A pre-registered decoded neurofeedback intervention for specific phobias

Cody A. Cushing[1], Hakwan Lau[2*], Mitsuo Kawato[3,4], Michelle G. Craske[1], Vincent Taschereau-Dumouchel[5,6*]

Affiliations:

1. Department of Psychology, UCLA, Los Angeles, CA, USA
2. RIKEN Center for Brain Science, Wako, Saitama, Japan
3. Brain Information Communication Research Laboratory Group, Advanced Telecommunications Research Institute International, Kyoto, Japan
4. XNef, Inc., Kyoto, Japan
5. Department of Psychiatry and Addictology, Université de Montréal, Montreal, Quebec, Canada
6. Centre de Recherche de l'Institut Universitaire en Santé Mentale de Montréal, Montreal, Quebec, Canada

* Corresponding Author

Abstract

Background: Treatment attrition rates can be high for specific phobia, partly due to the subjectively aversive nature of exposure therapy that involves direct exposure to fear- and panic-inducing stimuli. A new closed-loop fMRI method called multi-voxel neuro-reinforcement has the potential to alleviate the subjective aversiveness of interventions by directly inducing phobic representations in the brain, outside of conscious awareness. The current study seeks to test this method as an intervention for specific phobia.

Methods: In a pre-registered clinical trial, individuals (N=18) with at least two animal subtype specific phobias underwent double-blind multi-voxel neuro-reinforcement for one of the two feared animals, with the untargeted one serving as control. Unaware of the target of neuro-reinforcement (i.e., the target animal), participants were guided with visual feedback and rewarded for implicit activation of the target representation. Amygdala response to phobic stimuli was assessed pre-treatment and post-treatment using photographic image presentations. Attentional capture to phobic stimuli was assessed using an affective Stroop task.

Results: Confirming our pre-registered hypothesis, a significant interaction between phobia type (target/control) and time (pre-treatment/post-treatment) was found for amygdala response. There was also a nonsignificant trend (*p*=0.055) for the hypothesized attentional capture during the affective Stroop. In both measures, responding to the phobia targeted with neuro-reinforcement was selectively reduced compared to the placebo control.

Conclusions: Results suggest multi-voxel neuro-reinforcement has the potential to be a successful intervention for specific phobia. Multi-voxel neuro-reinforcement decreased physiological and behavioral responses to specific phobia through reduced amygdala activation and attentional capture by phobic stimuli. Consequently, multi-voxel neuro-reinforcement may complement current conventional psychotherapy approaches while providing a non-distressing experience for patients seeking treatment.

Introduction

Fear-based disorders such as specific phobia and post-traumatic stress disorder (PTSD) are among the most difficult mental disorders to treat. The most widely empirically supported treatment is ‘exposure therapy’, which involves direct exposure to fear-causing or panic-inducing stimuli (1). This treatment is highly effective in reducing fear. However, conscious exposure to feared stimuli is a disturbing and unpleasant experience for the patient, leading to high rates of attrition (2,3). As a result, only a small percentage of patients can effectively benefit from an otherwise effective treatment.

Due to these treatment difficulties, neurofeedback has been explored as a way of directly regulating brain activity in a number of mental health disorders (4–10). A promising new fMRI method called multi-voxel neuro-reinforcement (11–13) has demonstrated the ability to lessen physiological defensive responses to both laboratory-conditioned fears and pre-existing fears through a kind of ‘unconscious exposure’ (14–17). By using a machine-learning classifier (also referred to as a ‘decoder’), neuro-reinforcement can be provided based on a specific stimulus category (e.g. spider) rather than average brain activity alone (17). Importantly, this can be accomplished at an implicit level as participants undergoing neuro-reinforcement are simply trying to make a feedback disc on the screen grow in size with no specific instruction as to what makes the disc grow (18,19). In reality, the feedback provided is contingent on real-time ‘decoding’ of BOLD activity indicating for instance how closely brain activity represents a feared stimulus (e.g. spider). As participants are unaware of the relation between the feedback score and the feared stimulus category, their brain is able to activate a nonconscious representation of the feared stimulus outside of the patient’s awareness. Critically, this results in no subjective discomfort for the patient, but yet can still lead to lasting reduction of fear (5,6,16,20–22).

A more consistent activation of the targeted brain representation is thought to be achieved through reinforcement learning as a reward becomes paired with the activation of the targeted brain pattern (23–25). Through this process, neural and behavioral responses to feared stimuli can then be altered. While the exact mechanism of action is not yet fully understood, early results are consistent with an exposure mechanism such as extinction (7) indicating that the successful activation of the targeted brain pattern may be the main factor driving the decrease in neural and behavioral responses.

Regardless of the precise mechanism, multi-voxel neuro-reinforcement has shown early promise as a clinical intervention that can be applied outside of conscious awareness, eliminating the need for fearful conscious exposures (14,15). This kind of intervention can potentially target any neural pattern that can be identified reliably with multivariate-pattern analysis (MVPA) (11). Typically, the construction of such machine-learning decoders in a patient's brain involves repeated visual presentations of specific stimuli. Such explicit exposures to the feared stimuli would seemingly nullify the entire appeal of the multi-voxel neuro-reinforcement procedure. However, recent advances in fMRI methodology have enabled leveraging the data of "surrogate" participants in order to train such brain decoders (Fig. 1). This can be achieved by conducting functional alignment of fMRI brain data, allowing them to be moved from the native space of one person into another (26). Functional alignment methods, such as hyperalignment, have been shown to be superior to simple anatomical registration, possibly because cortical regions tend to be organized more functionally rather than strictly structurally (i.e., as a function of structural landmarks) (26).

By leveraging functional alignment approaches, a decoder can be built for a patient with a phobia using brain data from a group of healthy controls for whom viewing repeated images of a target representation (e.g. spider) produces no fear reaction (Fig. 1). The patient simply needs to undergo a similar task (minus the phobic images) while fMRI data are collected in order to calculate the necessary functional alignment. Training between-subject decoders this way enables “nonconscious exposure” in patients with phobias, without exposing them to feared stimuli. This surrogate data approach was explored in our previous proof-of-concept study (15), but there participants still saw the feared images during the decoder construction task. Here, we test for the first time whether multi-voxel neuro-reinforcement can succeed using a decoder trained completely on surrogate data where the participant undergoing neuro-reinforcement has never seen the feared images.

The specificity of the decoder also allows the opportunity for a within-subject placebo control provided the patient has more than one phobia. For example, if a patient has a snake and a spider phobia, a decoder can be built specifically for spiders while snakes remain a placebo control. Such within-subject placebo controls are not possible with other forms of neurofeedback: for example, increasing univariate BOLD signal within a region of interest cannot be specifically related to one image category. Here, we describe a pre-registered (https://osf.io/rj34q/?view_only=b6827aa394f143aeb29b99c095bd4183) double-blind placebo-controlled clinical trial of this method as an intervention in a population with specific phobia. We pre-registered 5 hypotheses (H1-5). We hypothesized that amygdala responses (H1) and skin conductance responses (H2) to phobic stimuli would selectively decrease for the targeted phobia relative to the control phobia following neuro-reinforcement. We focus on amygdala responding as our primary outcome due to its canonical role in learning and extinction of threat and fear responses (27–31). Additionally, we hypothesized that subjective fear ratings would stay the same following neuro-reinforcement (H3), despite the predicted changes in physiological responses, based on our previous findings in a non-clinical population (15). Secondarily, we introduce a modified affective Stroop task in which participants make rapid size judgments about phobic and neutral stimuli. In this task, we hypothesized that reaction times would be slower for phobic stimuli (H4i) and that following neuro-reinforcement there would be a selective reduction in reaction times (H4ii) and amygdala responses (H4iii) in response to the targeted phobia category compared to the control phobia. Finally, we randomly assigned participants to receive either 1, 3, or 5 sessions of neuro-reinforcement. We hypothesized that those receiving the most neuro-reinforcement would demonstrate the largest effects (H5).

To anticipate, we did not manage to collect the full amount of data (N=30) as planned, due to pandemic-related circumstances. However, despite the reduced sample size (N=18), our main hypothesis about amygdala response reduction (H1) was confirmed. Unfortunately, we lacked the statistical power to adequately assess the between-group differences for the amount of neuro-reinforcement received (H5).

