Implicit Decoded Neurofeedback training as a clinical tool

Ai Koizumi\(^1,2,3\) & Mitsuo Kawato\(^1,4\)

1. The Department of Decoded Neurofeedback, Computational Neuroscience Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan
2. Sony Computer Science Laboratories, Inc., Tokyo, Japan
3. Graduate School of Media and Governance, Keio University
4. RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

Abstract
This chapter will concentrate on the implicit nature of Decoded Neurofeedback (DecNef), which can serve as a potential alternative to exposure therapy or counter conditioning for alleviating fear and anxiety. DecNef has been shown to be successful in reducing fear responses to specific stimuli by inducing neural representations of feared stimuli without inducing conscious awareness of the stimuli. Such implicit DecNef may be suitable for the treatment of PTSD, phobia and other anxiety disorders, especially when patients are intolerant to aversive conscious exposure to their feared stimuli during conventional therapies. Our hope is not to replace the existing clinical practice with DecNef training, but rather to make some additional treatment options available to patients so that they have more flexibility in choosing how they would like to overcome their issues with fear and anxiety.
**Implicit nature of Decoded neurofeedback (DecNef)**

Decoded Neurofeedback (DecNef) is a relatively novel type of neurofeedback which allows real-time neurofeedback based on spatial activation patterns in targeted brain areas (Watanabe et al., 2018). Combined with fMRI multivoxel decoding techniques to dissociate different neural representations (e.g., Yamashita et al., 2008), DecNef provides feedback regarding the likelihood that a certain representation is activated in a given brain area. For example, it provides feedback to a participant based on whether his or her visual cortex is likely to be representing a red rather than a green target stimulus (Amano et al., 2016, Koizumi et al., 2016) or whether their prefrontal areas are likely to be representing a high rather than a low state of perceptual confidence (Cortese et al., 2017).

One of the features which makes DecNef distinguishable from many other neurofeedback techniques is its implicitness (Watanabe et al., 2018). That is, participants remain unaware of what neural representation is being induced throughout DecNef training sessions (Figure 1). They are generally instructed to somehow manipulate their brain activation in order to receive better feedback (e.g., a circle size on a screen) through trial-and-error. Yet, they are never provided with any explicit instruction e.g., to imagine a red stimulus. Despite the fact that participants generally learn to successfully induce the targeted brain activation patterns, post-training questionnaires and/or forced-choice questions have consistently revealed that participants do remain unaware of the identity of targeted neural activation patterns (Amano et al., 2016, Cortese et al., 2017, Koizumi et al., 2016, Shibata et al., 2011, Taschereau-Dumouchel et al., 2018) (Shibata et al., 2019) for review.

The fact that participants remain unaware of the induced neural representations may sound surprising to some. However, it is in line with a large body of literature on consciousness: Neural representations in confined brain areas need to be connected with wider parts of the brain to reach conscious awareness (Brown et al., 2019), although the exact mechanisms involved in consciousness remain a matter of active debate. Previous studies with DecNef have shown that the neural representation induced in a targeted brain area is highly localized (Shibata et al., 2011). For example, even when an intended stimulus representation of a right tilted grating is successfully induced in the early visual cortex targeted during DecNef training, activation patterns in other brain areas often fail to corepresent the right tilted grating (Shibata et al., 2011). Because the induced representation is generally confined to the targeted brain area, this representation may remain outside of conscious awareness.

![Figure 1](image_url)  
*Figure 1. Schematics of the implicit DecNef training to help overcome fear and anxiety.*

DecNef may benefit the treatment of fear-related disorders with its implicit nature.
Its implicit nature makes DecNef a unique tool to potentially treat anxiety disorders such as post-traumatic stress disorder and phobias. One of the common treatments for anxiety disorders is exposure therapy. During exposure therapy, a patient often experiences strong aversive emotions because he or she needs to observe or imagine the scenes and objects related to their disorders, e.g., an image of a car that reminds them of a traumatic car accident. Exposure therapy is thought to rely on a fear extinction process through which a participant eventually learns that the feared objects are no longer associated with pain or trauma. Despite its effectiveness, the distress of therapy leads to a non-negligible rate of dropouts, with an estimated range from 0 to as high as 70 % (Loerinc et al., 2015, Zayfert et al., 2005).

