Towards an unconscious neural reinforcement intervention for common fears

Vincent Taschereau-Dumouchel[a,b,1], Aurelio Cortese[a], Toshinori Chiba[a], J. D. Knotts[a,b], Mitsuo Kawato[a,c,1], and Hakwan Lau[a,b,d,e]

[a]Department of Decoded Neurofeedback, Computational Neuroscience Laboratories, Advanced Telecommunications Research Institute International, Kyoto 619-0288, Japan; [b]Department of Psychology, University of California, Los Angeles, CA 90095; [c]Faculty of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan; [d]Brain Research Institute, University of California, Los Angeles, CA 90095; and [e]Department of Psychology, University of Hong Kong, Hong Kong

Edited by Joseph E. LeDoux, New York University, New York, NY, and approved February 13, 2018 (received for review December 11, 2017)

Can "hardwired" physiological fear responses (e.g., for spiders and snakes) be reprogramed unconsciously in the human brain? Currently, exposure therapy is among the most effective treatments for anxiety disorders, but this intervention is subjectively aversive to patients, causing many to drop out of treatment prematurely. Here we introduce a method to bypass the subjective unpleasantness in conscious exposure, by directly pairing monetary reward with unconscious occurrences of decoded representations of naturally feared animals in the brain. To decode physiological fear representations without triggering excessively aversive reactions, we capitalize on recent advancements in functional magnetic resonance imaging decoding techniques, and use a method called hyperalignment to infer the relevant representations of feared animals for a designated participant based on data from other "surrogate" participants. In this way, the procedure completely bypasses the need for a conscious encounter with feared animals. We demonstrate that our method can lead to reliable reductions in physiological fear responses, as measured by skin conductance as well as amygdala hemodynamic activity. Not only do these results raise the intriguing possibility that naturally occurring fear responses can be "reprogrammed" outside of conscious awareness, importantly, they also create the rare opportunity to rigorously test a psychological intervention of this nature in a double-blind, placebo-controlled fashion. This may pave the way for a new approach combining the appealing rationale and proven efficacy of conventional psychotherapy with the rigor and leverage of clinical neuroscience.

physiological fear response | real-time functional magnetic resonance imaging | neural reinforcement

One of the most effective methods for the treatment of phobias involves exposure, or repeated approaches toward feared stimuli (1). Exposure-based therapies are effective in reducing symptoms, but their success depends on the individual's capacity or willingness to consciously confront feared objects. The associated distress can prevent patients from seeking treatment and contributes to attrition from exposure once treatment begins. For a variety of anxiety and trauma disorders, estimated overall dropout rates generally range from 0 to 52% (mean, 15.6%; median, 14.0%) (2); in some extreme cases, dropout rates as high as 70% have been reported (3).

Here we propose a potential solution to this dropout problem. Recent advancements in neuroimaging and machine learning have made it possible for us to identify specific representations of commonly and naturally feared animals in the human brain (4–6). We tested the hypothesis that despite the supposed deep evolutionary basis of these neural representations (7), we can unconsciously reprogram their associations to reduce the relevant physiological fear responses. Previously, using closed-loop fMRI neural reinforcement, we have shown that physiological fear responses can be reduced by pairing rewards with the unconscious occurrences of decoded object representations (8). However, in that study, the artificial objects were feared only because they had been experimentally conditioned with electric shocks, and that procedure was itself conscious. Here we tested whether our neural reinforcement method may apply to naturally occurring fear in everyday stimuli (e.g., images of spiders or snakes) entirely outside of participants' awareness.

The standard method for building these object decoders involves the presentation of relevant images to the subjects while fMRI pattern activity is recorded. However, this kind of procedure would lead us back to the problem of requiring subjects to consciously encounter the feared objects. To decode fear representations without triggering excessively aversive reactions, we capitalized on recent advancements in fMRI decoding techniques and used a method called hyperalignment (9, 10) to infer the relevant representations of feared animals for a designated participant based on data from other "surrogate" participants. In this way, the procedure completely bypasses the need for conscious exposure.

Results

Study Outline. The first phase of this study involved an fMRI session (which we call the decoder construction session) that allowed us to determine the decoded animal representations to

Significance

Conventional therapies for the treatment of anxiety disorders are aversive, and as a result, many patients terminate treatment prematurely. We have developed an unconscious method to bypass the unpleasantness in conscious exposure using functional magnetic resonance imaging neural reinforcement. Using this method, participants learn to generate brain patterns similar to the multivariate brain pattern of a feared animal. We demonstrate in a double-blind placebo-controlled experiment that neural reinforcement can lead to reliable reductions in physiological fear responses. Crucially, this intervention can be achieved completely unconsciously and without any aversive reaction. Extending our approach to other forms of psychopathologies, such as posttraumatic stress disorders, might eventually provide another means of intervention for patients currently receiving insufficient exposure treatments.

Author contributions: V.T.-D., A.C., J.D.K., M.K., and H.L. designed research; V.T.-D., A.C., T.C., and J.D.K. performed research; V.T.-D. analyzed data; and V.T.-D., A.C., T.C., J.D.K., M.K., and H.L. wrote the paper.

Conflict of interest statement: M.K. is the inventor of patents related to the DecNef method used in this study, and the original assignee of the patents is Advanced Telecommunications Research Institute International, with which the authors are affiliated.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: vincenttd@ucla.edu or kawato@atr.jp.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1721572115/-/DCSupplemental.