Methods

I. Participant Screening

Recruitment was accomplished through flyers, campus website announcements, and posting on online forums (e.g. Nextdoor, etc). Participants completed the modified Fear Survey Schedule (32) in order to identify healthy controls who reported no phobias and individuals who endorsed at least two specific phobias of animals from the ones included in our image dataset. Participants were excluded if they did not meet criteria for MRI scanning safety. Details of diagnostic screening and control vs phobia grouping can be found in *Supplemental Methods.*

For multi-voxel neuro-reinforcement, 23 participants (mean age (s.d.) = 26.5 (9.40), 69.6% female) with at least two specific animal phobias were enrolled for treatment. The informed consent of participants was obtained pursuant to the procedures of the Institutional Review Board at the University of California, Los Angeles. Participants were randomly assigned to complete either 1, 3, or 5 days of multi-voxel neuro-reinforcement to determine the doseresponse relationship with clinical outcomes. Of these 23 participants, 2 did not finish multi-voxel neuro-reinforcement (1 due to technical issues and 1 due to scheduling issues). Of the 21 participants who completed multi-voxel neuro-reinforcement, 1 experienced nausea during tasks and was excluded from further analysis. Then, 2 participants did not complete the pre-post "fear test" task for amygdala response (described below) and were excluded from analyses relevant to that task, leaving 18 subjects for our primary analyses (H1, H2, H3, and H5). This cohort of 18 participants falls short of our original goal of 30 participants due to shutdowns and recruitment difficulties resulting from the COVID-19 global pandemic. For secondary analysis of the affective Stroop task (H4), 2 of the 18 participants included in the fear test analysis did not complete the affective Stroop task, and one participant that did not complete the fear test task properly, but did complete the affective Stroop task, resulting in 17 participants analyzed (H4).

II. Decoder Construction

Prior to neuro-reinforcement, a between-subject machine learning decoder was trained for the target phobic image category (Fig. 1). The decoder was constructed using brain data from healthy controls (N=22) using a functional alignment method called hyperalignment (26). During an initial fMRI session (Fig. 2A), each healthy control viewed the same image dataset of 3600 images consisting of 40 categories of animals and objects (e.g. birds, butterflies, snakes, spiders). Conversely, participants with phobias viewed the same image dataset but with their specific phobias removed to avoid unnecessary exposure. Subject-specific decoders were developed using surrogate data based on previous methods (15), detailed in *Supplemental Methods* along with task details.

III. Pre- and Post-Neuro-Reinforcement Assessments

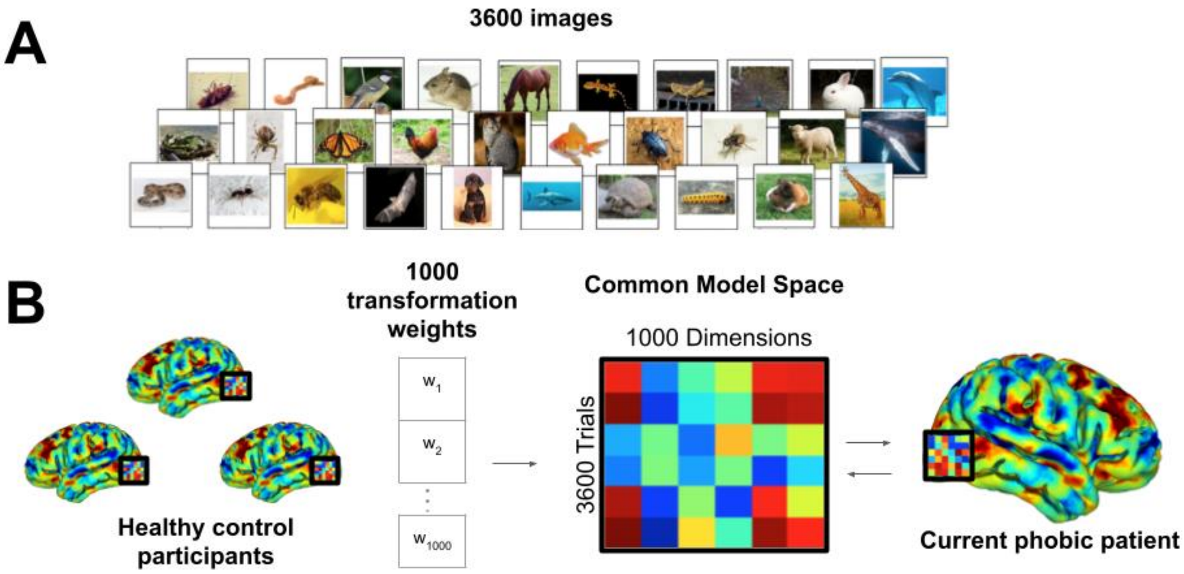


Figure 1. Functional alignment of brain data into phobic patient brain using hyperalignment. (A) All participants complete a near-identical task in the fMRI scanner where 3600 images are rapidly viewed during 0.98 second presentations. Phobic patients view happy human faces instead of their own phobic categories. Healthy controls view images from all categories. (B) Transformation parameters into the functionally aligned common model space are determined with phobic image trials withheld. Data from all participants for all categories (including phobic categories) are transformed into the common model space and then reverse transformed into the native space of the current phobic participant. A machine-learning classifier can then be trained on phobic images in the patient's native brain space despite the patient never having personally viewed the images.

Each phobia participant completed a pre-treatment and post-treatment fMRI session (Fig. 2A), during which they completed a fear test as well as an affective Stroop task while their BOLD activity was recorded.

Fear test. To assess neural and behavioral responses to phobic images, participants completed a task in which they rated how fearful they found images from select categories, following the previous proof-of-concept study (15). We refer to this task as the "fear test" and it is our primary test of neural and behavioral changes following neuro-reinforcement concerning hypotheses H1, H2, H3, and H5. During each trial, a fixation cross was presented for 3-7 seconds, followed by a static image for 6 seconds. After the static image, a blank screen was displayed for 4-12 seconds followed by a prompt to enter how fearful they found the image on a 7-point scale. These ratings were used as the subjective fear ratings to test hypothesis H3. Images displayed either belonged to the target phobia, control phobia, neutral animal, or neutral object categories. Neutral animals and objects were randomly selected based on categories for which a given participant reported no fear during their diagnostic interview. Participants completed two runs of 15 images each with a self-paced break between runs. Within each run, they viewed 5 target phobia images, 5 control phobia images, and 2-3 neutral animal/object images, counterbalanced across runs. The first image of each run was a neutral object, always immediately followed by either a target phobia or control phobia image, counterbalanced across runs. The remaining images within a run were randomly selected from each category.

Skin Conductance Response (H2). Skin Conductance Response (SCR) recordings were taken in the fMRI scanner during the fear test. Details of data collection and analysis are reported in *Supplemental Methods*.

Affective Stroop (H4). An affective Stroop task assessed reflexive attentional responses to phobic stimuli. The task started with a 1 second red fixation cross and then a brief (300 ms) image from either a phobic or neutral control category. As soon as the image appeared, participants were instructed to, as quickly and accurately as they could, make a size judgment about whether the presented animal could fit in their hand (i.e. is it the size of your hand or smaller?), by pressing one of two buttons with their index and middle finger to indicate yes or no. Response-key mappings were counterbalanced across participants. There was a 1.2 second response period (indicated by a blue fixation cross) following stimulus offset for response entry followed by a fixed 1 second inter-trial interval. Stimuli were selected from 7 animal categories: target phobia, control phobia, and 5 neutral animal categories. Similar to the fear test, neutral animal categories were selected from categories for which a given phobia participant reported no fear during their diagnostic interview. The task consisted of 210 randomly distributed trials split over 2 fMRI runs with a self-paced break between runs.

IV. Multi-voxel neuro-reinforcement

Using multi-voxel neuro-reinforcement, successful activation of the phobic image category was paired with reward (Fig. 2A). While participants laid in the fMRI scanner instructed to "use whatever mental strategy they can" to get the best feedback, a neuro-reinforcement method (15) was used to reward a nonconsciously represented phobic image category (e.g., spider). Feedback during these training sessions was based on real-time output of the decoder constructed for the individual corresponding to the specific animal phobia selected for neuro-reinforcement.

