nature human behaviour

LETTERS

PUBLISHED: 21 NOVEMBER 2016 | VOLUME: 1 | ARTICLE NUMBER: 0006

Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure

Ai Koizumi[1,2,3†], Kaoru Amano[3†], Aurelio Cortese[1,3,4,5†], Kazuhisa Shibata[1,6], Wako Yoshida[3,7,8], Ben Seymour[3,7,8*], Mitsuo Kawato[1,3,4*] and Hakwan Lau[2,5,9*]

Fear conditioning is a fundamentally important and preserved process across species[1,2]. In humans it is linked to fear-related disorders such as phobias and post-traumatic stress disorder (PTSD)[3,4]. Fear memories can be reduced by counter-conditioning, in which fear conditioned stimuli (CS+s) are repeatedly reinforced with reward[5] or with novel non-threatening stimuli[6]. However, this procedure involves explicit presentations of CS+s, which is itself aversive before fear is successfully reduced. This aversiveness may be a problem when trying to translate such experimental paradigms into clinical settings[7]. It also raises the fundamental question as to whether explicit presentations of feared objects is necessary for fear reduction[1,8]. Although learning without explicit stimulus presentation has been previously demonstrated[9–12], whether fear can be reduced while avoiding explicit exposure to CS+s remains largely unknown. One recently developed approach employs an implicit method to induce learning by reinforcing stimulus-specific neural representations using real-time decoding of multivariate functional magnetic resonance imaging (fMRI) signals[13–15] in the absence of stimulus presentation; that is, pairing rewards with the occurrences of multi-voxel brain activity patterns matching a specific stimulus (decoded fMRI neurofeedback (DecNef)[13,15]). It has been shown that participants exhibit perceptual learning for a specific visual stimulus feature through DecNef, without being given any strategy for the induction of specific neural representations, and without awareness of the content of reinforced neural representations[13]. Here we examined whether a similar approach could be applied to counter-conditioning of fear. We show that we can reduce fear towards CS+s by pairing rewards with the activation patterns in visual cortex representing a CS+, while participants remain unaware of the content and purpose of the procedure. This procedure may be an initial step towards novel treatments for fear-related disorders such as phobia and PTSD, via unconscious processing.

In our experiment (Fig. 1a), participants first acquired a fear response to two visual stimuli (the target CS+ and control CS+) in the acquisition session, and went through three daily sessions of neural reinforcement by DecNef during which only the activation patterns for the target CS+ were reinforced to reduce its associated fear, without physical presentations of the target CS+. On the following day, participants were presented with two CS+s and their fear response was measured in the test session. The details of these sessions are as follows.

In the acquisition session, Pavlovian aversive conditioning was performed in 17 healthy participants by pairing two visual cues (CS+s) with uncomfortable but tolerable electrical shocks (Fig. 1a). The CS+s were visual stimuli—vertical gratings of different colours (red and green), allowing them to be distinguished by multi-voxel pattern decoding in the visual cortex, V1/V2. Unbeknownst to the participant, one of the two CS+s was designated as the target CS+, meaning we intended to have its associated fear level subsequently reduced via DecNef. The other CS+ was designated as the control CS+, as a baseline comparison. The choice of stimuli was based on a previous study using similar procedures[15]. Towards the end of the acquisition session (Fig. 1a–d), both CS+s induced elevated skin conductance responses (SCRs) in comparison to an unreinforced cue (CS−), indicating successful conditioning of fear (Fig. 2a and Supplementary Fig. 1).

The activation patterns discriminating the target and non-target control CS+s were determined by conventional multivariate decoding in another session (that is, the fMRI session for multi-voxel pattern analysis (MVPA), see Supplementary Methods) before the acquisition session (mean of decoding accuracy estimated with leave-one-out cross-validation; 72.1% ± 9.2 s.d.), so that the likelihood that the target CS+ is represented in V1/V2 activation patterns could be calculated in real time during the subsequent neural reinforcement sessions.

In the three daily neural reinforcement sessions (Fig. 1e), participants viewed achromatic visual gratings. They were asked to use any mental strategy they liked to try and increase the diameter of a disc on a monitor to earn monetary reward. They were unaware that the diameter was proportional to how much multi-voxel patterns in the visual cortex resembled target CS+ representations. The likelihood of occurrence of the target CS+ pattern was associated with the magnitude of the reward, leading to counter-conditioning for the target CS+[5]. We hypothesized that this would reduce the ability of the target CS+ to elicit fear responses relative to the control CS+.

On day 1 of the neural reinforcement session, the occurrence of the target CS+ pattern was around chance level (50.2 ± standard error (s.e.) 4.1%; $t(16) = 0.05$, $P = 0.96$, paired t-test, two-tailed). Subsequently, target CS+ likelihood exceeded chance level on day

[1]Department of Decoded Neurofeedback, ATR Computational Neuroscience Laboratories, 2-2-2, Hikaridai, Seika-cho, Sorakugun, Kyoto, 619-0288, Japan. [2]Department of Psychology, Columbia University, 1190 Amsterdam Avenue 370 Schermerhorn Extension MC:5501, New York 10027, USA. [3]Center for Information and Neural Networks (CiNet), NICT, 1-4 Yamadaoka, Suita City, Osaka, 565-0871, Japan. [4]Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama, Ikoma Nara, 630-0192, Japan. [5]Department of Psychology, UCLA, Box 951563, Los Angeles, California 90095-1563, USA. [6]Department of Psychology, Graduate School of Environmental Studies, Nagoya University, Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan. [7]Department of Neural Computation for Decision-making, ATR Cognitive Mechanisms Laboratories, 2-2-2, Hikaridai, Seika-cho, Sorakugun, Kyoto, 619-0288, Japan. [8]Department of Engineering, University of Cambridge, Trumpington Street, Cambridge CB2 1PZ, UK. [9]Brain Research Institute, UCLA, Box 951761, Los Angeles, California 90095-1761, USA. †These authors contributed equally to this work.
*e-mail: kawato@atr.jp; bjs49@cam.ac.uk; hakwan@gmail.com

