecNef will be assessed with subjective measures of SUDS levels. One-sample/paired-samples t-test will be performed to assess whether the averaged SUDS during the experimental condition is lower than a) 50 (the strongest load that is still considered endurable), b) the baseline score measured before the first fMRI of each DecNef session. If SUDS during the DecNef is < 50 , we will consider that experimental condition carries at most an endurable load. If SUDS is lower than baseline, we will consider that the experimental condition does not enhance participant distress because of the contents of the target neural representation.

H_{DNF} 4: We will examine if the dropout rate in our study follows a distribution lower than the pooled dropout rate reported in the meta-analysis of RCTs of PTSD psychotherapy in general (mean = 16%,

95% CI: [14-18%])¹⁰. Specifically, we will compare the Null and Experimental hypotheses using Bayesian testing. The Null hypothesis assumes a dropout rate θ equivalent to the meta-analysis ($\theta = 0.16$)¹⁰. The Experimental hypothesis posits lower distribution ($\theta \sim U(0, 0.14)$), modeled as a uniform distribution over [0,0.14]. In both hypotheses, the number of dropouts follows a binomial distribution as a function of θ ($X \sim B(N, \theta)$). In the Bayesian testing, the evidence for the experimental hypothesis will be calculated using a Beta distribution Beta (1,1), rescaled over the interval [0, 0.14].

Three types of dropout (1-a, 1-b, and 2) will be defined, as follows:

1) Dropout at any point in the entire procedure, excluding genuine interval period:

1-a) Dropout for any reason

1-b) Dropout upon request of the participant or the doctor-in-charge (excluding reasons unrelated to participant health, such as MRI malfunction, below-chance-level decoding accuracy, excessive head movements in the scanners, family concerns, or traffic conditions).

2) dropout during one of the three consecutive days for both DecNef sessions.

The dropout rate in each definition will be defined as the percentage of total dropouts relative to all participants entering the study. Since the RCTs in the meta-analysis did not adopt a crossover design and therefore did not include an interval period, dropout definition 1) will not include participants who left the study during the genuine interval period, from post-training test 1 (+2 months) to pre-training test 2 (-1day). Dropout definition 1-a) is the most conservative definition. If this dropout rate is lower than 14-18%, we will consider DecNef superior to current psychotherapies in terms of therapy design. The dropout definition 1-b) provides a fair measure for comparison with current psychotherapies. If this dropout rate is lower than 14-18%, we will consider DecNef superior to current psychotherapies, at least in terms of therapy adherence. The dropout definition 2) provides insights into the tolerability of DecNef in comparison with current psychotherapies including prolonged exposure.

Supplementary Hypotheses (see Supplementary *Design Table*): We will conduct supplementary analyses to determine whether experimental and control conditions of DecNef produce differing conscious experiences. These assessments will examine distress levels (**SH_{DNF 1}**), dropout rates (**SH_{DNF 2}**), and the use of conscious strategies (**SH_{DNF 3}**) during DecNef. We will also compare physiological threat responses (**SH_{DNF 4}**) across conditions.

Evaluating consciousness and distress during decoder construction (**H_{DDC}**)

H_{DDC 1}: To evaluate distress during decoder construction, one sample/paired-samples t-test will be performed to assess whether the averaged SUDS during the decoding session is lower than 1) 50 (the strongest load that is still considered endurable), and 2) the baseline score that will be measured before the first fMRI session for the decoding in a similar manner to the SUDs analyses during neurofeedback.

Supplementary Hypotheses (see Supplementary *Design Table*): We will conduct supplementary analyses to determine whether presentations of trauma-related angry male faces during decoder construction produce different conscious experiences (**SH_{DDC 1}**) and physiological threat responses (**SH_{DDC 2}**) from those produced by presentations of happy female faces.