Each neuro-reinforcement run began with an extended rest period of 50 seconds while scanner image reconstruction processing caught up to real time. Then, an additional rest period of 10 seconds was collected to determine baseline BOLD activity levels followed by 16 trials of neuro-reinforcement. Each trial began with 6 seconds of rest, followed by 6 seconds of "induction" where participants modulated their brain activity in an attempt to receive high feedback. Following induction, real-time decoder output was calculated during a 4-second period and then displayed as a green disc for 2 seconds. The size of the disc directly corresponded to the likelihood estimate such that a 100% likelihood was associated with a maximum disc size (indicated by a visual boundary) and a 0% likelihood was associated with no disc display. The size of the disc also determined the amount of reward the participant received at the end of each run, with their average feedback score determining the percentage of that run's total bonus received. For example, an average feedback score of 60% resulted in 60% of the potential

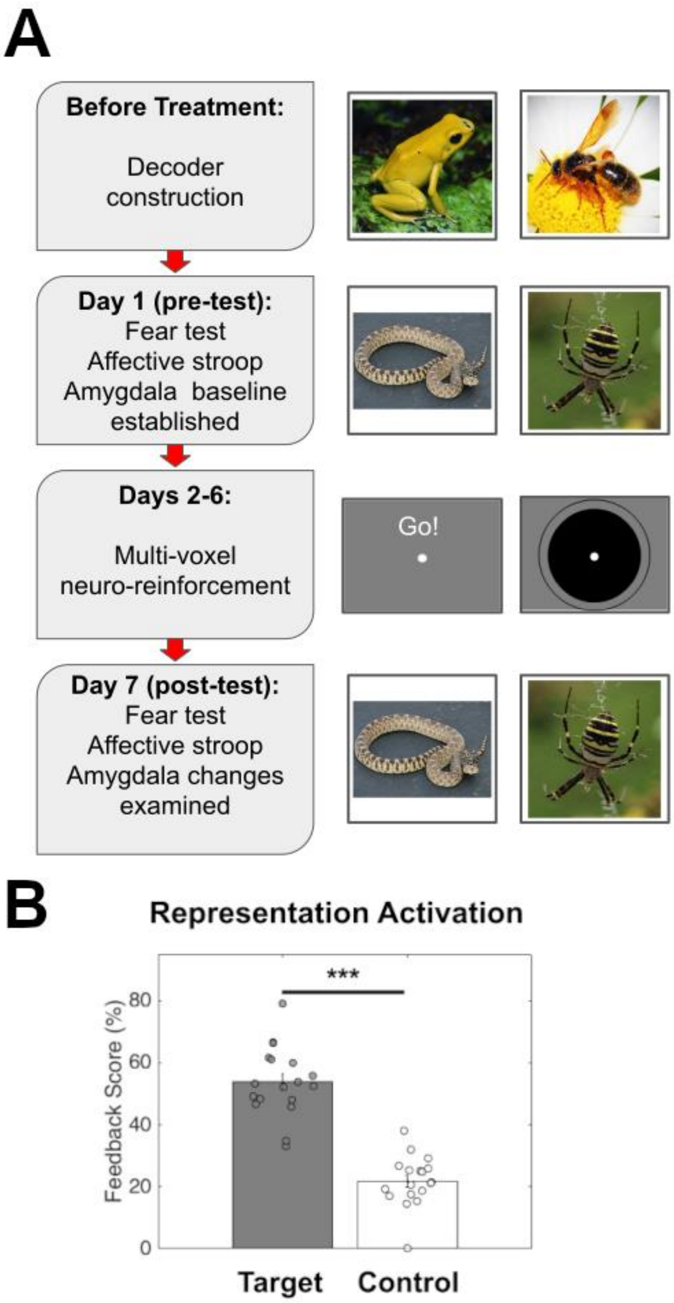


Figure 2. Study design and activation of Target and Control representations. (A) Timeline detailing patient activities during each day's fMRI session with sample stimuli from each day. Before beginning the treatment program patients undergo a decoder construction session where they view non-phobic images to enable hyperalignment with healthy control subjects. On day 1 of treatment, patients complete a pre-test in which phobic (and non-phobic) images are rated for fearfulness. Over the next 5 days, patients complete their assigned number of multi-voxel neuro-reinforcement sessions (1, 3, or 5 days). On day 7, patients complete the same task as a post-test to assess changes in amygdala and SCR response to treated and untreated phobias. (B) Representation pattern activation (measured by feedback score) for Target phobia compared to Control phobia, measured offline. Target phobia pattern was activated significantly more than Control during neuro-reinforcement. *** *p*<0.001

$6.00 bonus being received (i.e. $3.60). An additional bonus was also given when participants were able to generate a feedback score of 70% or more for 3 trials in a row. Participants were given an additional $2.00 per high-score streak bonus which was visually indicated by the feedback disc turning blue with a written message alerting them to their high-score streak.

V. Data Processing

Amygdala Response Analysis (H1, H4iii and H5). See Supplemental Methods.

Affective Stroop (H4). Reaction times were extracted for target phobia, control phobia, and neutral animal stimuli using custom scripts in MATLAB (Mathworks Inc., Natick, MA). Responses were coded as correct or incorrect based on unanimous agreement from 8 independent raters who judged whether each of the 30 potential animal categories was the size of their hand or smaller: unanimity was obtained for 24 animal categories; animal categories without consensus (bird, bat, fish, gecko, turtle, and guinea pig) were treated as correct as long as a response was recorded.

VI. Data Analysis Plan

Amygdala responses were tested with a 2 (condition: target phobia/control phobia) x 2 (time: pre-treatment/post-treatment) repeated-measures ANOVA using JASP software (JASP Team 2022). Due to limited sample size (from the COVID-19 pandemic), we were insufficiently powered to analyze neuro-reinforcement dosage groups separately, as we had initially pre-registered in hypothesis H5. Instead, neuro-reinforcement dosage (1, 3, or 5 days) was included as a covariate in the ANOVA as was each participant's total number of phobias, as a measure of clinical severity. The dosage group data are presented in *Supplemental Figures S1 and S2* for illustration purposes and the pre-registered statistical analysis for H5 is reported in *Supplemental Results*. To test the hypothesis of a significant reduction in amygdala response for the target phobia category post-treatment compared to the control phobia (H1) planned t-tests were performed on pre- and post-treatment activations for the target phobia and control phobia.

Planned t-tests were performed on pre- and post-treatment subjective fear ratings for the target phobia and control phobia, using custom scripts in Matlab, to test H3. One participant included in the amygdala analysis was excluded from this analysis due to not using the button box properly, resulting in 17 participants.

To verify phobic images were modulating attention as intended, a t-test was performed on reaction times to phobic images (grouping target and control) and neutral animal images pre-treatment (H4.i). For treatment effects (H4.ii), reaction times for correct trials were tested with a 2 (condition: target phobia/control phobia) x 2 (time: pre-treatment/post-treatment) repeated-measures ANOVA using JASP software (JASP Team 2022). Dosage group and number of phobias were included as covariates in the model. Planned t-tests were performed on pre- and post-treatment reaction times for the target phobia and control phobia.

Results

Double-blinded placebo control

After neuro-reinforcement, participants were unable to correctly guess the identity of their neuro-reinforcement target (43% accuracy in a two-alternative forced choice between the target and control phobic stimuli; chance level 50%) and participants reported strategies for neuro-reinforcement that were unrelated to the target and control categories. Collectively, this indicates neuro-reinforcement was carried out at an implicit level with participants being blind to the target of the intervention.

Target pattern induction

To assess the degree to which the desired pattern associated with the target phobic category was activated by patients during neuro-reinforcement, the feedback scores patients saw (representing degree of desired neural pattern activation) during neuro-reinforcement were compared to the scores patients would have seen if feedback had been based on the control phobic category pattern instead. The feedback was significantly higher for the target phobic category compared to what it would have been for the control phobic category (t(17)=12.63, p<0.001) (Fig. 2B). This result indicates that the desired target pattern was successfully activated by patients during neuro-reinforcement.

Amygdala Response (H1 and H5)

Before neuro-reinforcement, there was a significant amygdala response for both the target phobia (*t*(17)=2.20, *p*=0.042) and control phobia (*t*(17)=2.27, *p*=0.037) compared to neutral animals as confirmed by one-sample t-tests performed on the baselined parameter estimates. There was no difference in amygdala responses between the target and control phobias prior to neuro-reinforcement (*t*(17)=0.85, *p*=0.41). This indicates successful capturing of threat responding in the amygdala for phobic images.