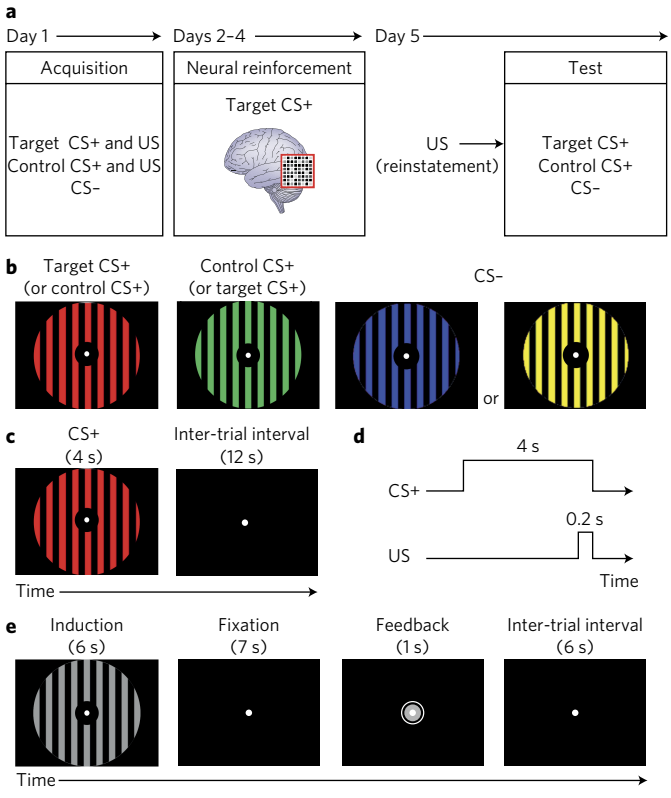

Figure 1 | Overall experimental design. a, After retinotopy and the fMRI session for MVPA (see Supplementary Methods), participants went through 5 days of main experimental sessions in the MRI scanner. **b**, The stimuli used as CSs were coloured vertical grating patterns, with the choice of colours (red and green) for the target CS+ and control CS+ counterbalanced across participants. A choice of colour (blue or yellow) for the CS− was also counterbalanced across participants. **c**, Timeline for a single trial in the acquisition or test session. In an acquisition trial, both target and control CS+s (red/green) were paired with the US (electric shock). **d**, CS+s were paired with the co-terminating US during acquisition at the contingency rate of 38%. **e**, During a neural reinforcement trial, participants were required to somehow regulate their brain activity upon seeing a grey vertical grating (induction cue). The size of the disc during the feedback period indicated the online-calculated likelihood of the target CS+ patterns in V1/V2. The disc size was proportional to the amount of monetary reward earned.

2 (58.9 ± 3.1%; $t(16) = 2.89$, $P = 0.01$) and day 3 (57.2 ± 3.3%; $t(16) = 2.20$, $P = 0.04$), providing evidence of successful DecNef of the target CS+ pattern. The effect of day was significant (analysis of variance (ANOVA); $F(2, 15) = 3.62$, $P = 0.038$), which was primarily due to the increase of target CS+ likelihood from day 1 to day 2 ($P = 0.089$, Bonferroni-corrected). Importantly, although the overall likelihood of target CS+ occurrence was modest, over the 3 days there was a sufficiently large variability in the trial-wise induction likelihood within each participant (across-participant mean of s.d. for the likelihood, 45.1%). Therefore, participants were exposed to a full range of contingency between the induction likelihood and its corresponding reward, which is critical for the facilitation of reinforcement learning. More detailed progress of the occurrence of the target CS+ pattern is shown in Supplementary Fig. 2.

A post-experimental questionnaire confirmed that participants remained unaware of the association between the disc's diameter and occurrence of the target CS+ representation, and did not consciously use strategies related to colour or grating (that is, appropriate imagery-based strategies, Supplementary Table 1). Neither were they able to guess the identity of the target CS+ versus control CS+ in a forced choice question afterwards (62.5% accuracy, chi-squared test $\chi^2 = 0.780$, $P = 0.377$). This result is in agreement with previous studies using a similar DecNef procedure[13,15]. Furthermore, there were also no reports of subjective fear during the neural reinforcement session (Supplementary Table 1), and SCRs during these sessions were significantly less when compared with those associated with the CS+s at the end of the acquisition session ($t(16) = 4.218$, $P = 0.001$, paired t-test, two-tailed), and did not correlate with induced target CS+ pattern likelihood (Fig. 2b,c).

Next, during the test session, four unsignalled shocks (unconditioned stimuli, USs) were presented to reactivate the fear memory (that is reinstatement; Fig. 1a) to assess potential fear reduction for the target CS+[16]. Given that the test session was performed several days after acquisition, this reinstatement procedure helped to ensure observable fear responses for the baseline condition (that is, the control CS+), to allow for a meaningful comparison. During the test session, the two CS+s and the CS− were presented alone to evaluate the associated fear response (SCR). Critically, we found that the SCR to the target CS+ was significantly reduced when compared with the control CS+ ($t(16) = 2.630$, $P = 0.018$, paired t-test, two-tailed) (Fig. 2a and Supplementary Fig. 1), suggesting that fear towards the CS+ was reduced only when it went through the DecNef procedure, where occurrence of the target CS+ was paired with reward, effectively counter-conditioning the previously acquired fear. Interestingly, the magnitude of this effect was similar to what was observed following conventional extinction procedures[17,18], even though the underlying mechanism may be different, not least as participants were unaware of the occurrence of the target CS+ representations during neural reinforcement.