Published online March 6, 2018.

target during neural reinforcement. The second phase involved a 5-d neural reinforcement session as well as two brief fMRI sessions conducted before and after neural reinforcement to assess the efficacy of the intervention. The experiment was conducted in six sessions carried out on different days: day 0, decoder construction session; day 1, prereinforcement session and neural reinforcement; days 2–4, neural reinforcement; day 5, neural reinforcement and postreinforcement session.

Building Accurate Hyperalignment Decoders. To construct across-subjects machine learning decoders for some of the most commonly feared animals, we designed an experiment (day 0, decoder construction session) in which normal healthy participants viewed 3,600 images of 30 different animals and 10 inanimate objects (Fig. 1*A*). The multivoxel fMRI responses to these individual images were recorded in the fMRI scanner and extracted at the single-trial level, using conventional analytic procedures (*SI Appendix*).

We then capitalized on a novel method of across-subjects multivoxel analysis known as hyperalignment, which allowed us to compare and translate patterns of fMRI activity between participants (10). Using this method, we exploited the data from as many as 29 "surrogate" participants (who viewed the target feared category) to construct the hyperalignment decoder for a designated participant as if he or she was never exposed to the target images (as would be ideal in a clinical setting) (Fig. 1*B*). To do so, an abstract common space was derived from the voxel activity in the ventral temporal cortex (fusiform, inferior temporal, lingual/parahippocampal cortex) (10) of each participant based on all image categories except the target category (e.g., snake). This allowed us to infer the decoders for the designated participant based on the decoders of surrogate participants (*SI Appendix*, Fig. S1). Specifically, we trained a decoder to discriminate multivoxel patterns for the target category from patterns for all of the other nontarget categories in the surrogate participants (target vs. nontarget), and through transformations via the common space, we inferred what such a decoder would be for the designated participant. We call this the hyperalignment decoder.

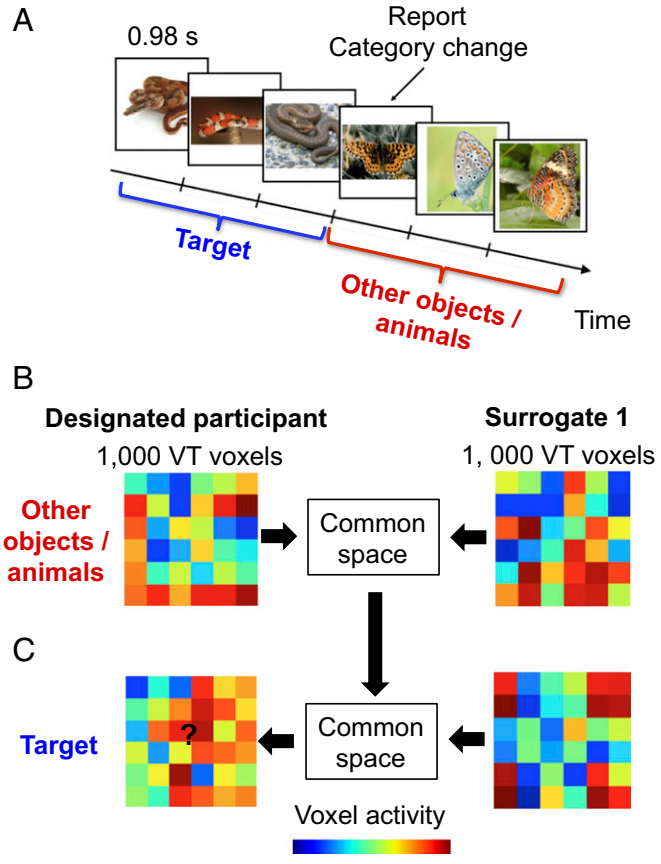

Fig. 1. Hyperalignment decoder construction. (*A*) An example sequence of events in the hyperalignment decoder construction session. (*B*) To mimic the situation wherein patients are not exposed to target images (to avoid excessive fear), the construction of hyperalignment decoders was based on the data from surrogate participants. To do so, we hyperaligned voxels in the ventral temporal area between a designated participant and surrogates into a "common space," using the representations for the 39 nontarget categories. (*C*) Through this space, we created a hyperalignment decoder for a target category in the designated participant based on the surrogates' representations for that target category.

One may worry that such an indirect inference strategy may provide only limited efficiency. However, for each designated participant, this procedure can benefit from the data of as many as 29 surrogate participants. As such, we harness the power of a much larger amount of data to train the decoders than is used in conventional (i.e., within-subject) fMRI decoding.

As in previous reports, here the hyperalignment decoders displayed decoding accuracies (mean, 82.4 ± 1.73% SD) even higher [t (39) = 9.55, $P < 0.0001$, two-sided t test; Wilcoxon signed-rank test: $z = 4.76$, $P < 0.0001$] than accuracies obtained with traditional within-subject decoders trained by presenting images to the participants themselves (mean, 71.7 ± 6.41% SD) (Fig. 2*B*). These results are also in accordance with previous data indicating that the topography of voxel selectivity appears to be preserved by hyperalignment (10) (Fig. 2*A* and *SI Appendix*, Table S1).