Following neuro-reinforcement, there was a significant interaction between phobia type (target/control) and time (pre/post) shown by a 2 (condition) x 2 (time) repeated-measures ANOVA (*F*(1,15)=5.52, *p*=0.033, Fig. 3A). This result indicates a greater reduction in amygdala response to target phobic images than to control phobic images following neuro-reinforcement. After neuro-reinforcement, the decrease in amygdala response trended towards significance for the target phobia (*t*(17)=1.87, *p*=0.079) but not the control phobia (*t*(17)=1.65, *p*=0.12). These findings support our pre-registered hypothesis H1 that amygdala activation would be selectively reduced for the target phobia following neuro-reinforcement, and indicate that physiological threat response to the target phobia is reduced by neuro-reinforcement (see supplementary material for the H5 results).

Skin Conductance Response (H2)

Our findings did not support hypothesis H2. We did not detect a pre-treatment phobia response in SCR data, using one-sample t-tests on baseline-corrected SCR values for either target (*t*(8)=0.86,*p*=0.42) or control (*t*(8)=0.38,*p*=0.71) phobias. Given no significant pre-existing SCR response to be changed via neuro-reinforcement, no further statistical testing was performed.

Self-Reported Fear (H3)

There was no significant change in self-reported fear levels in response to either the target phobia (*t*(16)=-1.52, *p*=0.15) or the control phobia (*t*(16)=-0.56, *p*=0.58), supporting our pre-registered hypothesis H3. These findings match previous findings that self-reported fear levels are not modulated by neuro-reinforcement (15).

Affective Stroop (H4)

Before treatment, reaction times for phobic stimuli were significantly slower compared to responses to neutral stimuli (*t*(16)=2.46, *p*=0.026), confirming our pre-registered hypothesis H4i. Slower reaction times for phobic stimuli indicate that attention is successfully captured by phobic stimuli in this task. Following neuro-reinforcement, there was a borderline significant interaction between phobia type (target/control) and time (pre/post) (*F*(1,14)=4.373, *p*=0.055), such that reaction times to the target phobia were faster following neuro-reinforcement than they were to the control phobia (Fig. 3B). Specifically, there was a trend towards significantly decreased reaction times to target phobia stimuli from pre-treatment to post-treatment (*t*(16)=1.92, *p*=0.072) but not for control phobic stimuli (*t*(16)=1.52), *p*=0.14). Selectively decreased reaction times for the target phobia indicate that attention is captured less by the target phobia following

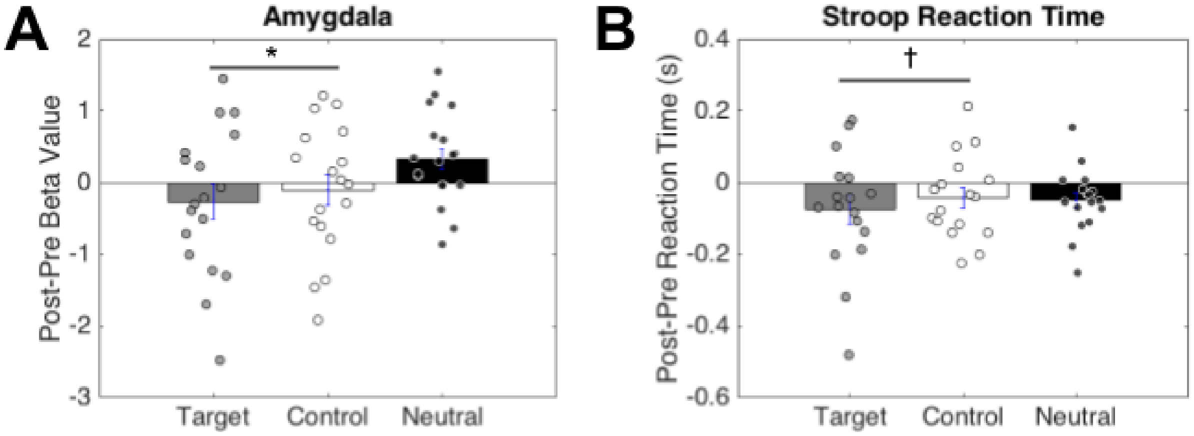


Figure 3. Changes in fear test amygdala responses and affective Stroop reaction times following neuro-reinforcement. (A) Amygdala response in the fear test showed a greater decrease in the Target than Control phobias following neuro-reinforcement. (B) Response times in the affective Stroop task showed a greater decrease in the Target than Control phobias following neuro-reinforcement. * p<0.05, † p<0.10 for interaction between Time (pre/post) and experimental condition (target/control) when controlling for days of neuro-reinforcement and clinical severity.

neuro-reinforcement. Amygdala responding during affective Stroop (H4.iii) is reported in *Supplemental Results.*

Discussion

In a double-blind placebo-controlled clinical trial, we investigated whether multi-voxel neuro-reinforcement could nonconsciously intervene on specific phobia. We found evidence of specific reduction in amygdala reactivity to the target phobia (H1) supporting previous findings (15) and reduced attentional capture by the target phobia following neuro-reinforcement in an affective Stroop task (H4). Importantly, these findings were obtained using neuro-reinforcement based on decoders trained on completely surrogate data. Consequently, this study supports the ability of decoded neuro-reinforcement to be performed without any previous exposure to the feared stimulus. Our findings were obtained using double-blind procedures. Most psychological interventions are difficult to test at such a rigorous double-blind level. This means that the efficacy of neuro-reinforcement as a clinical intervention was tested with a level of rigor that is rarely achieved by other interventions.

The changes we observed in amygdala responses and Stroop reaction times to phobic stimuli represent changes in physiological and reflexive responses to threat (33–35). These changes may represent 'preconscious' responses to feared stimuli due to their automatic and reflexive nature (36–42). No effects were observed with respect to explicit subjective ratings of fear, consistent with our hypotheses (H3) and prior findings (15). This pattern of results suggests that implicit neuro-reinforcement is more effective for automatic physiological responses to threat compared to the subjective experience of fear itself. The discordance across response modalities is consistent with a higher-order theory of emotion in which subjective mental experience operates via different mechanisms than physiological threat responses (33–35,43). While an effective treatment would ultimately aim to reduce subjective fear experiences when confronting phobic stimuli, neuro-reinforcement could represent an important first step in reducing overall subjective discomfort during traditional exposure treatments. For example, with reduced physiological threat responses, the reduction of subjective discomfort during traditional exposures may occur at an increased rate as subjective feelings come into alignment with already decreased physiological responding.

This notion is indirectly supported by the results from the affective Stroop task. Following neuro-reinforcement, there was a trend towards reaction times being significantly decreased for the target phobia relative to the control phobia (H4ii). In addition to providing further support for specific target engagement by neuro-reinforcement, this result suggests that individuals may be less reflexively avoidant of their phobia following neuro-reinforcement. If this is the case, patients may find traditional behavioral exposure treatments less aversive or at least be more likely to persist in exposure, following neuro-reinforcement leading to lower rates of attrition.

To test this hypothesis, future studies should complement neuro-reinforcement with a behavioral-approach task to investigate whether physiological symptoms are decreased when approaching the target phobia following neuro-reinforcement. If patients are more willing to approach the feared animal following neuro-reinforcement, then neuro-reinforcement may be a helpful complementary treatment alongside traditional exposure for ensuring the most comfortable treatment regimen possible. Additionally, future studies should explore how long neuro-reinforcement effects last following the intervention by re-testing participants weeks or months after neuro-reinforcement is completed.

The current study is not without limitations however. We were unable to detect an SCR in response to phobic stimuli in our group of participants. Consequently, we were unable to test one of our pre-registered hypotheses (H2) that neuro-reinforcement would lead to reduced phobic SCR responding. The lack of skin conductance threat response may have arisen from technical limitations, a large portion of participants being non-responders, or our relatively limited sample size. Similarly, the current study lacked the statistical power to test one of our other main hypotheses (H5); a between-subjects analysis of how much neuro-reinforcement is sufficient to achieve the desired outcomes. This limitation is directly due to our smaller-than-planned sample size (18 compared to 30 participants), a shortcoming that was due to the COVID-19 pandemic.

In summary, this study represents the first clinical trial of multi-voxel neuro-reinforcement for nonconscious brain-based psychotherapy. This procedure demonstrated the ability to lessen physiological, reflexive responses to specific phobia through reduced amygdala activation as well as less attentional capture by phobic stimuli. These findings provide a promising foundation to attempt larger-scale replications in clinical cohorts. Through advents in virtual reality, these responses can also be investigated in future studies using more realistic and immersive stimuli (44–48). This nonconscious procedure produces minimal discomfort in patients with very low rates of attrition. Consequently, neuro-reinforcement may serve to complement current conventional psychotherapy approaches while providing a more tolerable experience for patients seeking treatment.