We also recorded fMRI responses during the acquisition and test sessions, focusing in particular on responses in the amygdala and the ventral medial prefrontal cortex (VMPFC), which have been implicated in acquisition and extinction of fear memory[19,20]. The amygdala showed significant responses after conditioning for both CS+s, but a significant reduction in response to the target CS+ compared with the control CS+ in the test session (Fig. 3a), mirroring the specific pattern of fear reduction seen in the SCRs. VMPFC responses were significantly negative for both CS+s during acquisition as previously shown[18,19,21], but significantly less positive during neural reinforcement and test sessions for the target CS+ (Fig. 3b). While responses in the amygdala and VMPFC were reduced for the target CS+ following the neural reinforcement sessions, during these sessions the trial-wise response level in these regions did not correlate with the likelihood of target CS+ induction, suggesting that fear memory was not strongly reactivated by the spontaneous occurrence of target CS+ patterns in the visual cortex (Supplementary Fig. 3). The average responses in V1/V2 were similar between target and control CS+s both before and after the neural reinforcement sessions (Supplementary Fig. 4), undermining the possibility that the differential responses in the amygdala and VMPFC for the two CS+s during the test session were merely due to the altered visual processing for target CS+.

Although the likelihood of target CS+ occurrence was estimated selectively from the activation patterns in V1/V2 during the neural reinforcement sessions, it remains unclear whether such information reflecting the target CS+ was confined to V1/V2 or whether other brain regions were engaged to act in concert with V1/V2. To examine whether any brain regions outside the visual cortex were engaged, we conducted the whole-brain searchlight MVPA[22], which quantitatively measures the degree to which the activation patterns of other brain areas could predict the likelihood that the target CS+ patterns were induced in V1/V2. Such predictability of the target CS+ likelihood in V1/V2 would reflect 'information transmission' from V1/V2 to other areas[13,15]. For more detailed advantages of this approach, see