Importantly, whether the designated participant presented a similar fear profile (over all 40 categories of animals/objects) to the average fear profile of all participants in the hyperalignment did not modulate the accuracy of the hyperalignment decoder (*SI Appendix*, Fig. S5). Furthermore, we explicitly determined whether hyperalignment decoders built from "normal surrogate" participants could generalize to patients diagnosed with specific phobias, by recruiting patients to take part in our decoder construction session. Here we recruited three patients who met the *Diagnostic and Statistical Manual of Mental Disorders*, Fourth Edition criteria for specific phobia for at least one animal in our database and determined how their fear profiles affected the accuracy of the hyperalignment decoders. Here hyperalignment decoders constructed with our group of normal surrogate participants had similar decoding accuracy levels for patients diagnosed with a specific phobia as those established within the group of normal surrogates. This was specifically true for the phobic category (all within ±1 STD; *SI Appendix*, Fig. S6). This suggests that our method may be promising even for patients with atypical fear profiles.

Decrease in Physiological Fear Response Following Neural Reinforcement. We then conducted 5 d of our neural reinforcement procedure (Fig. 3*A*) using these hyperalignment decoders in 17 participants who presented with high (subclinical) levels of fear for at least two animal categories in our database. Details on the selection process are provided in *SI Appendix*. For each participant, one of the feared animal categories was selected at random by the computer to be the target of the intervention, while the other of the feared animal categories acted as a within-subject control to allow us to determine the specificity of the intervention. Physiological fear response was assessed before and after neural reinforcement sessions by presenting images of the two feared categories (target and control) and images of a neutral category (baseline condition) while measuring skin conductance response and amygdala hemodynamic response.

In each trial during the neural reinforcement procedure (Fig. 2*C*), the hyperalignment decoder was applied to fMRI images in real time to determine the likelihood that the pattern of activity corresponding to the target category was represented in the brain

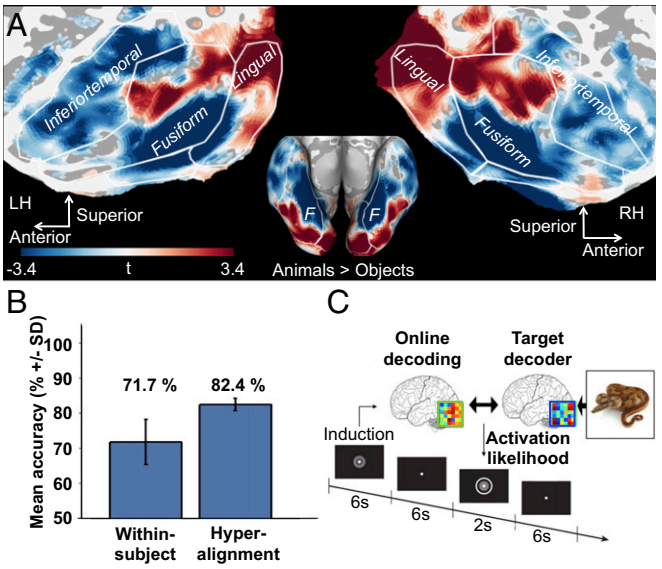

Fig. 2. Classification accuracies and neural reinforcement procedure. (*A*) Voxels contributing to the hyperalignment decoders. Plotted are the *t* values of the voxels' weights, which are consistent with a lateral-to-medial animacy continuum (10, 22) (critical *t* value = 3.4 for $P < 0.001$, uncorrected). (*B*) Hyperalignment decoders built using 29 surrogate participants had better accuracy than conventional within-subject methods. Error bars represent ±SD. (*C*) During neural reinforcement, online decoding was used to reinforce occurrences of the target (but not the control) multivoxel representation. The feedback was proportional to the likelihood of the target being represented, and to monetary gain. Crucially, both participants and experimenters were unaware of the identity of the target category throughout the experiment.

(*SI Appendix*, Fig. S2). This information was visually fed back to participants by varying the size of a disk image from trial to trial, which was directly proportional to the amount of money that the subject would earn on a trial. Following previous decoded neurofeedback (DecNef) procedures (8, 11–17), participants were explicitly informed of the association between disk size and monetary reward but received no instructions as to what brain activity patterns were necessary to maximize the size of the disk. Despite this, participants were able to learn to activate the target representation with statistical consistency above chance (*SI Appendix*, Fig. S2*A* and *SI Discussion*). This process was thus conducted in a double-blinded fashion, as neither the participants nor the experimenters were aware of the identity of the target category.

Confirming our hypothesis, we found a specific reduction of physiological fear response for the target category after neural reinforcement (Fig. 3 *B* and *C*): Amygdala response decreased for the target condition [t (16) = 2.41; $P = 0.028$] but remained unchanged for the control condition [t (16) = 0.40; $P = 0.69$], and showed a significant time-by-condition interaction [$F(1,16) = 5.57$; $P = 0.031$]. We likewise observed a significant decrease in skin conductance response for the target condition [$t(122) = -2.48$; $P = 0.014$], but not for the control condition [$t(122) = 0.016$; $P = 0.99$], with a significant time-by-condition interaction [$F(1,244) = 2.13$; $P = 0.033$]. The effect sizes were considered of medium size for both the amygdala (Cohen's $d = 0.62$) and the skin conductance response (Cohen's $d = 0.55$) (corrected for dependence between the means) (18) and are similar to effect sizes of exposure therapy (Cohen's $d = 0.42$–0.68) (19).