DecNef, with its implicitness, may help overcome such issue of distress and benefit the treatment of anxiety disorders. Specifically, instead of explicitly viewing or imagining a feared object, a patient could unknowingly induce a neural representation of the feared object through DecNef training. Repeated implicit reactivation of the neural representation of a feared object, when paired with reward as described below, may eventually alleviate fear towards the object in a manner similar to explicit exposure therapy. Here we describe a few studies that have directly examined this possibility.

**Progressive development of DecNef as a clinical tool**
The first study examined the effectiveness of DecNef for reducing fear-like responses among healthy participants (Figure 2. (Koizumi et al., 2016)). In this study, participants initially underwent a fear conditioning session to obtain fear-like response to two colored stimuli, red and green gratings, both of which were paired with uncomfortable electric shocks. Participants then went through a three-day DecNef session in which they were provided with rewarding feedback when their early visual cortices (V1/V2) were more likely to be representing either the red or green stimulus. Which color served as their target stimulus was counterbalanced across participants and was fixed throughout the session. Better induction of the activation patterns representing the target stimulus was followed by a larger disc presented on the monitor, whose size was proportional to the actual monetary reward provided to the participants at the end of each day.

After the three-day DecNef session, participants underwent a test session in which they were explicitly presented with the fear conditioned red and green stimuli. In the test session, participants showed reduced skin conductance responses and amygdala activation to the stimulus that served as target and had its neural representation paired with reward during the DecNef training, compared to the control stimulus whose neural representation was not paired with reward during DecNef training.

Importantly, when participants were asked to guess which color, red or green, served as their target neural representation after completing all the sessions, they could only guess the target color at chance level. Moreover, fear-like responses to the neurally induced activation patterns of feared objects during the DecNef session were significantly smaller than the responses observed when the stimuli were explicitly presented. These results suggest that DecNef is capable of reducing fear-like responses to feared objects without explicit and aversive exposure.
Figure 2. Schematics of the procedure in Koizumi et al. (2016) (Koizumi et al., 2016). In this study, healthy participants went through a fear conditioning session to form two fear memories to red and green colored gratings. They then underwent DecNef training (with grey gratings) to reduce one of these fear memories. In this example, the red memory is targeted during training, and the fear response to the red stimulus is reduced relative to the control (green) stimulus.

One limitation of the study by Koizumi et al. (Koizumi et al., 2016) was that, prior to the threat-conditioning session, the red and green stimuli were explicitly presented to the participants to enable the multivoxel pattern decoding of their neural representations. This explicit decoding session was not unpleasant to the participants as it was prior to the formation of aversive memory for the decoded representations. However, if the procedure were to be applied to the actual treatment of anxiety disorders, aversive exposure to the feared objects for the sake of decoding would undermine the key advantage of this approach: that is, the ability to reduce fear associations via the implicit and non-aversive nature of DecNef training.

To make the entire procedure implicit, it is necessary to also render the decoding process implicit. To achieve this, Taschereau-Dumouchel et al. (Taschereau-Dumouchel et al., 2018) has elaborated the aforementioned DecNef training procedure by integrating another novel technique, hyperalignment (Guntupalli et al., 2016, Haxby et al., 2011). With hyperalignment, this study aimed to alleviate fear-like responses to feared animals among sub-clinically phobic participants. Briefly, hyperalignment enabled the construction of a decoder for the representation of a feared animal in each participant's brain that was based on, or “borrowed from”, the brain activation patterns of surrogate participants.
The first stage of the study involved learning how to align the neural activation patterns of surrogate participants to the activation patterns of a participant with a subclinical level of phobia. This first stage was achieved using activation patterns obtained while the surrogates as well as phobic participants explicitly observed various images of animals that did not include the feared animals. For example, if the phobic participant is fearful of both snakes and spiders, the first stage used the activation patterns in ventral temporal cortex for various animals such as butterflies, dolphins, etc., but did not involve those for snakes or spiders. Data from this stage allowed computation of a mapping that could translate arbitrary activation patterns in ventral temporal cortex from the surrogate group to the participant’s brain.