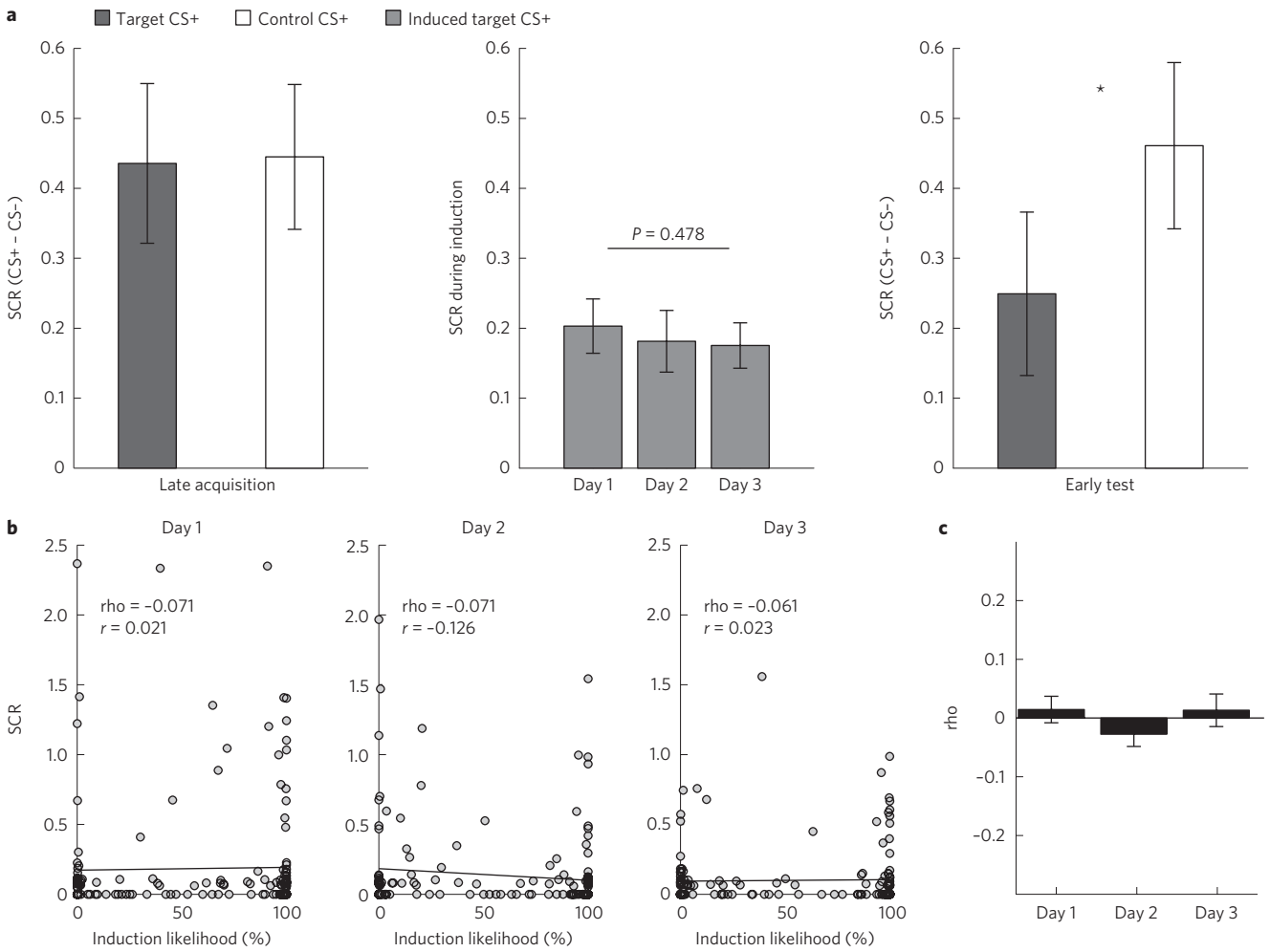

Figure 2 | Reduction of fear response as measured by SCR. a, In late acquisition (last two trials), participants developed positive responses for both target and control CS+s. During the neural reinforcement session (middle panel), such responses were lower than those associated with the CS+s at the end of acquisition ($t(16) = 4.218$, $P = 0.001$, paired t-test, two-tailed). In early test (first two trials; see Fig. 1a), the response to the target CS+ was reduced compared with the control CS+ ($t(16) = 2.630$, $P = 0.018$, paired t-test, two-tailed). *, $P < 0.05$. **b**, During neural reinforcement, trial-wise correlation between SCR and induction likelihood (that is, the degree to which activity in V1/V2 resembled the multi-voxel pattern for the target CS+) was negligible, indicating that the induced target CS+ did not lead to a fear response. Shown are scatter plots of a representative participant, where each dot represents a single induction trial. Rho denotes the Spearman correlation coefficient, while the Pearson correlation coefficient r is also shown as a reference. The solid line represents the result of a least-square regression. **c**, Plotted is the Fisher-transformed correlation coefficients for each day of the neural reinforcement session averaged across participants. Error bars in panels **a** and **c** represent s.e.

Supplementary Methods. Specifically, the searchlight MVPA estimated the degree to which the trial-by-trial induction likelihood of target CS+ in V1/V2 can be reconstructed from the multi-voxel patterns within a spherical region of interest (ROI; radius = 15 mm) centred at each voxel (see Supplementary Methods). The analysis revealed that the target CS+ likelihood (that is, the likelihood of red or green grating) was transmitted to many visual areas during the fMRI session for MVPA when the red and green gratings were physically presented (Fig. 4a), but it was largely confined to V1/V2 during the neural reinforcement sessions, with a few exceptions (Fig. 4b). In particular, the striatal area (caudate nucleus) had significant information transmission from V1/V2, in keeping with its possible role in reinforcement learning[23,24]. Such engagement of the striatal area was further supported by a psychophysiological interaction (PPI) analysis[25], showing enhanced functional connectivity between V1/V2 and the striatum, as a function of the increase of target CS+ likelihood in V1/V2 (Supplementary Methods and Supplementary Fig. 6). These results suggest that, besides the visual cortex where target CS+ was induced, the striatal area was engaged to some extent in the neural reinforcement sessions.

Previous studies with conventional extinction procedures have shown a role for the VMPFC for successful extinction[19,20]. Yet, the aforementioned searchlight analysis revealed no notable information transmission from V1/V2 to VMPFC, suggesting that the VMPFC may not be actively engaged in DecNef on average within the participants group. Moreover, across participants, the degree to which VMPFC patterns predicted V1/V2 patterns (information transmission) was negatively correlated with the success of fear reduction (Spearman's rho = −0.522, $P = 0.034$; Fig. 4c), suggesting that there was less VMPFC engagement for participants with more successful fear reduction. These results, as well as the whole-brain analysis examining the correlation between information transmission and fear reduction (Supplementary Fig. 5), suggest that VMPFC disengagement may have led to larger reduction of fear. Thus, our results are consistent with the view that the fear reduction observed here depended on a possibly different mechanism to that of conventional extinction procedures[20].

To summarize, we provide behavioural and neurophysiological evidence that rewards can be directly paired with the patterns of neural activity in the visual cortex to facilitate the reduction of

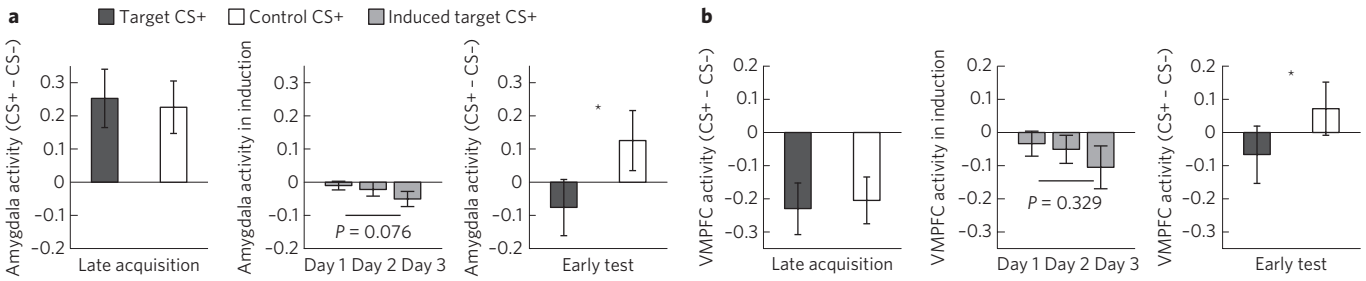

Figure 3 | Brain activity in the amygdala and VMPFC. a,**b**, The same labelling is used as in Fig. 2a except that the dependent measure here is the average level of activity (arbitrary units) in the amygdala (**a**) and VMPFC (**b**). Amygdala activity was reduced for the target CS+ compared with the control CS+ ($t(16) = 2.21$, $P = 0.042$, paired t-test, two-tailed) in early test (**a**). VMPFC activity was reduced for the target CS+ relative to the control CS+ ($t(16) = 2.13$, $P = 0.049$, paired t-test, two-tailed) in early test (**b**). While activity in the amygdala and VMPFC numerically decreased across the 3 days of the neural reinforcement sessions (middle panels of **a** and **b**, respectively), these decreases were not significant ($P = 0.076$ and $P = 0.329$, respectively). Asterisks in panels **a** and **b** indicate $P < 0.05$. Error bars represent s.e. Also see Supplementary Fig. 3.

fear. This DecNef procedure bypasses the requirement for physical visual presentation of a CS. As such, the procedure represents direct counter-conditioning of neural activity based on the fluctuations of the information content in the visual cortex activity. The results show that counter-conditioning of visual-cortical-based fear is sufficient to bring about the reduction of behavioural and amygdalar fear responses, even in the absence of the participants' awareness of the content of neural induction and purpose of the procedure[13,15,26].