Importantly, at the end of the procedure, participants were unable to guess the identity of the target category (47% accuracy in a two-alternative forced-choice question), and reported strategies for maximizing rewards that were generally unrelated to the target and purpose of the procedure. (*SI Appendix*, Table S2 presents induction strategies reported by participants.) This confirms that our treatment effects can be obtained outside of participants' conscious awareness (*SI Appendix*, Figs. S3 and S8).

What are the possible mechanisms underlying these results? To better understand the nature of neural reinforcement, we conducted information transmission analyses (8, 12, 14) to investigate whether multivoxel patterns in other brain regions can predict, on a trial-by-trial basis, the likelihood of target representation in the ventral temporal cortex. These analyses indicated that during neural reinforcement, the voxels tracking the likelihood of target representation were contained primarily in the fusiform, inferior temporal, and lingual cortices (Fig. 4). These results were compared with normal conscious viewing of target images (day 0, decoder construction session), wherein the voxels tracking the decoders' likelihood were distributed more broadly in the fusiform and lingual regions, as well as outside of the ventral temporal cortex in areas such as the amygdala (*SI Appendix*, Fig. S4), cuneus, and parietal and occipital regions (Fig. 4). Overall, these observations are consistent with our previous report showing that during neural reinforcement, the induced target representations were relatively localized and disconnected from the rest of the fear-related circuitry. Such a disconnect may be an important aspect of our fear reduction procedure (8).

Discussion

Using our neural reinforcement method, we have shown that training participants to activate the neural representations of a feared animal can lead to a decrease in the relevant physiological fear response. This intervention appears to have a specific effect on the targeted animal, as the physiological fear response to a feared control animal remained relatively unchanged.

The fact that this procedure can be achieved unconsciously provides some significant advantages. First, our procedure can be achieved without triggering conscious aversive reactions, which might help decrease physiological fear reactivity in individuals who are unwilling to enroll in conventional treatment or those who dropped out of treatment because of the aversive nature of the intervention. In the present study, because the

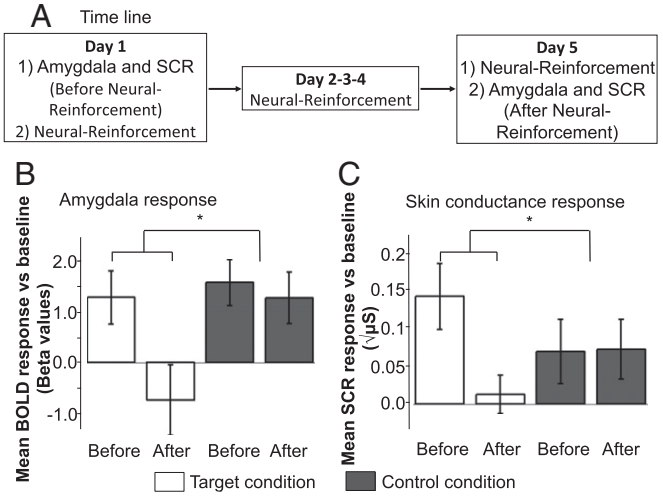

Fig. 3. Decrease in physiological fear responses following neural reinforcement. (*A*) To assess changes in physiological fear responses, on days 1 and 5, participants viewed images of two animal categories that they feared (target and control) and one animal that they did not fear (baseline). (*B*) The results indicate a significant decrease in amygdala response for the target condition while the control condition remained unchanged by the procedure. (*C*) Likewise, the results indicate a significant decrease in skin conductance response in the target condition and no decrease in the control condition. BOLD, blood oxygen level-dependent; SCR, skin conductance response. Error bars represent ±1 SEM. *$P < 0.05$.