Acknowledgments

The authors would like to thank study coordinators Ana Costello, Annelise Murillo, and Shawn Wang for their assistance to this study through participant recruitment and data collection.
HL received financial support from the US National Institute of Mental Health (R61MH113772) and The Templeton World Charity Foundation (RA537-01). MK was supported by AMED under grant number JP18dm0307008. MGC received financial support from the US National Institute of Mental Health (R61MH113772). VT-D received financial support from the Fonds de Recherche du Québec - Santé (FRQS).

Disclosures

Author Mitsuo Kawato is an inventor of patents owned by the Advanced Telecommunications Research Institute International related to the present work (PCT/JP2012/078136 [WO2013/06 871 9517] and PCT/JP2014/61543 [WO2014/178322]).

References

1. Craske MG, Kircanski K, Zelikowsky M, Mystkowski J, Chowdhury N, Baker A (2008): Optimizing inhibitory learning during exposure therapy. *Behav Res Ther* 46: 5–27.

2. Loerinc AG, Meuret AE, Twohig MP, Rosenfield D, Bluett EJ, Craske MG (2015): Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clin Psychol Rev* 42: 72–82.

3. Zayfert C, DeViva JC, Becker CB, Pike JL, Gillock KL, Hayes SA (2005): Exposure utilization and completion of cognitive behavioral therapy for PTSD in a "real world" clinical practice. *Journal of Traumatic Stress* 18: 637–645.

4. Young KD, Zotev V, Phillips R, Misaki M, Yuan H, Drevets WC, Bodurka J (2014): Real-Time fMRI Neurofeedback Training of Amygdala Activity in Patients with Major Depressive Disorder. *PLOS ONE* 9: e88785.

5. Gerin MI, Fichtenholtz H, Roy A, Walsh CJ, Krystal JH, Southwick S, Hampson M (2016): Real-Time fMRI Neurofeedback with War Veterans with Chronic PTSD: A Feasibility Study. *Frontiers in Psychiatry* 7. Retrieved March 22, 2023, from https://www.frontiersin.org/articles/10.3389/fpsyt.2016.00111

6. Scheinost D, Stoica T, Saksa J, Papademetris X, Constable RT, Pittenger C, Hampson M (2013): Orbitofrontal cortex neurofeedback produces lasting changes in contamination anxiety and resting-state connectivity [no. 4]. *Transl Psychiatry* 3: e250–e250.

7. Chiba T, Kanazawa T, Koizumi A, Ide K, Taschereau-Dumouchel V, Boku S, *et al.* (2019): Current Status of Neurofeedback for Post-traumatic Stress Disorder: A Systematic Review and the Possibility of Decoded Neurofeedback. *Frontiers in Human Neuroscience* 13. https://doi.org/10.3389/fnhum.2019.00233

8. Stoeckel LE, Garrison KA, Ghosh SS, Wighton P, Hanlon CA, Gilman JM, *et al.* (2014): Optimizing real time fMRI neurofeedback for therapeutic discovery and development. *NeuroImage: Clinical* 5: 245–255.

9. Mennen AC, Turk-Browne NB, Wallace G, Seok D, Jaganjac A, Stock J, *et al.* (2021): Cloud-Based Functional Magnetic Resonance Imaging Neurofeedback to Reduce the Negative Attentional Bias in Depression: A Proof-of-Concept Study. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging* 6: 490–497.

10. Young KD, Siegle GJ, Zotev V, Phillips R, Misaki M, Yuan H, *et al.* (2017): Randomized Clinical Trial of Real-Time fMRI Amygdala Neurofeedback for Major Depressive

Disorder: Effects on Symptoms and Autobiographical Memory Recall. *AJP* 174: 748–755.

11. Cohen JD, Daw N, Engelhardt B, Hasson U, Li K, Niv Y, *et al.* (2017): Computational approaches to fMRI analysis [no. 3]. *Nat Neurosci* 20: 304–313.
12. Watanabe T, Sasaki Y, Shibata K, Kawato M (2017): Advances in fMRI Real-Time Neurofeedback. *Trends in Cognitive Sciences* 21: 997–1010.
13. deBettencourt MT, Cohen JD, Lee RF, Norman KA, Turk-Browne NB (2015): Closed-loop training of attention with real-time brain imaging [no. 3]. *Nat Neurosci* 18: 470–475.
14. Koizumi A, Amano K, Cortese A, Shibata K, Yoshida W, Seymour B, *et al.* (2017): Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nature Human Behaviour*. https://doi.org/10.1038/s41562-016-0006
15. Taschereau-Dumouchel V, Cortese A, Chiba T, Knotts JD, Kawato M, Lau H (2018): Towards an unconscious neural reinforcement intervention for common fears. *Proceedings of the National Academy of Sciences of the United States of America* 115: 3470–3475.
16. Siegel P, Cohen B, Warren R (2022): Nothing to Fear but Fear Itself: A Mechanistic Test of Unconscious Exposure. *Biological Psychiatry* 91: 294–302.
17. Taschereau-Dumouchel V, Cushing CA, Lau H (2022): Real-Time Functional MRI in the Treatment of Mental Health Disorders. *Annual Review of Clinical Psychology* 18: 125–154.
18. Taschereau-Dumouchel V, Cortese A, Lau H, Kawato M (2021): Conducting decoded neurofeedback studies. *Social Cognitive and Affective Neuroscience* 16: 838–848.
19. Taschereau-Dumouchel V, Liu K-Y, Lau H (2018): Unconscious psychological treatments for physiological survival circuits. *Current Opinion in Behavioral Sciences* 24: 62–68.
20. Siegel P, Warren R (2013): Less is still more: Maintenance of the very brief exposure effect 1 year later. *Emotion* 13: 338–344.

21. Siegel P, Warren R (2013): The effect of very brief exposure on experienced fear after in vivo exposure. *Cognition and Emotion* 27: 1013–1022.

22. Rance M, Walsh C, Sukhodolsky DG, Pittman B, Qiu M, Kichuk SA, *et al.* (2018): Time course of clinical change following neurofeedback. *NeuroImage* 181: 807–813.

23. Shibata K, Lisi G, Cortese A, Watanabe T, Sasaki Y, Kawato M (2019): Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback. *NeuroImage* 188: 539–556.

24. Oblak EF, Lewis-Peacock JA, Sulzer JS (2017): Self-regulation strategy, feedback timing and hemodynamic properties modulate learning in a simulated fMRI neurofeedback environment. *PLOS Computational Biology* 13: e1005681.

25. Momennejad I, Otto AR, Daw ND, Norman KA (2018): Offline replay supports planning in human reinforcement learning ((D. Badre & M. J. Frank, editors)). *eLife* 7: e32548.

26. Haxby JV, Guntupalli JS, Connolly AC, Halchenko YO, Conroy BR, Gobbini MI, *et al.* (2011): A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* 72: 404–416.

27. Phelps EA, Delgado MR, Nearing KI, LeDoux JE (2004): Extinction Learning in Humans: Role of the Amygdala and vmPFC. *Neuron* 43: 897–905.

28. Delgado MR, Nearing KI, LeDoux JE, Phelps EA (2008): Neural Circuitry Underlying the Regulation of Conditioned Fear and Its Relation to Extinction. *Neuron* 59: 829–838.

29. LaBar KS, Gatenby JC, Gore JC, LeDoux JE, Phelps EA (1998): Human Amygdala Activation during Conditioned Fear Acquisition and Extinction: a Mixed-Trial fMRI Study. *Neuron* 20: 937–945.

30. Wen Z, Raio CM, Pace-Schott EF, Lazar SW, LeDoux JE, Phelps EA, Milad MR (2022): Temporally and anatomically specific contributions of the human amygdala to threat and safety learning. *Proceedings of the National Academy of Sciences* 119: e2204066119.

31. Lissek S, Powers AS, McClure EB, Phelps EA, Woldehawariat G, Grillon C, Pine DS (2005): Classical fear conditioning in the anxiety disorders: a meta-analysis. *Behaviour Research and Therapy* 43: 1391–1424.

32. Wolpe J, Lang PJ (1964): A FEAR SURVEY SCHEDULE FOR USE IN BEHAVIOUR THERAPY. *Behav Res Ther* 2: 27–30.

33. LeDoux JE, Pine DS (2016): Using Neuroscience to Help Understand Fear and Anxiety: A Two-System Framework. *AJP* 173: 1083–1093.