Neural mechanisms of DecNef effects (**H_{NM}**)

H_{NM 1a}: The aim of this analysis is to test whether DecNef's underlying mechanisms map onto extinction learning or counter-conditioning. To do so, we will examine whether core regions in extinction learning, vmPFC^{59,73}, and counter-conditioning, caudate and ventral striatum^{74,75}, are engaged when the angry face pattern is induced in STS during DecNef. To estimate engagement of these areas, we will calculate their relative information transmission with the STS. More specifically, we will calculate the degree to which their activity patterns predict the angry face likelihood in STS activity patterns^{11,13}. To this end, we will train a SLR decoder⁷⁰ so that activation patterns in each ROI reconstruct the STS likelihood. Here, the likelihood in STS will be identical to the likelihood to-be-fed back to participants during DecNef. Leave one-run out cross-validation will be performed to

estimate the validity of the constructed decoder. For each ROI, we will analyze the across-participant Pearson correlation of symptoms reduction (pre-training test versus post-training test (+1 week)) with information transmission. We will then test the difference between these correlations. If DecNef shares its mechanism with exposure therapy, then vmPFC should be positively correlated with symptom reduction. If DecNef shares its mechanism with counter-conditioning, then information transmission in caudate and ventral striatum should be positively correlated with symptom reduction.

To comprehensively examine information transmission in a data-driven manner, we will perform a whole-brain MVPA using a searchlight method, as described in the Supplementary methods. Anticipating the interests of the general audience, we will conduct supplementary analyses with the amygdala, as well as caudate and ventral striatum separated.

H_{NM} 1b: To illuminate mechanisms underlying DecNef, we will further examine how angry face representations induced in STS may lead to changes in the STS connectivity with other brain areas. Specifically, we expect fear/reward circuits to interact with the STS during the successive DecNef trials if DecNef is involved in fear extinction/reinforcement learning. We will conduct a psychophysiological interaction (PPI) analysis⁷⁶ to examine the differential functional connectivity of the STS with other brain regions, as a function of the angry face representation likelihood in the STS, i.e., psychological variable, for each ROI, i.e., vmPFC, and caudate and ventral striatum, as well as for each voxel in the whole brain. To this end, we will test the difference between these correlations.

The GLM model in the PPI analysis will include regressors for the angry face representation likelihood (psychological variable), the z-normalized time course of the seed ROI, and the PPI term (seed time course x likelihood). We will code the psychological variable as 1 or -1 to cover the induction and rest periods of trials (total 12 s) with high (> 50%) versus low (< 50%) likelihood (the disc size reflecting the induction likelihood) respectively, convolved with a canonical HRF. Additional regressors of no interest will be the initial rest period (30 s), the feedback period, and the fixation period (each coded separately for high and low likelihood trials), as well as six motion parameters, the day of DecNef sessions, and the experimental run number. We will use voxels within

the STS that are activated with a liberal threshold (uncorrected $P < .1$) during the induction period (irrespective of angry face representation likelihood) relative to the fixation period. We will conduct these analytical steps within native coordinates of each participant. Then, in the MNI space, we will conduct a group level analysis to obtain a map of voxels showing significant differential connectivity with the STS as a function of angry male face representation likelihood ($P < 0.05$; corrected with permutation procedure⁷⁷).

In a similar vein to **H_{NM} 1a**, we will conduct supplementary analyses with the amygdala, as well as caudate and ventral striatum separated.

T_{NM} 2: There are two possible learning processes underlying the fear-reduction effects of DecNef. One is through effects of exposure, where neural representations of feared stimuli are induced without actual fear or distress. The other is through effects of counter-conditioning, where such representations are paired with reward. In order to dissociate potential effects of exposure from those of counter-conditioning, we will apply a mathematical model, as in our previous study²². The model dissociates DecNef effects derived from exposure therapy ($V_{X(EB)}$) and counter-conditioning ($V_{X(CC)}$) separately, on the basis of the Rescorla-Wagner model and synaptic plasticity rules. Detailed information can be found elsewhere²². Briefly, $V_{X(EB)}$ is assumed to be linearly proportional to the number of “successful” trials, where the likelihood for the target activity pattern is above chance. $V_{X(CC)}$ of each trial is also assumed to be proportional to the induction likelihood of the target pattern multiplied by the amount of the reward that the participant obtains in the trial. Finally, we assume that the DecNef effect is caused by the weighted linear summation of $V_{X(EB)}$ and $V_{X(CC)}$.