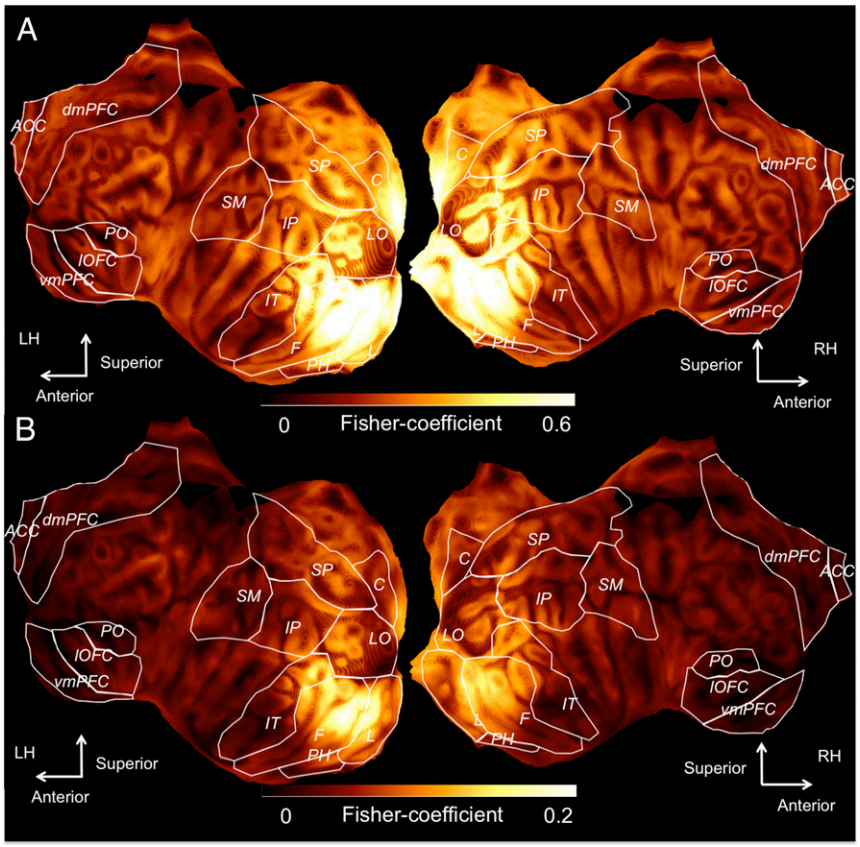

Fig. 4. (*A*) Information transmission analyses during normal conscious perception (decoder construction) and during neural reinforcement (unconscious occurrences of target). (*B*) In a searchlight procedure (39), sparse linear regression was used to predict the linearized likelihood of target representation in the ventral temporal cortex from multivoxel patterns within each sphere. Plotted in the MNI space are the mean Fisher-transformed correlation coefficients representing the accuracy of this prediction. The maximums of data scales were adjusted to reflect significant voxels determined using a permutation test. Overall transmission was lower during neural reinforcement than in normal conscious viewing. ACC, anterior cingulate cortex; C, cuneus; dmPFC, dorsomedial prefrontal; F, fusiform; lOFC, lateral orbitofrontal; IP, inferior parietal; IT, inferior temporal; LO, lateral occipital; PH, parahippocampal; PO, pars orbitalis; SM, supramarginal; SM, supramarginal gyrus; SP, superior parietal; TP, temporal pole.

purpose is one of proof of concept, participants in fact saw the images of the feared animals during the decoder construction session, which helped demonstrate high performance in the actual accuracy of the hyperalignment decoders. However, we found that these trials are not necessary for decoder construction if we apply the hyperalignment procedure, meaning that future interventions could be conducted completely without presenting these aversive images. One may worry whether these decoders could be sufficiently general to be relevant to actual everyday encounters with the target animals, given that they are constructed based on a specific set of images (90 per category); however, numerous lines of evidence suggest that voxel patterns in the ventral temporal cortex identified in similar fashion represent generic object category information not specific to particular images or viewpoints (10, 20–22). In the context of neural reinforcement, the mechanisms for the generalizability of these effects have been discussed in a recent review (17). Second, the fact that the participants are unaware of the nature of the procedure could help prevent the use of conscious "safety signals" or "safety behaviors" during exposure, which are known to interfere with the efficacy of exposure-based therapies (23).

Perhaps most importantly, our procedure was conducted in a double-blinded fashion, which in itself is a methodological advancement, since conventional psychological and neurofeedback interventions can rarely be conducted at this level of experimental rigor. In general, many psychotherapeutic treatments are known to be effective (19), but conducting double-blind placebo controls to assess the effectiveness of such treatments is challenging, owing to the very nature of therapy. The current experimental approach may be extended to other psychotherapeutic methods, to allow more complete integration of clinical psychology with standard medicine.

Despite these advantages of our "unconscious" intervention, it has also been suggested that this kind of procedure may impact only the physiological reactivity to feared stimuli, not necessarily the behavioral outcomes and subjective experiences (24). Strictly speaking, such physiological responses may be more correctly identified as "threat"-related rather than reflective of the subjective experience of fear per se (25). Our results are exactly in accordance with this prediction. Nonetheless, it is worth noting that the present study is one of an experimental demonstration, an investigation of basic mechanisms. In implementing the double-blinded procedure, we aimed to ensure that the procedure was fully unconscious; as such, we did not inform the participants about the purpose of the experiment. In contrast, had our purpose been merely to pass standard placebo-control requirements, it would have been sufficient to inform the subjects that there was a 50-50 chance of receiving placebo intervention or of receiving a specific intervention, i.e., fear reduction for a particular target object. In that scenario, perhaps the level of partial awareness with uncertainty may be sufficient to bring the underlying physiological changes to a conscious level. Ultimately, for the method to be applied in the clinic, patients will be fully aware of the purpose of the procedure while they are receiving treatment as well. While these considerations do not undermine the validity and intended scope of the current study, it would be advantageous and interesting for future studies to test how our intervention procedure may be combined with conventional (conscious) psychotherapeutic treatments to produce synergistic and long-lasting effects on clinical outcomes as well as on conscious experiences. As in studies of other mental illnesses, one finding is that combining different methods sometimes leads to the best overall outcomes. Such is the case in depression, where the combination of medication (i.e., serotonin reuptake inhibitors) and cognitive-behavioral therapy has been shown to be advantageous (26). As such, neural reinforcement may benefit from concurring cognitive restructuring treatments as well (27). In such a scenario, the self-monitoring of changes in physiological fear reactivity might be a key aspect to emphasize

to potentiate the effect of neural reinforcement to bring such changes to the level of conscious awareness.