The pattern of activity in the VMPFC contrasts with that observed in conventional studies of extinction, whereby it typically exhibits enhanced responses to extinguished fear cues[18,19]. In conventional extinction, fear memory is thought to be inhibited by extinction learning of a context-specific 'safety' state, mediated by the VMPFC, via the exertion of an inhibitory influence on amygdala-based fear memories[18,19]. Similarly, in counter-conditioning, presenting rewards in place of previously paired aversive outcomes reduces conditional fear by a putatively similar inhibitory mechanism[5]. The fact that we see reduced VMPFC activity in association with a reduced fear response to the target CS+ (Fig. 3b) suggests that the mechanism of fear reduction demonstrated here may differ from conventional extinction procedures[2,18,19,21]. This is further supported by the fact that greater interaction between the visual cortex and VMPFC during the neural reinforcement sessions as measured by information transmission is associated with less fear reduction (Fig. 4c and Supplementary Fig. 5).

Given that successful fear reduction requires VMPFC involvement during conventional extinction training[20], this result hints at the possibility that VMPFC disengagement may be key to the success of our procedure[21]. Consistent with this view, some previous studies have also suggested that disengagement of the VMPFC could in some cases counter-intuitively lead to more robust extinction of fear. For example, a human lesion study has shown that VMPFC damage prevents development of PTSD after traumatic experiences[27], implying robust fear extinction. Similarly, infant rats achieve more robust fear extinction than adult rats despite the fact that infant rats do not rely on medial prefrontal areas of the cortex, including an area homologous to the human VMPFC[28]. Such VMPFC disengagement may have made the effect of DecNef robust, potentially making the fear reduction effect context-invariant (see Supplementary Fig. 7).

Although the exact neural mechanism underlying this neural counter-conditioning procedure remains to be further elucidated, it is likely to critically involve the striatum, an area implicated in

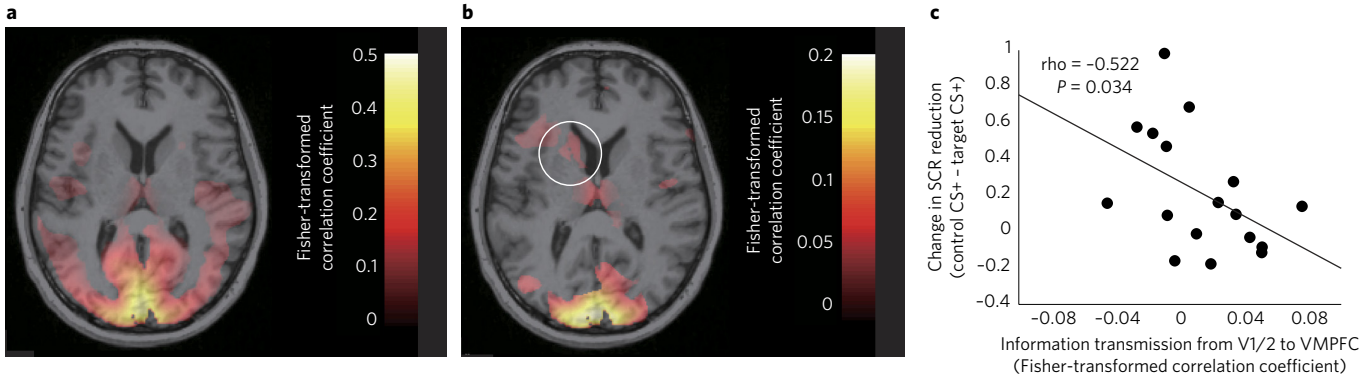

Figure 4 | Engagement of striatum and disengagement of VMPFC during neural reinforcement. a, Whole-brain searchlight MVPA quantitatively evaluated information transmission from V1/V2 to the whole brain during the fMRI session for MVPA. Information transmission was estimated as the correlation between the target CS+ likelihood in V1/V2 and its reconstructed value obtained from the multi-voxel patterns within each searchlight spherical ROI (radius = 15 mm). Information was transmitted mostly within, although not confined to, many visual areas. This result ensures the sensitivity and power of the whole-brain searchlight MVPA to detect information transmission outside V1/V2, if there is any. **b**, Whole-brain MVPA during the neural reinforcement sessions. Overall, information transmission from V1/V2 was mostly confined to the early visual cortex. However, there was a significant transmission to the striatum, mostly caudate, consistent with its role in reinforcement learning[22,23]. The white circle highlights the significant information transmission in striatum. In **a** and **b**, Fisher-transformed correlation coefficients for the significant voxels are shown ($P < 0.05$; multiple corrections with permutation procedure[42]). **c**, Disengagement of VMPFC and successful fear reduction. Less information transmission between V1/V2 and VMPFC during neural reinforcement was related with larger reduction of SCR (control CS+ – target CS+) in early test (Spearman's rho = –0.522, $P = 0.034$). Each data point corresponds to each participant. The solid line represents a least-square regression line. See also Supplementary Fig. 5.

reinforcement learning[23,24]. During the neural reinforcement sessions, occurrence of the target CS+ in V1/V2 could be predicted from the activation patterns of the striatum, mainly the caudate nucleus (Fig. 4b). Similarly, the functional coupling between V1/V2 and striatum was modulated by the occurrence of the target CS+ in V1/V2 (Supplementary Fig. 6). Activation in striatal areas has been found in previous studies on neurofeedback training[29], which suggests that neurofeedback procedures in general may rely to some extent on striatal reinforcement.