34. Taschereau-Dumouchel V, Kawato M, Lau H (2019): Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. *Molecular Psychiatry*. https://doi.org/10.1038/s41380-019-0520-3

35. Taschereau-Dumouchel V, Michel M, Lau H, Hofmann SG, LeDoux JE (2022): Putting the “mental” back in “mental disorders”: a perspective from research on fear and anxiety. *Mol Psychiatry* 1–9.

36. Mobbs D, Marchant JL, Hassabis D, Seymour B, Tan G, Gray M, *et al.* (2009): From Threat to Fear: The Neural Organization of Defensive Fear Systems in Humans. *J Neurosci* 29: 12236–12243.

37. Larson CL, Schaefer HS, Siegle GJ, Jackson CAB, Anderle MJ, Davidson RJ (2006): Fear Is Fast in Phobic Individuals: Amygdala Activation in Response to Fear-Relevant Stimuli. *Biological Psychiatry* 60: 410–417.

38. McFadyen J, Mermillod M, Mattingley JB, Halász V, Garrido MI (2017): A Rapid Subcortical Amygdala Route for Faces Irrespective of Spatial Frequency and Emotion. *The Journal of Neuroscience* 37: 3864–3874.

39. Méndez-bértolo C, Moratti S, Toledano R, Lopez-sosa F, Martínez-alvarez R, Mah YH, *et al.* (2016): A fast pathway for fear in human amygdala. *Nature Neuroscience* 19: 1041–1049.

40. Cushing CA, Im HY, Adams RB, Ward N, Kveraga K (2019): Magnocellular and parvocellular pathway contributions to facial threat cue processing. *Social Cognitive and Affective Neuroscience* 14: 151–162.

41. Cushing CA, Im HY, Adams RB, Ward N, Albohn DN, Steiner TG, Kveraga K (2018): Neurodynamics and connectivity during facial fear perception: The role of threat exposure and signal congruity. *Scientific Reports* 8: 2776–2776.

42. Adams RB, Franklin RG, Kveraga K, Ambady N, Kleck RE, Whalen PJ, *et al.* (2012): Amygdala responses to averted vs direct gaze fear vary as a function of presentation speed. *Social Cognitive and Affective Neuroscience* 7: 568–577.

43. Brown R, Lau H, LeDoux JE (2019): Understanding the Higher-Order Approach to Consciousness. *Trends in Cognitive Sciences* 23: 754–768.

44. Morina N, Ijntema H, Meyerbröker K, Emmelkamp PMG (2015): Can virtual reality exposure therapy gains be generalized to real-life? A meta-analysis of studies applying behavioral assessments. *Behaviour Research and Therapy* 74: 18–24.

45. Bohil CJ, Alicea B, Biocca FA (2011): Virtual reality in neuroscience research and therapy [no. 12]. *Nat Rev Neurosci* 12: 752–762.

46. Dunsmoor JE, Ahs F, Zielinski DJ, LaBar KS (2014): Extinction in multiple virtual reality contexts diminishes fear reinstatement in humans. *Neurobiology of Learning and Memory* 113: 157–164.

47. Shiban Y, Reichenberger J, Neumann ID, Mühlberger A (2015): Social conditioning and extinction paradigm: A translational study in virtual reality. *Frontiers in Psychology* 6. https://doi.org/10.3389/fpsyg.2015.00400

48. Tröger C, Ewald H, Glotzbach E, Pauli P, Mühlberger A (2012): Does pre-exposure inhibit fear context conditioning? A Virtual Reality Study. *Journal of Neural Transmission* 119: 709–719.

Supplemental Information

A pre-registered decoded neurofeedback intervention for specific phobias

Cody A. Cushing, Hakwan Lau, Mitsuo Kawato, Michelle G. Craske, Vincent Taschereau-Dumouchel

Supplemental Methods

Participants

Diagnostic Assessment

All participants underwent a diagnostic interview, using the Anxiety Disorders Interview Schedule-5 (1), administered by trained and reliability certified study staff (Bachelors degree), with each interview reviewed for final consensus by the Principal Investigator (MGC).

Participants were excluded if they 1) did not have normal/corrected to normal vision or hearing; 2) unable to understand informed consent or could not complete the consent form correctly; 3) unable to respond adequately to screening questions; 4) unable to maintain focus/stillness during assessment; 5) had a history of neurological disease or defect; 6) were diagnosed with PTSD, OCD, SUD, current MDD, Bipolar, Psychosis, or any other neurologic diagnoses or unstable serious medical conditions (all assessed using the ADIS-5); 7) currently prescribed psychotropic medication.

Groups

Healthy Control Group: No animal type specific phobias or fears, ascertained from administration of the ADIS-5.

Phobia Group: Met diagnostic criteria for at least two animal type specific phobias, assessed using the ADIS-5, with the exception of the functional impairment/distress criterion. Animal phobias were only eligible if interviewer ratings of fear or avoidance for a given phobia were at least 4 on a 0-8 point scale (0 = no fear/never avoids, 8=very severe fear/always avoids). For the 23 participants that were enrolled and started a pre-treatment session, participants had a mean (s.d.) of 2.39 (0.65) phobias and target/control phobias had a mean (s.d.) fear rating of 5.54 (0.76) and avoidance rating of 5.54 (1.11).

MRI scanning parameters

All fMRI data were acquired on a 3T Siemens Prisma scanner using a 32-channel head coil at the UCLA Ahmanson-Lovelace Brain Mapping Center.

Decoder Construction

Across 6 task runs during decoder construction, fMRI data were collected with a multi-band sequence with an acceleration factor of 8 and phase encoding in the posterior (P) to anterior (A) direction in order to minimize dropout in the ventral temporal brain area. Voxel sizes were 2.0x2.0x2.0mm^3 with a 208x208mm^2 Field of View. Images were collected across 72 interleaved slices with a TR of 800ms, TE of 37.00 ms, and flip angle of 52 degrees. Anatomical data were collected using a T1-weighted imaging sequence with volumetric navigators (vNAV) with prospective motion correction (TR: 2500ms/TI: 1000ms/Flip Angle: 8.0 degrees/Voxel Size: 0.8x0.8x0.8mm/Matrix Size: 256x256/Num. Slices: 208/Slice Thickness: 0.8mm).

Multi-voxel neuro-reinforcement

Prior to the cessation of data collection due to the COVID-19 pandemic, fMRI data during the fear test task and affective Stroop task were collected across 2 runs each using the same sequence described for *Decoder Construction* for 7 participants. However, during the COVID-19 shutdown, this sequence was replaced with a similar but modified sequence better tailored for capturing BOLD activity in subcortical regions such as the amygdala. This replacement sequence used for the remaining 11 participants was a multi-band sequence with an acceleration factor of 6 and phase encoding in the A-P direction. Voxel sizes were 2.0x2.0x2.0mm^3 with a 192x192mm^2 Field of View. Images were collected across 72 interleaved slices with a TR of 1000ms, TE of 30.00ms and flip angle of 60 degrees. Accompanying Spin Echo Field Maps were collected in opposing phase encoding directions (A-P/P-A) before functional runs in order to be used for offline distortion correction. FMRI data during online neuro-reinforcement were collected using a multi-band sequence with an acceleration factor of 6 and phase encoding in the P-A direction to minimize dropout in the ventral temporal area. Additional parameters were voxel size: 2.0x2.0x2.0mm^3, FOV: 208x208mm^2, num. slices: 72, TR: 1000ms, TE: 37.00 ms, and flip angle: 60 degrees.

Decoder Construction

Decoder Construction: task

In place of phobic images, phobic participants viewed happy human faces using stimuli from the Chicago Face Database and NimStim Set of Facial Expressions (2,3). These stimuli have their emotional expression verified by independent raters and were used to provide a non-disturbing stimulus replacement that was sufficiently orthogonal to the task image set of animals and objects. The decoder construction task consisted of 6 runs of 600 trials each. Each trial consisted of a .98 second image presentation with no inter-trial interval. This rapid event-related design was used to maximize the number of images each participant viewed. To ensure attention, participants were given the task of pressing a button each time the image category changed (i.e. a 1-back task). Image categories were presented in chunks of 2, 3, 4, or 6 consecutive images.