In conclusion, we have exploited an opportunity to apply recent advances in fMRI decoding to move one step closer to our ultimate goal of creating a method for an unconscious brain-based psychotherapy for anxiety disorders. This study provides the first evidence that physiological fear responses to specific, subclinical, naturally occurring fears can be reduced unconsciously with hyperalignment decoders, completely outside of the awareness of human subjects. The staggering progress of current neuroimaging decoding technology (4–6), combined with in-scanner virtual reality experiments (28, 29), may mean that we can eventually extend our approach to other forms of fear, such as acrophobia, anxiety induced by public speaking, fear associated with specific persons or episodic memories, and so on. Future studies should rigorously test whether the current neural reinforcement approach may generalize to different forms of anxiety-related illnesses. In particular, for posttraumatic stress disorder, it has been estimated that as few as 2% of patients receive sufficient treatment in the form of traditional exposure (30–32). Our unconscious brain-based method may eventually alleviate this challenging and critical problem of patient attrition and pave the way for a novel approach combining the appealing rationale and proven efficacy of conventional psychotherapy with the rigor and leverage of clinical neuroscience.

Methods

Participants. Thirty participants (eight females, mean age 24.0 ± 3.97 years) took part in a decoder construction session and were included in the hyperalignment procedure. Seventeen participants (five females, mean age 21.92 ± 1.54 years) also took part in the neural reinforcement experiment. Participants in the neural reinforcement experiment first participated in the decoder construction session and were selected for neural reinforcement if they reported, on a 7-point Likert scale, "high" or "very high" fears of at least two animals included in the database. We predetermined the number of participants based on a previous study (8). The experiment was conducted in a double-blinded fashion; i.e., neither the participants nor the experimenters were aware of the target category of the neural reinforcement procedure. All participants provided written informed consent, and the study was approved by the Institutional Review Board of Advanced Telecommunications Research Institute International, Japan.

Decoder Construction Session and Hyperalignment. Hyperalignment decoders were trained to discriminate the brain representation of a feared animal from those of other animals and objects. To do so, each participant underwent a 1-h fMRI decoder construction session during which they were presented with images of various animals and objects. We aimed to present images from 40 categories because previous studies have shown that similar numbers of object/animal categories can be decoded from fMRI patterns (21, 33). For this decoding to be robust, each category requires a reasonably high number of samples. By presenting images for 0.98 s each, it was possible to present a total of 3,600 images within an imaging session of typical duration (approximately 1 h). To optimize the tradeoff between the number of different categories sampled and the number of trials in each category, we chose to present 90 different images per category and to include 30 animal categories and 10 object categories (*SI Appendix*, Fig. S7 and *SI Methods*).

We constructed the target decoders for a designated participant from the data of 29 surrogate participants. This method was chosen to determine how effective this procedure could be if participants were never exposed to the target category during the decoder construction session, as would be ideal in a clinical setting. To do so, we iteratively performed a new hyperalignment for each category and for each participant.

We conducted this procedure on the voxels of the ventral temporal cortex (fusiform, lingual/parahippocampal, and inferior temporal cortex), as described in more detail in *SI Appendix*. We first set aside, for each designated participant, the multivoxel patterns elicited by the target category plus an equal number (90 trials) of randomly selected patterns associated with the remaining nontarget categories. This was done to prevent circularity (34), as the set-aside data for the designated participant was later used to test the accuracy of the hyperalignment decoders. The remaining data from all participants were used to carry out hyperalignment and to develop the abstract common decoder space. This procedure involved determining a set of geometric transformations (rotation, translation, and scaling) that brought data from the native space of each participant (where individual voxels are dimensions) into a common space where brain representations could be optimally aligned between participants. Importantly, this transformation can be reversed such that data from the common space can be projected back into the native space of participants. Another important feature of hyperalignment is that new data can be transformed into the common space even if they were previously withheld from hyperalignment. We capitalized on both of these features to build our training dataset; we brought all of the data from all participants (which included the target category previously set aside) back into the native space of the designated participant by first transforming it into the common space. This allowed us to construct the hyperalignment decoder in the native space of the designated participant. The decoder was trained to discriminate target trials from nontarget trials using the data of 29 surrogate participants and was tested on the data of the target participant (Fig. 1). Based on previous procedures (10), hyperalignment was conducted in pyMVPA 2.4 (www.pymvpa.org) in the NeuroDebian environment (35).

We used sparse logistic regression (36, 37) to select the most discriminant voxels for the target category (average of 141.9 voxels; SEM ±4.0) and to identify a linear hyperplane that would maximally separate voxel patterns associated with the target category from those associated with the randomly selected nontarget images. We trained these decoders on the data from the surrogate participants averaged within runs (six runs) and categories. Thus, the training dataset consisted of 348 multivoxel patterns distributed over 1,000 ventral temporal voxels. The performance of the hyperalignment decoders was then tested using the multivoxel patterns of the designated participant that had been held out from hyperalignment and decoder construction (i.e., the 90 trials of the target category and 90 trials selected at random from the nontarget categories). Fig. 2*A* shows the contrast of the sparse logistic regression weights on each voxel between animal vs. object categories (*SI Appendix, SI Methods*). Fig. 2*B* shows the accuracies of the hyperalignment decoders constructed iteratively with 29 surrogate participants and averaged over the 30 participants and the 40 object categories.