While this is a human study to show that fear can be reduced by directly pairing reward with the induced activation patterns for CS+ in visual cortex, an animal study has previously shown that pairing reward with the optogenetically reactivated hippocampal traces of a fear conditioned location (that is, CS+) could reduce the associated fear response[12]. The study further showed that, after the optogenetic reactivation, the hippocampal memory trace lost its ability to activate the previously associated amygdala neurons. Similarly, in our study, reinforcing the target CS+ pattern in the visual cortex may have weakened its previously acquired fear association with the amygdala, resulting in reduction of the amygdalar response to the physically presented target CS+ (Fig. 3a).

Overall, the fear reduction effect achieved with the DecNef procedures appears robust, as it tolerated the challenge provided by reinstatement to reactivate the fear memory[16]. Tolerance of the fear reduction effect to such a challenge is ecologically meaningful, as it may capture resistance to relapse of fear in real-life fear memories[16]. Here we used reinstatement primarily to allow a sufficient magnitude of fear response after the multiple days of the training period, but in principle it would be interesting to look at the effect on the fear response without reinstatement. From a theoretical perspective, showing fear reduction both with and without reinstatement would allow better clarification as to whether fear reduction was driven primarily by an inhibition of the original memory[16]. As fear reduction effects are often weakened after reinstatement[30–33], it is likely that the observed fear reduction effect would still have been present even if tested without reinstatement, but future studies could directly investigate this.

Our current findings may eventually benefit clinical treatments for fear-related disorders. From a translational perspective, the traditional application of fear extinction to anxiety-related disorders faces several challenges. One difficulty in applying traditional associative learning concepts to exposure therapy is that some participants may not comply with explicit encounters with feared objects in the first place[7], because of their intrinsic aversiveness. Here, the induction of brain patterns are not accompanied by conscious awareness of the relevant content, and so this may alleviate the problem of patient attrition. However, to achieve this goal, one needs to pass several technical hurdles. For instance, for ecological validity, one would need to develop procedures to decode images with rich real-life content, and the learning of the relevant multi-voxel patterns for individual patients would also need to be done without conscious presentation to the patients. These may be overcome by building decoders with subliminal presentation, or adapting them from other individuals' brain activity[34]. Despite these challenges, the present results hopefully represent an initial step towards a potential new avenue for treatment.

Methods

The entire experiment consisted of the five main sessions (Fig. 1a), acquisition, neural reinforcement × 3 and test, which were conducted after the two preparatory sessions, retinotopy, and the fMRI session for MVPA. All sessions were conducted on different days separated by at least 24 h. All of the experiments were conducted with fMRI measurement.

Participants. Twenty-four participants (15 males, 23.2 ± 2.5 years old) were initially enrolled. Seven participants were eliminated prior to participating in the neural reinforcement sessions because six of them failed to show a measurable fear response (see 'Acquisition session') and one participant did not complete the acquisition session due to excessive anxiety. The remaining 17 participants (11 males, 23.5 ± 2.8 mean years old) completed all experimental sessions. We predetermined the number of participants to complete all the sessions based on our pilot study. Participants gave written consent prior to participating in each session. The study was approved by the Institutional Review Board of ATR, Japan.

Retinotopy session. We first conducted a standard retinotopic mapping experiment to localize V1/V2 in each participant[35] (see Supplementary Methods).

fMRI session for MVPA. The aim of the fMRI session for MVPA was to obtain fMRI data for constructing a decoder to classify the activation patterns in V1/V2 (see 'Retinotopy session') evoked by isoluminant red versus green vertical gratings (Fig. 1b), which were to serve as the CS+s in the subsequent acquisition session. The decoder was used in the following neural reinforcement sessions to evaluate the trial-by-trial likelihood that participants could induce brain activation patterns for the target CS+ (red or green grating, counterbalanced across participants). During this session each trial consisted of a fixation disc (6 s), followed by a grating that flickered at 0.5 Hz (6 s total).

The preprocessed fMRI signals from the localized V1/V2 subregions were then used to construct a decoder to classify the activation patterns for red versus green grating (see Supplementary Information). We used sparse logistic regression (SLR)[36] to automatically select the voxels that were relevant for classification. We trained the decoder using 192 data points obtained from 192 trials (across all 12 fMRI runs).

Acquisition session. The aim of the acquisition session was to establish a fear memory for red and green gratings (conditioned stimulus (CS+)) by pairing them with an uncomfortable but tolerable electric shock (unconditioned stimulus (US)). These two gratings were identical to the fMRI session for MVPA. A grating with a novel colour (blue or yellow) was introduced as the CS−, which was never paired with the US. The choice of colour for the CS− was counterbalanced across participants orthogonal to the choice of colour for the target CS+ (red or green) (for example, approximately half of the participants with green target CS+ were assigned blue CS−, while the other half of participants were assigned yellow CS−). With such counterbalancing, we avoided a situation where the activation patterns for CS− would always be more similar to one of the CS+s (for example, the target) than to the other CS+ (for example, the control). The experimenter was not blind to the colour assignments in the acquisition session or subsequent sessions, as handling with blindness was difficult due to the complexity of our procedures. Two CS+s were presented either with or without the US (five and eight times, respectively), and the CS− was always presented without the US (eight times)[21]. Trial order was randomized. Each trial started with a presentation of a CS (4 s) followed by a fixation disc (12 s). On trials with the US, a CS+ co-terminated with a burst of electric shocks (36 impulses across 200 ms total). Skin conductance response (SCR) was recorded using BrainAmp Ag/AgCl sintered MR electrodes (Brain Products) attached to the distal phalanges of the index and middle fingers of the right hand. Among 24 participants who completed the acquisition session, 6 participants were excluded because no SCR was detected for the CS+s. Another participant did not complete the acquisition session due to excessive anxiety. The remaining 17 participants proceeded to the subsequent sessions. To estimate fear response in late acquisition, we calculated the mean SCR during the last two trials for each CS (Fig. 2a, left).