In the second stage, the activation patterns obtained while the surrogate participants explicitly observed the feared animals (e.g., snakes and spiders) were aligned to the neural space of the phobic-participant using this mapping. Since this study involved only sub-clinically phobic participants, the accuracy of this ‘borrowed’ decoder could be tested with the actual neural activation patterns obtained while they explicitly observed their feared animals. The decoding accuracy was as high as 82.4 %, which was better than the accuracy of the decoder (71.7%) built directly from the neural data of the phobic participant. This relatively high accuracy may be because the benefit of using a larger data set from surrogate participants (N = 29) outperforms the potential information loss that occurs when aligning the activation patterns between the participants. Even with the borrowed decoder, this study showed that fear-like response could be reduced through DecNef training. Importantly, the effect of DecNef was selective to one of the feared animals that went through the DecNef training as a target (e.g., snakes) and did not generalize to the other, control feared animal (e.g., spiders) that never underwent the training. We here highlight that this was a double-blind, placebo controlled, randomized trial, where neither the experimenters nor the participants were aware of which of the animals served as a target of DecNef and which was the control animal. Thus, the effect of DecNef training could not be accounted for by potential placebo effects, experimenter effects, and/or voluntary strategies of participants. Critically, decoding with hyperalignment can render the entire procedure of DecNef training implicit (and thus not aversive), while maintaining the effectiveness of training for reducing fear responses.

Another way to render the decoding process implicit is to use subliminal presentation of feared objects. When only weak sensory inputs are presented, such inputs are processed in the brain to some extent but often fail to evoke conscious awareness of their presence. One of the established procedures to present visual stimuli subliminally is to use continuous flash suppression (CFS; (Tsuchiya and Koch, 2005)). With CFS, a target visual stimulus is presented to a non-dominant eye while dynamic Mondrian patterns are continuously presented to the other, dominant eye. When the stimulus properties such as contrast and duration are carefully manipulated, the target stimulus is often rendered unconscious.

Currently, Chiba et al. (Chiba et al., 2019) are developing a DecNef training protocol that integrates implicit decoding using CFS. This work targets a clinical sample of female PTSD patients who are fearful of aggressive men because of their traumatic experiences of male violence. Instead of explicitly presenting the visual images of angry male faces, this study implicitly presents such images with CFS for decoding purposes. The study then builds a decoder which can estimate the likelihood that the target of an angry man is being represented in a high-level visual area, i.e., superior-temporal sulcus. Although the data collection is still in progress and thus the sample size is currently small, this DecNef training has so far shown success in alleviating PTSD symptoms (Chiba et al., 2019). The
effectiveness of this implicit DecNef procedure will be reported in the future after testing has been completed on the full pre-planned sample size.

**Potential mechanisms behind DecNef training**

The mechanisms by which DecNef training alleviates fear remain to be further examined. One potential mechanism is similar to the mechanism of a typical fear extinction procedure or exposure therapy. That is, repetitive reactivation of the neural representation of a feared object in the absence of threat forms a new association memory between the object and safety (Bouton, 2004). Another potential mechanism of the DecNef training is counter-conditioning induced when an aversive object is repetitively reinforced with positive rewards (Newall et al., 2017). The DecNef training provides monetary reward which is proportional to the success in inducing the target neural representation. Thus, the representation of a feared object is not only repeated in the absence of threat but is also repetitively paired with reward.

To dissociate the two possible mechanisms, Chiba et al. (Chiba et al., 2019) examined whether the trial-by-trial progress of the DecNef induction success (i.e., likelihood of target stimulus) was better fit by a model based on fear extinction learning or by a model based on counter-conditioning. The study revealed that the fear extinction model fit the DecNef performance of the previous two studies (Koizumi et al., 2016, Taschereau-Dumouchel et al., 2018) better than the counter-conditioning model. This result suggests that DecNef training may rely more on a mechanism similar to fear extinction.