Based on these assumptions, the total effect is given as follows:

$$V_{X(EB)} = \sum_i H(L_{target(i)} - L_{control(i)})$$

$$V_{X(CC)} = \sum_i threshold(L_{target(i)} - 0.5)^2$$

Then, we will compare following three models:

$$V_{X(SEV)} = \beta_{EB} V_{X(EB)} \quad (\text{model 1})$$

$$V_{X(SEV)} = \beta_{CC} V_{X(CC)} \quad (\text{model 2})$$

$$V_{X(SEV)} = \beta_{EB} V_{X(EB)} + \beta_{CC} V_{X(CC)} \quad (\text{model 3})$$

where $V_{X(SEV)}$ is symptom reduction from the pre- to post-training test. The β_{EB} and β_{CC} are the coefficients of $V_{X(EB)}$ and $V_{X(CC)}$, respectively. $H(X)$ is the Heaviside step function, which is 1 if $X > 0$ and 0 otherwise. The $\text{threshold}(X) = X$ if $X > 0$, and 0 otherwise. We will test these models using fixed effect models. The model comparison will be conducted using Bayesian Information Criterion (BIC).

Data availability

All data and materials will be made publicly available.

Code availability

All analysis code will be made publicly available.

References

1. Kessler, R. C. *et al.* Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. *Arch. Gen. Psychiatry* **62**, 593–602 (2005).
2. Van Ameringen, M., Mancini, C., Patterson, B. & Boyle, M. H. Post-traumatic stress disorder in Canada. *CNS Neurosci. Ther.* **14**, 171–181 (2008).
3. Koenen, K. C. *et al.* Posttraumatic stress disorder in the World Mental Health Surveys. *Psychol. Med.* **47**, 2260–2274 (2017).
4. Goldstein, R. B. *et al.* The epidemiology of DSM-5 posttraumatic stress disorder in the United States: results from the National Epidemiologic Survey on Alcohol and Related Conditions-III. *Soc. Psychiatry Psychiatr. Epidemiol.* **51**, 1137–1148 (2016).
5. Rauch, S. L., Shin, L. M. & Phelps, E. A. Neurocircuitry Models of Posttraumatic Stress

- Disorder and Extinction: Human Neuroimaging Research—Past, Present, and Future. *Biological Psychiatry* vol. 60 376–382 Preprint at <https://doi.org/10.1016/j.biopsych.2006.06.004> (2006).
6. Foa, E. B., Zoellner, L. A., Feeny, N. C., Hembree, E. A. & Alvarez-Conrad, J. Does imaginal exposure exacerbate PTSD symptoms? *J. Consult. Clin. Psychol.* **70**, 1022–1028 (2002).
 7. Hembree, E. A. *et al.* Do Patients Drop Out Prematurely from Exposure Therapy for PTSD? *J. Trauma. Stress* **16**, 555–562 (2003).
 8. Schnurr, P. P. *et al.* Cognitive behavioral therapy for posttraumatic stress disorder in women: a randomized controlled trial. *JAMA* **297**, 820–830 (2007).
 9. Bradley, R., Greene, J., Russ, E., Dutra, L. & Westen, D. A multidimensional meta-analysis of psychotherapy for PTSD. *Am. J. Psychiatry* **162**, 214–227 (2005).
 10. Lewis, C., Roberts, N. P., Gibson, S. & Bisson, J. I. Dropout from psychological therapies for post-traumatic stress disorder (PTSD) in adults: systematic review and meta-analysis. *Eur. J. Psychotraumatol.* **11**, 1709709 (2020).
 11. Shibata, K., Watanabe, T., Sasaki, Y. & Kawato, M. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* **334**, 1413–1415 (2011).
 12. deBettencourt, M. T., Cohen, J. D., Lee, R. F., Norman, K. a. & Turk-Browne, N. B. Closed-loop training of attention with real-time brain imaging. *Nat. Neurosci.* **18**, (2015).
 13. Amano, K. *et al.* Learning to associate orientation with color in early visual areas by associative decoded fMRI neurofeedback. *Curr. Biol.* **26**, 1861–1866 (2016).
 14. Cortese, A., Amano, K., Koizumi, A., Kawato, M. & Lau, H. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat. Commun.* (2016) doi:10.1038/ncomms13669.
 15. Shibata, K., Watanabe, T., Kawato, M. & Sasaki, Y. Differential Activation Patterns in the Same Brain Region Led to Opposite Emotional States. *PLoS Biol.* **14**, 1–27 (2016).
 16. Watanabe, T., Sasaki, Y., Shibata, K. & Kawato, M. Advances in fMRI Real-Time Neurofeedback. *Trends Cogn. Sci.* **21**, 997–1010 (2017).
 17. Taschereau-Dumouchel, V. *et al.* Towards an unconscious neural reinforcement intervention for common fears. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 3470–3475 (2018).