Neural reinforcement session. The neural reinforcement sessions were conducted for three consecutive days. The aim of the session was to repetitively induce V1/V2 activation patterns for one of the CS+s (red, $N = 9$; green, $N = 8$) without participants' awareness of the induced target CS+. We reinforced participants with monetary reward for inducing the patterns for one of the CS+s, given the capacity of reward to reinforce behaviour[37] as well as neural activity[38]. Participants were not attached to an electrode for electric shock.

Each trial had a sequence of an induction period (6 s), a fixation period (7 s), a feedback period (1 s), and an inter-trial interval (6 s) (Fig. 1e). During the induction period, participants were instructed to somewhat regulate their brain activity so as to maximize the size of the white disc that served as feedback. Feedback was presented after 6 s of the fixation period following the induction period. In the induction period, a grey vertical grating was presented. The grey grating flickered at 0.5 Hz (7 s total; three repetitions of a grating (1.5 s) and a fixation (0.5 s)). Participants were not informed as to what the feedback disc size represented (that is, target CS+ likelihood in V1/V2).

The V1/V2 activation pattern during the induction period was analysed online to estimate the likelihood that the currently achieved brain activation patterns represented the patterns for the target CS+ (red or green) that were previously decoded from the fMRI session for MVPA. A haemodynamic delay of 6 s was taken into account.

Test session. A day after the last day of the neural reinforcement session, we conducted the test session to measure fear responses to the target CS+, control

CS+ and CS−. On the basis of our preliminary studies, we presented four unsignalled USs before the test session to activate the fear memory (that is, reinstatement) in a similar manner as a previous study[21]. Following reinstatement, each CS was presented 11 times in a semi-randomized order: CS− was always presented on the first trial to capture irrelevant SCR due to orienting effect[21]. Data corresponding to this first CS− were discarded from the subsequent analyses. A trial sequence was identical to the acquisition session, except that there was no trial with the US. SCR was recorded in the same manner as in the acquisition session.

MRI parameters. Participants were scanned in a 3T MRI scanner (Trio, Siemens) with a head coil at the ATR Brain Activation Imaging Center. See Supplementary Methods for more detailed parameters.

Definition of ROIs. Along with SCR, we measured the response in the amygdala and VMPFC to track the fear-related activity in these areas with fMRI. To determine the amygdala ROI, we first defined the anatomical boundary of the amygdala with freesurfer segmentation, and selected voxels within this anatomical boundary that showed a greater response for all US trials and the last two trials of each CS+ (that is, fear-relevant trials) relative to fixation during the acquisition session. To define the VMPFC ROI, we first created an anatomical mask of a sphere with 15 mm radius centred around previously reported MNI coordinates [0, 40, −12][39], which was estimated based on the representative literature[18,19,40]. We then selected voxels within the spherical ROI that showed a smaller response for all US trials and the last two trials of each CS+ relative to fixation during the acquisition session, which was the expected direction of activity based on previous literature[21]. Caudate and ventral striatum were defined using the FSL Structural Striatal Atlas[41].

Data availability. The data that support the findings of this study are available from the corresponding author upon reasonable request.