Decoder construction: fMRIprocessing

Decoder construction fMRI data were processed using a combination of SPM12 (Statistical Parametric Mapping; www.fil.ion.ucl.ac.uk/spm) and custom python scripts using pyMVPA and sklearn packages (4,5). All 6 runs of the task were concatenated and preprocessed in SPM using default parameters unless otherwise explicitly specified. Data were realigned to the first image from the first run of the task and segmented into tissue classes. Anatomical and functional data were coregistered using the gray matter image from segmentation as a reference. Motion was then regressed out of the functional data using the 6 head motion parameters from realignment. Single-trial estimates were then generated with pyMVPA using the least-squares 2 (LS-2) method (6) in which a separate GLM is computed for each trial where the current trial is assigned to one regressor while the remaining trials are equally split between two “rest” regressors.

Using hyperalignment, single-trial estimates from healthy controls in the target brain region (ventral temporal cortex) were functionally transformed to the current phobic participant’s brain and used to train a machine-learning pattern classifier (decoder) using the phobic images that the participant did not see (Fig. 1). To ensure double-blind treatment target selection, the target for treatment was automatically selected by a computer program that calculated which phobic category had the highest cross-validated area under the receiver operating characteristic curve (AUC) during binary one vs. all classification.

To determine AUC metrics, a 6-fold cross-validation (CV) procedure was used. FMRI data for each participant were loaded and masked to the ventral temporal (VT) area in their own native space using an anatomical mask derived from Freesurfer parcellations of the fusiform, lingual, parahippocampal, and inferior temporal areas (7). Single-trial parameter estimates were standardized by feature within subject and within each of the 6 task runs. The data were split into 6 folds for training and testing based on the 6 runs completed by each participant. That is, for each CV split, the withheld testing set consisted of all the data from each participant for one of the six task runs. The remaining preprocessing was calculated using only the training data to avoid overfitting. As hyperalignment requires a stable number of features across participants, 1000 voxels were selected within the VT area via F-test to select which voxels accounted for the most variance elicited by all image categories across all training trials. For each phobic participant, a unique set of hyperalignment transformation parameters into the common model space was calculated for the current phobic participant and all healthy controls. The fitting of the hyperalignment parameters was done using trials for all image categories except the current participant’s phobias. For example, if a phobic participant had spider and snake phobias, all spider and snake trials were withheld from all participants when fitting the transformation parameters.

After hyperalignment transformation parameters were determined, the data from all healthy controls were moved into the native space of the current phobic participant by transforming the data into the common model space and then reverse transforming the data from the common model space into the native space of the current participant. The transformed data included the previously withheld phobic category images from the healthy controls as well as the testing dataset.

With all data in the current participant's native space, class sizes (target vs. non-target image categories) were balanced by random undersampling balanced between the 39 non-target image categories. Following previous work (8), a Sparse Multinomial Logistic Regression (SMLR) classifier was trained to perform binary (one-vs-rest) classification between the potential target category and all remaining categories (9). AUC scores for each CV split were calculated based on classifier estimates.

Of the potential phobic categories to be selected for treatment for the current participant, the phobia with the highest AUC scores across all 6 CV splits was blindly selected via computer program as the target for treatment. The within-subject control was also blindly selected through automated random selection from the remaining phobic categories if the participant had more than two phobias. For the final decoder to be used in neuro-reinforcement, the same procedure was performed but trained using all 6 runs of data.

Pre/Post Test

Fear test: fmri processing

FMRI task runs were distortion corrected using FSL's topup (10,11) according to spin echo field map sequences collected in opposite phase-encoding directions. Due to technical issues with spin echo field map collection, 5 participants were excluded from distortion correction. Anatomical T1 images were brain extracted using bet (12). Then, preprocessing and ICA-decomposition were performed using FSL's melodic and FEAT (FMRIB's Software Library, www.fmrib.ox.ac.uk/fsl). During preprocessing, fMRI data were motion corrected using mcflirt (13), brain extracted using bet (12), spatially smoothed with a Gaussian kernel of FWHM 4.0mm, intensity normalized, and highpass filtered with a gaussian-weighted least-squares straight line fitting with sigma=50.0s. Images were then registered to the standard MNI space using FLIRT and then refined using nonlinear registration with FNIRT (13,14). Registration of multi-band images were improved by using a high-contrast single-band reference image collected at the start of each functional run as an initial reference image for registration.

ICA components were then manually investigated with components resulting from movement or other sources of noise removed. To further account for movement, data were processed with the Artifact Detection Tools (ART, https://www.nitrc.org/projects/artifact_detect) toolbox to generate motion regressors and identify outlier timepoints for censoring. First-level GLMs were then calculated in SPM12 with a temporal derivative to account for slice-timing differences. Regressors were fit for the onset of target phobia, control phobia, neutral animal, and neutral object images with a duration of 0 seconds to model the event-related response. Following previous work (8), only the first 2 trials within each run were analyzed for target phobia and control phobia images.

Bilateral amygdala masks were generated from the automatic Freesurfer segmentation of the T1 image and transformed into the participant's native functional space. Average parameter estimates were extracted from the Amygdala using marsbar (15). Average parameter estimates for phobic stimuli were then corrected to baseline by subtracting the average amygdala response to the neutral animal from the target phobia and control phobia, within runs. Baseline-corrected phobia responses were then averaged across runs for pre-treatment and post-treatment sessions.

Skin Conductance Response (SCR)

Data collection

Skin Conductance Response (SCR) was recorded in Acknowledge software via Biopac MP-150 system using the EDA-100C module and Ag/AgCl electrodes placed distally on the index and middle fingers of the left hand. SCR recordings were taken during pre-treatment and post-treatment MRI scanning sessions. Of the 18 participants analyzed in our main analyses, 5 participants had technical issues during data collection and 4 participants were non-responders showing no discernable SCR. Consequently, 9 participants were analyzed for SCR.

Data analysis

SCR recordings were analyzed with custom code in python utilizing the bioread package. SCR data were filtered with a 1st-order 5 Hz low-pass Butterworth filter to account for influences of the magnetic field in the MRI environment. SCR recordings were then epoched according to stimulus onset times during the Fear Test task from 2 seconds preceding stimulus onset to 5 seconds following stimulus onset. Epoch timecourses were baseline corrected according to the average activity during the 2 seconds before stimulus onset. Peak SCR values were then extracted for each trial epoch by taking the maximum SCR value in the time period of 1 second to 5 seconds following stimulus onset. If the peak SCR value was less than 0.02 microsiemens then it was coded as 0 following previous research (ref). Peak SCR values were then square root transformed in preparation for statistical analysis.

Self Report Questionnaires. The following self report questionnaires were administered at pre-treatment and post-treatment:

Depression, Anxiety, and Stress Scale (DASS-21) (33), Behavioral Inhibition/Behavioral Activation Scale (BIS/BAS) (34), Sheehan Disability Scale (SDS) (35), and Modified Fear Survey Schedule (32).

Supplemental Results

Amygdala response during Stroop task

Amygdala responding during the affective Stroop task did not demonstrate the same interaction we observed during the fear test task (*F*(1,14)=1.075, *p*=0.317) counter to our pre-registered hypothesis H4iii. Additionally, in the affective Stroop task, a phobia response was not observed in response to the target phobia pre-treatment as tested with a one-sample t-test on the baselined parameter estimates (*t*(16)=0.19, *p*=0.85). This lack of significant phobia response pre-treatment could be due to the increased cognitive load of this task which required rapid, reflexive judgments as soon as the stimulus appeared (compared to fear ratings in the fear test which were input many seconds after the original stimulus disappeared). Additionally or alternatively, the amygdala may have habituated during the affective Stroop task as it was always immediately preceded by the fear test.