18. Shibata, K. *et al.* Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback. *Neuroimage* **188**, 539–556 (2019).
19. Cortese, A., Lau, H. & Kawato, M. Unconscious reinforcement learning of hidden brain states supported by confidence. *Nat. Commun.* **11**, 4429 (2020).
20. Koizumi, A. *et al.* Fear extinction without fear : Direct reinforcement of neural activity bypasses the need for conscious exposure. *Nature Human Behaviour* **1**, 1–7 (2016).
21. Cushing, C. A., Lau, H., Kawato, M., Craske, M. G. & Taschereau-Dumouchel, V. A double-blind trial of decoded neurofeedback intervention for specific phobias. *Psychiatry Clin. Neurosci.* **78**, 678–686 (2024).
22. Chiba, T. *et al.* Current Status of Neurofeedback for Post-traumatic Stress Disorder: A Systematic Review and the Possibility of Decoded Neurofeedback. *Front. Hum. Neurosci.* **13**, 233 (2019).
23. Krystal, J. H. *et al.* Adjunctive Risperidone Treatment for Antidepressant-Resistant Symptoms of Chronic Military Service–Related PTSD: A Randomized Trial. *JAMA* **306**, 493–502 (2011).
24. Todd, R. M. *et al.* Soldiers With Posttraumatic Stress Disorder See a World Full of Threat: Magnetoencephalography Reveals Enhanced Tuning to Combat-Related Cues. *Biol. Psychiatry* **78**, 821–829 (2015).
25. Blake, D. D. *et al.* The development of a Clinician-Administered PTSD Scale. *J. Trauma. Stress* **8**, 75–90 (1995).
26. Haxby, J. V. *et al.* A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72**, 404–416 (2011).
27. Guntupalli, J. S. *et al.* A Model of Representational Spaces in Human Cortex. *Cereb. Cortex* **26**, 2919–2934 (2016).
28. Tsuchiya, N. & Koch, C. Continuous flash suppression reduces negative afterimages. *Nat. Neurosci.* **8**, 1096–1101 (2005).
29. Milad, M. R. *et al.* Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biol. Psychiatry* **66**, 1075–1082 (2009).
30. 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).
31. Dickinson A. Dearing, M. Mechanism of learning and motivation. (1979).
 32. Rescorla, R. A. A theory of Pavlovian conditioning : Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory* 64–99 (1972).
 33. Homan, P. *et al.* Neural computations of threat in the aftermath of combat trauma. *Nat. Neurosci.* (2019) doi:10.1038/s41593-018-0315-x.
 34. Olf, M. Sex and gender differences in post-traumatic stress disorder: an update. *Eur. J. Psychotraumatol.* **8**, (2017).
 35. Breslau, N. *et al.* Trauma and posttraumatic stress disorder in the community: the 1996 Detroit Area Survey of Trauma. *Arch. Gen. Psychiatry* **55**, 626–632 (1998).
 36. Creamer, M., Burgess, P. & McFarlane, A. C. Post-traumatic stress disorder: findings from the Australian National Survey of Mental Health and Well-being. *Psychol. Med.* **31**, 1237–1247 (2001).
 37. Santiago, P. N. *et al.* A systematic review of PTSD prevalence and trajectories in DSM-5 defined trauma exposed populations: intentional and non-intentional traumatic events. *PLoS One* **8**, e59236 (2013).
 38. Tolin, D. F. & Foa, E. B. Sex differences in trauma and posttraumatic stress disorder: a quantitative review of 25 years of research. *Psychol. Bull.* **132**, 959–992 (2006).
 39. Cortese, A., Amano, K., Koizumi, A., Lau, H. & Kawato, M. Decoded fMRI neurofeedback can induce bidirectional confidence changes within single participants. *Neuroimage* **149**, 323–337 (2017).
 40. Rance, M. *et al.* Time course of clinical change following neurofeedback. *Neuroimage* (2018) doi:10.1016/j.neuroimage.2018.05.001.
 41. Megumi, F., Imamizu, H., Yamashita, A. & Kawato, M. Functional MRI neurofeedback training on connectivity between two regions induces long-lasting changes in intrinsic functional network. *Front. Hum. Neurosci.* **9**, (2015).
 42. Weiss, D. S. The Impact of Event Scale: Revised. in *Cross-Cultural Assessment of Psychological Trauma and PTSD* (eds. Wilson, J. P. & Tang, C. S.-K.) 219–238 (Springer US, Boston, MA,