This modeling result, however, may need to be elaborated further, and more data on brain activity are needed to better elucidate the mechanisms underlying this form of DecNef training. For example, Koizumi et al. (Koizumi et al., 2016) showed that the striatum, which is involved in reward-related learning (Cox and Witten, 2019), was activated during DecNef training in a manner related to the trial-wise success in inducing the feared object representation in visual cortices. Meanwhile, the involvement of the ventromedial prefrontal cortex, which is widely implicated in fear extinction (Milad et al., 2007, Phelps et al., 2004), was not seen in the data from Koizumi et al. (Koizumi et al., 2016). These neuroimaging results suggest at least some additional role of reward and/or involvement of a non-typical extinction process in DecNef training. Better understanding of the neural mechanisms of DecNef training may help to optimize the training for more effective alleviation of fear and anxiety.

**Alleviating both implicit and explicit symptoms of anxiety disorders with DecNef**

The implicit nature of the DecNef training is beneficial in reducing the aversiveness of exposure when treating anxiety disorders. Given its implicit nature, however, one concern is that the effect of the training might be confined to more implicit symptoms of the disorders (Schiller, 2016). That is, participants might still feel a subjective experience of fear that is as strong as in the pre-training phase, even when they show reduced autonomic responses to the feared objects. Such dissociations between implicit and explicit measurements of fear have been demonstrated in the past (LeDoux, 2019).

Indeed, the earlier study with DecNef training [4] only demonstrated the effects on objective measurements of fear-like responses, such as skin conductance response and amygdala activation. Results on the additional effect of DecNef training on the subjective fear are still scarce and mixed, and thus such an effect needs to be further investigated. For example, the aforementioned study in progress (Chiba et al., 2019) has so far been successful in reducing PTSD symptoms, which are assessed through explicit, subjective
reports by the patients. Meanwhile, more explicit, subjective experiences of fear with animals were shown to remain unchanged after DecNef training with the hyperalignment technique (Taschereau-Dumouchel et al., 2018, Taschereau-Dumouchel, 2020).

Further investigations may more directly examine the effects of DecNef training on both implicit and explicit components of fear- and anxiety-related symptoms, in order to improve the training to better treat them both. For example, the DecNef training may be improved by directly targeting the neural activation patterns related to each component of symptoms. One recent study (Taschereau-Dumouchel et al., 2019) has demonstrated that while the degree of fear-like response measured via skin conductance could be decoded from the activation patterns in the amygdala, the degree of explicit reporting of experiencing fear could be decoded from the prefrontal areas. Based on this dissociation, DecNef training may target the neural activation patterns in the amygdala and/or its associated areas such as sensory cortices to reduce implicit fear-like responses, while simultaneously targeting the patterns in the prefrontal areas to additionally alleviate the subjective experience of fear.

**Summary and future directions**

Implicit DecNef training may prove beneficial for the treatment of other disorders involving anxiety besides phobias and PTSD. One example may be obsessive-compulsive disorder (OCD). OCD is a disorder which involves obsessions with certain unwanted and intrusive thoughts inducing anxiety (e.g., I may have left the house unlocked) and/or compulsions to repetitively perform certain behaviors to neutralize such obsessions (e.g., repeated checking) (Stein et al., 2019). One of the common treatments for OCD is Exposure Response Prevention (ERP, (Wheaton et al., 2016)), in which patients need to face their fear through exposure to images and situations related to their symptoms (e.g., viewing a door) while inhibiting their urge to respond (e.g. refraining from checking the lock). As one potential new application, DecNef training may elaborate on this ERP procedure. Specifically, DecNef training may help patients to implicitly induce the neural activation patterns representing the symptom-provoking stimuli. Because of the implicit nature of the activated representations, compulsive responses may be less likely to be evoked and cannot be executed within MRI scanners during the training. Thus, inhibition of such responses during DecNef training may be less effortful and stressful than during standard ERP.

The procedure of DecNef training has been progressively modified across studies, and future studies may further customize the training in a flexible manner to meet the needs of targeted disorders. Our hope is not to replace the existing clinical practice with DecNef training, but rather to provide some additional treatment options to patients so that they can more flexibly choose how they would like to overcome their disorders.
References