Prereinforcement and Postreinforcement Sessions. To assess changes in physiological fear responses, we used brief visual presentations of animals from two feared categories (i.e., target and control categories), before and after the neural reinforcement sessions. Each session included the presentation of 30 images divided in two short blocks: 10 images of the target condition, 10 images of the control condition, 5 images of a neutral animal (determined on a 7-point Likert scale), and 5 images of a neutral object. These presentations were carried out in the MRI scanner while electrodermal activity (i.e., skin conductance response) and fMRI images were acquired. The images presented during prereinforcement and postreinforcement were never presented during the hyperalignment procedure and were created following the same procedure (*SI Appendix, SI Methods*). Each trial included the presentation of a fixation cross for 3–7 s (mean, 5 ± 2 s), presentation of the image for 6 s, and then a blank screen for 4–12 s (mean, 8 ± 3 s). Each block started with 20 s of rest, followed by the presentation of the image of a neutral object (e.g., a chair). The next two images were randomly set to be from the target or control category, and their order was counterbalanced between blocks. The remaining images were then presented randomly during the rest of the block. To estimate physiological fear responses, we built on previous methodologies (8, 38), and calculated skin conductance and amygdala responses during the first two trials of the two feared categories within each block. These mean responses were then baseline-corrected by subtracting the mean responses to the neutral animal category. More details are provided in *SI Appendix*.

Neural Reinforcement Session. The aim of the neural reinforcement sessions was to allow participants to associate a reward with the activation of the neural representation of a feared animal in the ventral temporal cortex (target category). To select participants for neural reinforcement, we chose individuals who self-reported "high" or "very high" fear of at least two animals in our database. One of these two animal categories was randomly selected (by a computer) to be the target of the intervention and the other was selected to be the control condition. This within-subject control condition allowed for a double-blinded procedure, since neither the experimenters nor the participants were aware of the target of the intervention during the procedure. Since participants frequently reported high fears of more than two categories (average of 4.8 feared categories; SEM ±0.86), we selected within the "high" and "very high" fear categories the two categories presenting the hyperalignment decoders with the highest accuracies. For two participants, the multivoxel representations of these two categories were also correlated with one another ($r > 0.25$ both within subjects and in

the hyperalignment data of surrogate participants). In these situations, we selected (*i*) the feared category with the highest accuracy and (*ii*) the next feared category in terms of accuracy that was not correlated with the first category selected ($r < 0.25$).

Online neural reinforcement was conducted across five sessions on five different days. Following previous procedures (8, 12–15), each trial began with an induction period (6 s), during which participants were instructed to "activate a pattern in their brain" to maximize the size of a subsequently presented feedback disk (i.e., the diameter of the inner gray circle) (Fig. 2*C*). Online decoding was achieved using the hyperalignment decoder for the target category, while the decoder for the control category was never used for reinforcement. The diameter of the circle during the feedback period was a function of the "activation likelihood" of the target category, i.e., the likelihood that the hyperalignment decoder predicted the target category from the multivoxel pattern in ventral temporal cortex. Participants were informed that their monetary gain would be a function of the overall success in correctly activating brain patterns (i.e., activation likelihood) during each session, but—critically—they were not told what the target multivoxel pattern represented.

Information Transmission Analyses. The hyperalignment decoder likelihood computed based on voxels in the ventral temporal region could be associated with the transmission of information to other brain regions. This process could occur both during the actual presentation of the target category during decoder construction as well as during pattern induction in the neural reinforcement sessions. To compare the flow of information in the brain between the decoder construction and induction phases, we used information transmission analysis (8, 12–14). This analysis uses a searchlight approach (39) in which a sphere (radius, 15 mm; mean, 266 voxels) is iteratively centered around each voxel of the gray matter mask in the native space of each participant. Within each sphere, sparse linear regression machine learning classification is used to determine if it is possible to use the activity of the voxels within the sphere to predict, on a trial-by-trial basis, the linearized induction likelihood for the target category predicted by the hyperalignment decoder constructed in the ventral temporal cortex. The predicted values are then correlated with the true linearized likelihood of the ventral temporal decoders. The correlation coefficients for each sphere are then Fisher-transformed and assigned to the central voxel of the sphere. The coefficients were then projected in the MNI space and smoothed using a Gaussian filter (FWHM = 6 mm). The results of the information transmission analysis are presented on the MNI brain during decoder construction (Fig. 4*A*) and during neural reinforcement (Fig. 4*B*). PyCortex (40) was used for data presentation in Fig. 4. More information is provided in *SI Appendix*.

Data supporting the findings of this study, along with the custom code used to generate these data, are available from the corresponding authors.

ACKNOWLEDGMENTS. We thank K. Nakamura and M. Miuccio for their help with scheduling and conducting the experiment; N. Hiroe and H. Moriya for assisting with the equipment; Y. Shimada and A. Nishikido for operating the fMRI scanner; K. Ide for helping with recruitment of patients; B. Maniscalco, M. Peters, and B. Odegaard for comments on the manuscript; and M. Craske, S. Guntupalli, and M. Sun for discussions about the experiments. 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. This study was also partly funded by the US National Institute of Neurological Disorders and Stroke of the National Institutes of Health (Grant R01NS088628, to H.L.). V.T-D. is supported by a fellowship from the Fond de Recherche du Québec-Santé. M.K. is the inventor of patents related to the DecNef method used in this study, and the original assignee of the patents is Advanced Telecommunications Research Institute International, with which some of the authors are affiliated.