- 2007).
43. Beck, A. T., Steer, R. A., Brown, G. K. & Others. Manual for the beck depression inventory-II. Preprint at (1996).
 44. Spielberger, C. D. State-Trait Anxiety Inventory for Adults. (1983) doi:10.1037/t06496-000.
 45. Shin, L. M. *et al.* A functional magnetic resonance imaging study of amygdala and medial prefrontal cortex responses to overtly presented fearful faces in posttraumatic stress disorder. *Arch. Gen. Psychiatry* **62**, 273–281 (2005).
 46. Milad, M. R. *et al.* Presence and acquired origin of reduced recall for fear extinction in PTSD: Results of a twin study. *J. Psychiatr. Res.* **42**, 515–520 (2008).
 47. Yehuda, R. *et al.* Post-traumatic stress disorder. *Nat Rev Dis Primers* **1**, 15057 (2015).
 48. Lanius, R. A. *et al.* Emotion modulation in PTSD: Clinical and neurobiological evidence for a dissociative subtype. *Am. J. Psychiatry* **167**, 640–647 (2010).
 49. Chiba, T. *et al.* A reciprocal inhibition model of alternations between under-/overemotional modulatory states in patients with PTSD. *Mol. Psychiatry* (2020) doi:10.1038/s41380-020-0827-0.
 50. Grasser, L. R. *et al.* Skin conductance response to trauma interview as a candidate biomarker of trauma and related psychopathology in youth resettled as refugees. *Eur. J. Psychotraumatol.* **13**, 2083375 (2022).
 51. White, S. F., Costanzo, M. E., Blair, J. R. & Roy, M. J. PTSD symptom severity is associated with increased recruitment of top-down attentional control in a trauma-exposed sample. *Neuroimage Clin* **7**, 19–27 (2015).
 52. Weathers, F. W., Keane, T. M. & Davidson, J. R. Clinician-administered PTSD scale: a review of the first ten years of research. *Depress. Anxiety* **13**, 132–156 (2001).
 53. Bryant, R. A. *et al.* Trajectory of post-traumatic stress following traumatic injury: 6-year follow-up. *Br. J. Psychiatry* **206**, 417–423 (2015).
 54. Nash, W. P. *et al.* Posttraumatic stress in deployed Marines: prospective trajectories of early adaptation. *J. Abnorm. Psychol.* **124**, 155–171 (2015).
 55. Peelen, M. V., Atkinson, A. P. & Vuilleumier, P. Supramodal representations of perceived