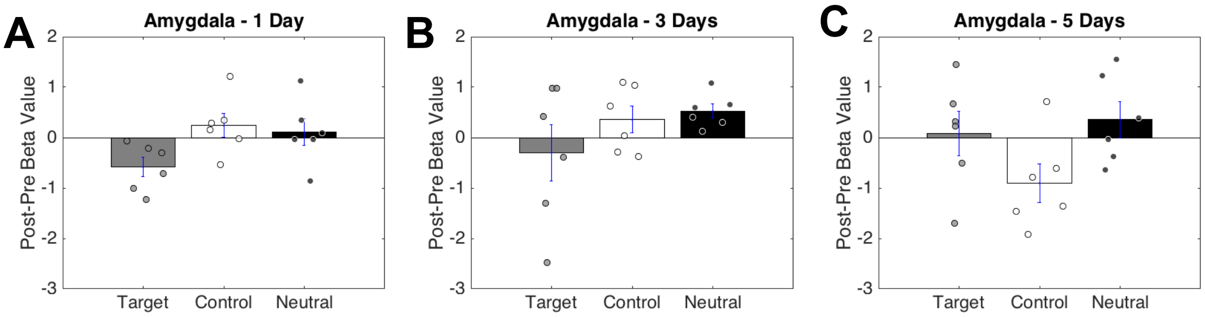
Between-subjects analysis of dosage effects (H5)

Although circumstances outside of our control (detailed in methods) prevented us from collecting a sufficient sample size to analyze the between-subject effect of dosage with sufficient power as we initially pre-registered, we report the pre-registered analysis here. When dosage (1, 3, or 5 days of neuro-reinforcement) is treated as a between-subjects factor in a 3 (between-subjects dosage: 1, 3, or 5 days of neuro-reinforcement) x 2 (within-subjects condition: target, control phobia) x 2 (within-subjects time: pre-treatment, post-treatment) repeated-measures ANOVA, we fail to find evidence in support of H5. The 3-way interaction between dosage, condition, and time is not significant (*F*(2,14) = 3.236, *p*=0.07). This lack of evidence in support of our pre-registered hypothesis H5 is most likely due to insufficient power to detect such between-subjects effect in the current design. Future studies will be needed to address the question of how the number of neuro-reinforcement sessions an individual receives affects reduced amygdala responses to feared stimuli.

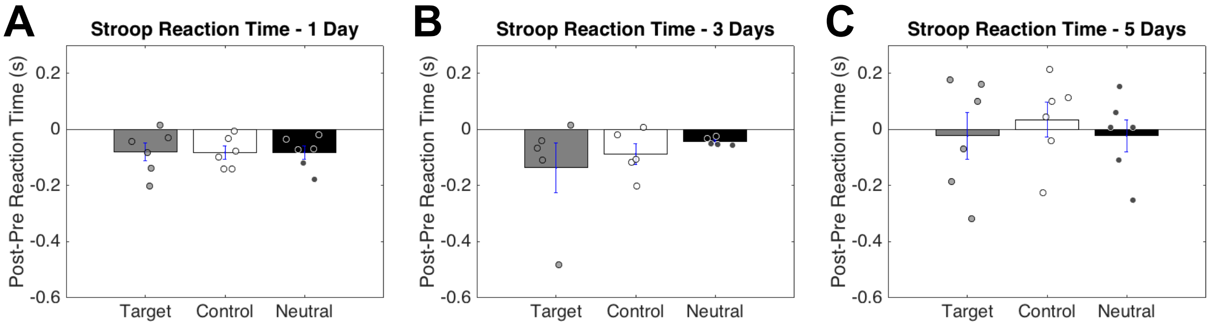
Self-Report Questionnaires

A paired sample t-test for Depression, Anxiety and Stress Anxiety Subscale was marginally significant, t(17) = 2.06, p = .055: pre-test (M = 8.9, SD = 2.7) and post-test (M = 8.2, SD = 1.6) indicating a marginal decrease in anxiety following neuro-reinforcement. There were no effects for the depression subscale or stress subscale or the total DASS score.

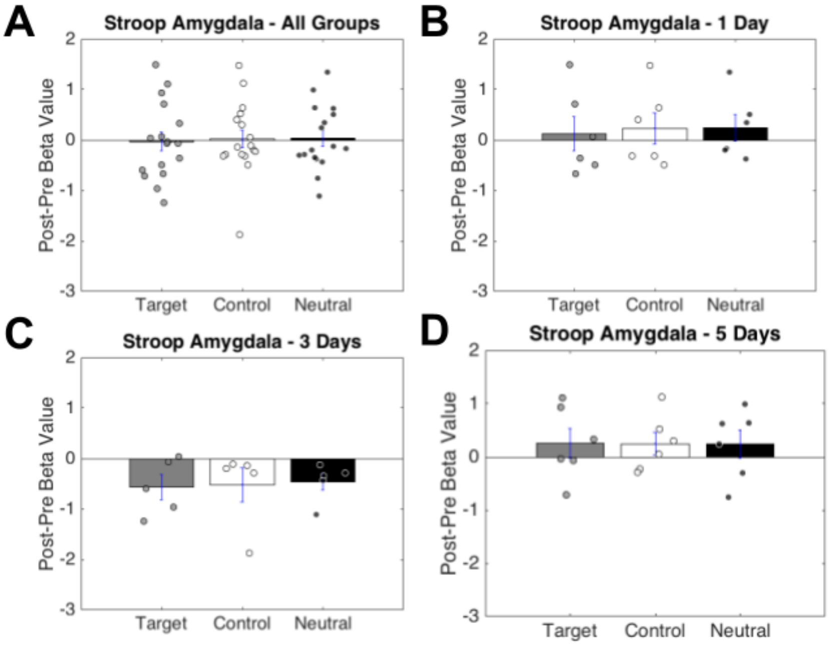
Supplemental Figures



Supplemental Figure S1. Amygdala response following neuro-reinforcement by number of days of neuro-reinforcement received. Panels show changes in responses to target phobia, control phobia, and neutral animal images from pre-neuro-reinforcement to post-neuro-reinforcement, quantified as post minus pre difference. Results for participants that received 1 day (A), 3 days (B), or 5 days (C) of neuro-reinforcement.



Supplemental Figure S2. Reaction times in affective Stroop task following neuro-reinforcement by number of days of neuro-reinforcement received. Panels show changes in reaction times to target phobia, control phobia, and neutral animal images from pre-neuro-reinforcement to post-neuro-reinforcement, quantified as post minus pre difference. Results for participants that received 1 day (A), 3 days (B), or 5 days (C) of neuro-reinforcement.



Supplemental Figure S3. Amygdala response following neuro-reinforcement in affective Stroop task. Results for all dosage groups combined (A) showing post-treatment minus pre-treatment amygdala responses to target phobia, control phobia, and neutral animals in the affective Stroop task. Also, the 1 day (B), 3 days (C), and 5 days (D) of neuro-reinforcement are also shown for illustrative purposes.

References

1. Brown T, Barlow D (2014): *Anxiety and Related Disorders Interview Schedule for DSM-5 (ADIS-5)® - Adult Version: Client Interview Schedule 5-Copy Set*. Oxford, New York: Oxford University Press.

2. Ma DS, Correll J, Wittenbrink B (2015): The Chicago face database: A free stimulus set of faces and norming data. *Behav Res Methods* 47: 1122–1135.

3. Tottenham N, Tanaka JW, Leon AC, McCarry T, Nurse M, Hare T a., *et al.* (2009): The NimStim set of facial expressions: Judgements from untrained research participants. *Psychiatry Res* 168: 242–249.

4. Hanke M, Halchenko YO, Sederberg PB, Hanson SJ, Haxby JV, Pollmann S (2009): PyMVPA: a Python Toolbox for Multivariate Pattern Analysis of fMRI Data. *Neuroinformatics* 7: 37–53.

5. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al.* (2011): Scikit-learn: Machine Learning in Python. *J Mach Learn Res* 12: 2825–2830.

6. Turner BO, Mumford JA, Poldrack RA, Ashby FG (2012): Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. *NeuroImage* 62: 1429–1438.

7. Fischl B, Van Der Kouwe A, Destrieux C, Halgren E, Ségonne F, Salat DH, *et al.* (2004): Automatically Parcellating the Human Cerebral Cortex. *Cereb Cortex* 14: 11–22.

8. Taschereau-Dumouchel V, Cortese A, Chiba T, Knotts JD, Kawato M, Lau H (2018): Towards an unconscious neural reinforcement intervention for common fears. *Proc Natl Acad Sci U S A* 115: 3470–3475.

9. Krishnapuram B, Carin L, Figueiredo MAT, Hartemink AJ (2005): Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Trans Pattern Anal Mach Intell* 27: 957–968.

10. Andersson JLR, Skare S, Ashburner J (2003): How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *NeuroImage* 20: 870–888.

11. Smith SM, Jenkinson M, Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, *et al.* (2004): Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage* 23 Suppl 1: S208-219.

12. Smith SM (2002): Fast robust automated brain extraction. *Hum Brain Mapp* 17: 143–155.

13. Jenkinson M, Bannister P, Brady M, Smith S (2002): Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage* 17: 825–841.

14. Jenkinson M, Smith S (2001): A global optimisation method for robust affine registration of brain images. *Med Image Anal* 5: 143–156.

15. Brett M, Anton J-L, Valabregue R, Poline J-B (2002): Region of interest analysis using an SPM toolbox. *NeuroImage* 16: 497.