1. Craske MG, et al. (2008) Optimizing inhibitory learning during exposure therapy. *Behav Res Ther* 46:5–27.
2. Loerinc AG, et al. (2015) Response rates for CBT for anxiety disorders: Need for standardized criteria. *Clin Psychol Rev* 42:72–82.
3. Zayfert C, et al. (2005) Exposure utilization and completion of cognitive behavioral therapy for PTSD in a "real world" clinical practice. *J Trauma Stress* 18:637–645.
4. Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci* 37:435–456.
5. Diedrichsen J, Kriegeskorte N (2017) Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput Biol* 13:e1005508.
6. Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532:453–458.
7. Ohman A, Mineka S (2001) Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychol Rev* 108:483–522.
8. Koizumi A, et al. (2016) Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nat Hum Behav* 1:0006.
9. Guntupalli JS, et al. (2016) A model of representational spaces in human cortex. *Cereb Cortex* 26:2919–2934.
10. Haxby JV, et al. (2011) A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron* 72:404–416.
11. Sitaram R, et al. (2017) Closed-loop brain training: The science of neurofeedback. *Nat Rev Neurosci* 18:86–100.
12. Amano K, Shibata K, Kawato M, Sasaki Y, Watanabe T (2016) Learning to associate orientation with color in early visual areas by associative decoded fMRI neurofeedback. *Curr Biol* 26:1861–1866.
13. Cortese A, Amano K, Koizumi A, Kawato M, Lau H (2016) Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat Commun* 7:13669.
14. Shibata K, Watanabe T, Sasaki Y, Kawato M (2011) Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* 334:1413–1415.
15. Shibata K, Watanabe T, Kawato M, Sasaki Y (2016) Differential activation patterns in the same brain region led to opposite emotional states. *PLoS Biol* 14:e1002546.
16. Alegria AA, et al. (2017) Real-time fMRI neurofeedback in adolescents with attention deficit hyperactivity disorder. *Hum Brain Mapp* 38:3190–3209.
17. Watanabe T, Sasaki Y, Shibata K, Kawato M (2017) Advances in fMRI real-time neurofeedback. *Trends Cogn Sci* 21:997–1010.
18. Morris SB, DeShon RP (2002) Combining effect size estimates in meta-analysis with repeated-measures and independent-groups designs. *Psychol Methods* 7:105–125.
19. Wolitzky-Taylor KB, Horowitz JD, Powers MB, Telch MJ (2008) Psychological approaches in the treatment of specific phobias: A meta-analysis. *Clin Psychol Rev* 28: 1021–1037.
20. Haxby JV, et al. (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–2430.
21. Kriegeskorte N, et al. (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60:1126–1141.
22. Grill-Spector K, Weiner KS (2014) The functional architecture of the ventral temporal cortex and its role in categorization. *Nat Rev Neurosci* 15:536–548.
23. Sloan T, Telch MJ (2002) The effects of safety-seeking behavior and guided threat reappraisal on fear reduction during exposure: An experimental investigation. *Behav Res Ther* 40:235–251.
24. LeDoux J (2015) *Anxious: The Modern Mind in the Age of Anxiety* (Oneworld Publications, London).
25. LeDoux JE (2017) Semantics, surplus meaning, and the science of fear. *Trends Cogn Sci* 21:303–306.
26. Hollon SD, et al. (2014) Effect of cognitive therapy with antidepressant medications vs. antidepressants alone on the rate of recovery in major depressive disorder: A randomized clinical trial. *JAMA Psychiatry* 71:1157–1164.
27. Craske MG, Barlow DH (2005) *Mastery of Your Anxiety and Worry: Client Workbook* (Oxford Univ Press, New York), 2nd Ed.
28. Powers MB, Emmelkamp PMG (2008) Virtual reality exposure therapy for anxiety disorders: A meta-analysis. *J Anxiety Disord* 22:561–569.
29. Parsons TD, Rizzo AA (2008) Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *J Behav Ther Exp Psychiatry* 39: 250–261.
30. Watts BV, et al. (2014) Implementation of evidence-based psychotherapies for posttraumatic stress disorder in VA specialty clinics. *Psychiatr Serv* 65:648–653.
31. Belleau EL, et al. (2017) Pre-treatment predictors of dropout from prolonged exposure therapy in patients with chronic posttraumatic stress disorder and comorbid substance use disorders. *Behav Res Ther* 91:43–50.
32. Najavits LM (2015) The problem of dropout from "gold standard" PTSD therapies. *F1000Prime Rep* 7:43.
33. Charest I, Kievit RA, Schmitz TW, Deca D, Kriegeskorte N (2014) Unique semantic space in the brain of each beholder predicts perceived similarity. *Proc Natl Acad Sci USA* 111:14565–14570.
34. Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI (2009) Circular analysis in systems neuroscience: The dangers of double dipping. *Nat Neurosci* 12:535–540.
35. Halchenko YO, Hanke M (2012) Open is not enough. Let's take the next step: An integrated, community-driven computing platform for neuroscience. *Front Neuroinform* 6:22.
36. 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.
37. Yamashita O, Sato M-A, Yoshioka T, Tong F, Kamitani Y (2008) Sparse estimation automatically selects voxels relevant for the decoding of fMRI activity patterns. *Neuroimage* 42:1414–1429.
38. Schiller D, et al. (2010) Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature* 463:49–53.
39. Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103:3863–3868.
40. Gao JS, Huth AG, Lescroart MD, Gallant JL (2015) Pycortex: An interactive surface visualizer for fMRI. *Front Neuroinform* 9:23.