- emotions in the human brain. *J. Neurosci.* **30**, 10127–10134 (2010).
56. Zhang, H. *et al.* Face-selective regions differ in their ability to classify facial expressions. *Neuroimage* **130**, 77–90 (2016).
57. Hesselmann, G. & Malach, R. The link between fMRI-BOLD activation and perceptual awareness is stream-invariant in the human visual system. *Cereb. Cortex* **21**, 2829–2837 (2011).
58. Fogelson, S. V., Kohler, P. J., Miller, K. J., Granger, R. & Tse, P. U. Unconscious neural processing differs with method used to render stimuli invisible. *Front. Psychol.* **5**, 1–11 (2014).
59. 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. U. S. A.* **110**, 20040–20045 (2013).
60. Zarnhofer, S. *et al.* The Influence of verbalization on the pattern of cortical activation during mental arithmetic. *Behav. Brain Funct.* **8**, 13 (2012).
61. Seijas, O. *et al.* Ocular dominance diagnosis and its influence in monovision. *Am. J. Ophthalmol.* **144**, 209–216 (2007).
62. Sripada, R. K. & Rauch, S. A. M. Between-session and within-session habituation in Prolonged Exposure Therapy for posttraumatic stress disorder: a hierarchical linear modeling approach. *J. Anxiety Disord.* **30**, 81–87 (2015).
63. Albers, C. & Lakens, D. When power analyses based on pilot data are biased: Inaccurate effect size estimators and follow-up bias. *J. Exp. Soc. Psychol.* **74**, 187–195 (2018).
64. Ros, T. *et al.* Consensus on the reporting and experimental design of clinical and cognitive-behavioural neurofeedback studies (CRED-nf checklist). *Brain* **143**, 1674–1685 (2020).
65. Green, P. & MacLeod, C. J. SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* **7**, 493–498 (2016).
66. McLean, C. P., Levy, H. C., Miller, M. L. & Tolin, D. F. Exposure therapy for PTSD: A meta-analysis. *Clin. Psychol. Rev.* **91**, 102115 (2021).
67. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. *Behav. Res. Methods* **41**, 1149–1160 (2009).
68. O’Brien, P. C. & Fleming, T. R. A multiple testing procedure for clinical trials. *Biometrics* **35**,

- 549–556 (1979).
69. Marshall, A. D., Occhipinti, S. & Loxton, N. J. Tipping the analytical scales, investigating the use of frequentist equivalence analyses in psychology: a scoping review. *Qual. Quant.* **58**, 2929–2955 (2024).
 70. 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).
 71. Lissek, S. *et al.* Classical fear conditioning in the anxiety disorders: A meta-analysis. *Behav. Res. Ther.* **43**, 1391–1424 (2005).
 72. Galeano, D., Li, S., Gerstein, M. & Paccanaro, A. Predicting the frequencies of drug side effects. *Nat. Commun.* **11**, 4575 (2020).
 73. Do-Monte, F. H., Manzano-Nieves, G., Quinones-Laracuente, K., Ramos-Medina, L. & Quirk, G. J. Revisiting the Role of Infralimbic Cortex in Fear Extinction with Optogenetics. *Journal of Neuroscience* **35**, 3607–3615 (2015).
 74. Haruno, M. & Kawato, M. Different neural correlates of reward expectation and reward expectation error in the putamen and caudate nucleus during stimulus-action-reward association learning. *J. Neurophysiol.* **95**, 948–959 (2006).
 75. O’Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304**, 452–454 (2004).
 76. Friston, K. J. *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6**, 218–229 (1997).
 77. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).

Acknowledgements

This study is funded by the Innovative Science and Technology Initiative for Security (JPJ004596), ATLA, AMED (JP20dm0307008), and KDDI collaborative research contract. The funders had no role

in the study design, data collection and analysis, the decision to publish, or in preparation of the manuscript.

We thank Kaori Nakamura and Chika Hosomi for their help with scheduling and conducting the pilot experiment.

Author contributions

TC, HL, MK, and AK conceived the research. TC, KI, JET, AC, TK, KA, HL, MK, AK, and TK designed the research. TC developed the decoding and DecNef task in consultation with AC, TK, KA, BS, and AK. TC analyzed the pilot data in consultation with KA, AK, and MK. TC conducted the power analysis in consultation with KA, AK, and MK. TC, IK, KA, HL, MK, AK, and KT developed the analysis plan. TC and JET prepared the manuscript with feedback from all co-authors. TC, KI, JET, SB, HT, MS, and TK coordinated the implementation of the project. TC, KI, JET, KN, SB, HT, MS, and TK will contribute to data collection. All co-authors will review and approve the final manuscript.

Competing interests

Author Ai Koizumi is employed by Sony Computer Science Laboratories, Inc. The other authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. Author Mitsuo Kawato is an inventor of patents owned by the Advanced Telecommunications Research Institute International related to the present work (PCT/JP2012/078136 [WO2013/069517] and PCT/JP2014/61543 [WO2014/178322]). This study was funded by KDDI Corporation, which played no scientific role in the project. There is nothing else to disclose.