Received 10 May 2016; accepted 10 October 2016; published 21 November 2016

References

1. LeDoux, J. *Anxious* (Oneworld Publications, 2015).
2. Schiller, D. *et al.* Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* **463,** 49–53 (2010).
3. Lissek, S. *et al.* Classical fear conditioning in the anxiety disorders: a meta-analysis. *Behav. Res. Ther.* **43,** 1391–1424 (2005).
4. Yehuda, R. & LeDoux, J. Response variation following trauma: a translational neuroscience approach to understanding PTSD. *Neuron* **56,** 19–32 (2007).
5. Dickinson, A. & Dearing, M. in *Mechanisms of Learning and Motivation* (eds Dickinson, A. & Boakes, R. A.) Ch. 8 (Psychology Press, 1979).
6. Dunsmoor, J. E., Campese, V. D., Ceceli, A. O., LeDoux, J. E. & Phelps, E. A. Novelty-facilitated extinction: providing a novel outcome in place of an expected threat diminishes recovery of defensive responses. *Biol. Psychiatry* **78,** 203–209 (2015).
7. Schnurr, P. P. *et al.* Cognitive behavioral therapy for posttraumatic stress disorder in women: a randomized controlled trial. *JAMA* **297,** 820–830 (2007).
8. Siegel, P. & Weinberger, J. Less is more: the effects of very brief versus clearly visible exposure. *Emotion* **12,** 394–402 (2012).
9. Esteves, F., Parra, C., Dimberg, U. & Ohman, A. Nonconscious associative learning: Pavlovian conditioning of skin conductance responses to masked fear-relevant facial stimuli. *Psychophysiology* **31,** 375–385 (1994).
10. Knight, D. C., Nguyen, H. T. & Bandettini, P. A. Expression of conditional fear with and without awareness. *Proc. Natl Acad. Sci. USA* **100,** 15280–15283 (2003).
11. Raio, C. M., Carmel, D., Carrasco, M. & Phelps, E. A. Nonconscious fear is quickly acquired but swiftly forgotten. *Curr. Biol.* **22,** R477–R479 (2012).
12. Redondo, R. L. *et al.* Bidirectional switch of the valence associated with a hippocampal contextual memory engram. *Nature* **513,** 426–430 (2014).
13. Shibata, K., Watanabe, T., Sasaki, Y. & Kawato, M. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* **334,** 1413–1415 (2011).
14. 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,** 470–475 (2015).
15. Amano, K., Shibata K., Kawato, M., Sasaki, Y. & Watanabe, T. Learning to associate orientation with color in early visual areas by associative decoded fMRI neurofeedback. *Curr. Biol* **26,** 1861–1866 (2016).
16. Bouton, M. E. Context and behavioral processes in extinction. *Learn. Mem.* **11,** 485–494 (2004).
17. Milad, M. R. *et al.* Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. *Biol. Psychiatry* **66,** 1075–1082 (2009).
18. 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).
19. Phelps, E. A., Delgado, M. R., Nearing, K. I. & LeDoux, J. E. Extinction learning in humans: role of the amygdala and vmPFC. *Neuron* **43,** 897–905 (2004).
20. Do-Monte, F. H., Manzano-Nieves, G., Quiñones-Laracuente, K., Ramos-Medina, L. & Quirk, G. J. Revisiting the role of infralimbic cortex in fear extinction with optogenetics. *J. Neurosci.* **35,** 3607–3615 (2015).
21. Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M.-H. & Phelps, E. A. Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proc. Natl Acad. Sci. USA* **110,** 20040–20045 (2013).
22. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103,** 3863–3868 (2006).
23. 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).
24. O'Doherty, J. *et al.* Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304,** 452–454 (2004).
25. Friston, K. J. *et al.* Psychophysiological and modulatory interactions in neuroimaging. *Neuroimage* **6,** 218–229 (1997).
26. Sarrazin, J.-C., Cleeremans, A. & Haggard, P. How do we know what we are doing? Time, intention and awareness of action. *Conscious. Cogn.* **17,** 602–615 (2008).
27. Koenigs, M. *et al.* Focal brain damage protects against post-traumatic stress disorder in combat veterans. *Nat. Neurosci.* **11,** 232–237 (2008).
28. Kim, J. H., Hamlin, A. S. & Richardson, R. Fear extinction across development: the involvement of the medial prefrontal cortex as assessed by temporary inactivation and immunohistochemistry. *J. Neurosci.* **29,** 10802–10808 (2009).
29. Emmert, K. *et al.* Meta-analysis of real-time fMRI neurofeedback studies using individual participant data: how is brain regulation mediated? *Neuroimage* **124,** 806–812 (2016).
30. Brooks, D. C., Beth, H., Nelson, J. B. & Bouton, M. E. Reinstatement after counterconditioning. *Anim. Learn. Behav.* **23,** 383–390 (1995).
31. LaBar, K. S. & Phelps, E. A. Reinstatement of conditioned fear in humans is context dependent and impaired in amnesia. *Behav. Neurosci.* **119,** 677–686 (2005).
32. Norrholm, S. D. *et al.* Conditioned fear extinction and reinstatement in a human fear-potentiated startle paradigm. *Learn. Mem.* **13,** 681–685 (2006).
33. Hermans, D. *et al.* Reinstatement of fear responses in human aversive conditioning. *Behav. Res. Ther.* **43,** 533–551 (2005).
34. Haxby, J. V. *et al.* A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* **72,** 404–416 (2011).
35. Engel, S. A., Glover, G. H. & Wandell, B. A. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb. Cortex* **7,** 181–192 (1997).
36. Yamashita, O., Sato, M.-A., Yoshioka, T., Tong, F. & Kamitani, Y. Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *Neuroimage* **42,** 1414–1429 (2008).
37. Kobayashi, S. *et al.* Influences of rewarding and aversive outcomes on activity in macaque lateral prefrontal cortex. *Neuron* **51,** 861–870 (2006).
38. Bray, S., Shimojo, S. & O'Doherty, J. P. Direct instrumental conditioning of neural activity using functional magnetic resonance imaging-derived reward feedback. *J. Neurosci.* **27,** 7498–7507 (2007).
39. Lonsdorf, T. B., Haaker, J. & Kalisch, R. Long-term expression of human contextual fear and extinction memories involves amygdala, hippocampus and ventromedial prefrontal cortex: a reinstatement study in two independent samples. *Soc. Cogn. Affect. Neurosci.* **9,** 1973–1983 (2014).
40. Kalisch, R. *et al.* Context-dependent human extinction memory is mediated by a ventromedial prefrontal and hippocampal network. *J. Neurosci.* **26,** 9503–9511 (2006).
41. Tziortzi, A. C. *et al.* Imaging dopamine receptors in humans with [^{11}C]-(+)-PHNO: dissection of D3 signal and anatomy. *Neuroimage* **54,** 264–277 (2011).
42. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* **15,** 1–25 (2002).

Acknowledgements

We thank K. Nakamura for her help in scheduling and conducting the experiment, N. Hiroe for assistance with equipment, Y. Shimada and A. Nishikido for operating the fMRI scanner, H. Ban for technical advice, and M. Craske, M. Treanor, M. Sun, A. Izquierdo and F. Krasne for their comments on the manuscript. The study was conducted in the ImPACT Program of Council for Science, Technology and Innovation (Cabinet Office, Government of Japan). This work was partially supported

by 'Brain Machine Interface Development' under the Strategic Research Program for Brain Sciences supported by the Japan Agency for Medical Research and Development (AMED), the ATR entrust research contract from the National Institute of Information and Communications Technology, and the US National Institute of Neurological Disorders and Stroke of the National Institutes of Health (grant no. R01NS088628 to H.L.). B.S. is funded by the Wellcome Trust, UK. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

A.K., H.L., B.S. and M.K. designed the study while actively discussing with other co-authors, A.K., K.A. and A.C. implemented the experiment, A.K. conducted the experiment, A.K., K.S., A.C., H.L. and M.K. analysed the results with the support of K.A. and W.Y. A.K., B.S., H.L. and M.K. wrote the manuscript.

Additional information

Supplementary information is available for this paper.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials should be addressed to M.K.

How to cite this article: Koizumi, A. *et al.* Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nat. Hum. Behav.* **1,** 0006 (2016).

Competing interests

K.S. and M.K. are the inventors of patents related to the DecNef method used in this study, and the original assignee of the patents is ATR, with which some of the authors are affiliated.