Unconscious manipulation of conceptual representations with decoded neurofeedback impacts search behaviour

Pedro Margolles[1,2], Patxi Elosegi[1,2], Ning Mei[1], and David Soto[1,3]

[1] *Basque Center on Cognition, Brain and Language, San Sebastian, Spain*

[2] *University of the Basque Country- UPV/EHU, Spain*

[3] *Ikerbasque, Basque Foundation for Science, Bilbao, Spain*

July 4, 2023

Author Note

Correspondence should be addressed to: pedromargolles@gmail.com, d.soto@bcbl.eu, Basque Center on Cognition, Brain and Language, Paseo Mikeletegi 69, 2nd Floor 20009 San Sebastian.

Abstract

The necessity of conscious awareness in human learning has been a long-standing topic in psychology and neuroscience. Previous research on non-conscious associative learning is limited by the low signal-to-noise ratio of the subliminal stimulus, and the evidence remains controversial, including failures to replicate. Here, functional MRI decoded neurofeedback (DecNef) allowed participants to generate brain activity patterns consistent with a perceptual state without awareness of the represented content. We aimed to instill perceptual meaning into a meaningless hiragana by conditioning it through reinforcing neural patterns associated with perceptual content (i.e. living), thereby bypassing the conscious mapping between symbol and the perceptual referent. Despite being unaware of the goal of DecNef training, participants learned to induce the target representation in fusiform cortex. Then, the behavioural significance of DecNef training was evaluated in a visual search task. DecNef and control participants searched for living or non-living targets that were pre-cued by the targeted hiragana during DecNef or by a novel hiragana. The targeted hiragana did not prime search for the associated target but, strikingly, participants were impaired at searching for the perceptual content targeted by DecNef. Hence, conscious awareness may function to support higher-order associative learning, while lower-level forms of re-learning, modification, or plasticity in established neural representations can occur unconsciously, with behavioural consequences outside the original training context. The work also offers an account of DecNef effects in terms of neural representational drift.

Significance Statement

The present study examines the role of conscious awareness in human learning by using real-time functional MRI and decoded neurofeedback (DecNef) techniques to allow participants self-regulate brain activity patterns according to a representation of the world, without awareness of the represented content. We demonstrate that established conceptual representations in the human brain can be modified without awareness and impact visual search behaviour for the targeted content through a mechanism of neural representational drift. However, the work implicates conscious awareness in higher-order forms of associative learning. The study further provides novel methodological insights for the advancement of protocols for DecNef and pave the way for exploring the potential of DecNef in personalized interventions, enabling targeted guidance to rectify maladaptive conceptual representations.

Keywords: decoded neurofeedback, associative learning, representational drift, consciousness

Introduction

Learning new concepts (e.g.. word meaning) relies on associative learning mechanisms to map a symbol with an episodic referent [1, 2]. A longstanding question in psychology and neuroscience relates to the role of conscious awareness in higher-level associative learning. It remains uncertain whether awareness is a prerequisite for learning new concepts and, if so, what specific purpose it serves.

Previous studies have provided evidence of unconscious associative learning between subliminal visual items and conscious rewards [3, 4] (but see [5]) and fear conditioning responses to unseen stimuli [6]. Moreover, studies utilizing evaluative learning paradigms have demonstrated the possibility of conditioning affective value into neutral visual stimuli, like nonsense syllables, and suggested that such semantic associations can occur implicitly, even without participants' awareness of the visual items [7–9]. However, it is important to note that the aforementioned evidence remains controversial and has faced substantial criticism [10–13]. It has been argued that conditioning effects in evaluative learning paradigms rely on demand characteristics of the experiments and explicit awareness rather than non-conscious knowledge [12, 14, 15].

Research on unconscious associative learning is severely constrained by the low signal to noise ratio of the unconscious stimulus, which is typically very brief and masked. If the unconscious item is poorly represented in the brain, associative learning is likely to falter. Addressing this challenge requires an increase in signal strength in the unconscious condition [16]. Here, we tackle these issues by means of functional MRI decoded neurofeedback (DecNef), which allows participants to generate brain activity patterns in keeping with a target state without awareness of the content represented in those patterns [17]. Prior DecNef research has primarily focused on modulating affective responses to faces [18], fear [19, 20] or subjective confidence [21]. Here, we aim to use DecNef to manipulate and generate novel conceptual knowledge. Thereby, the primary goal of the experiment was to incept perceptual meaning associated with real-world referents into the representation of a meaningless hiragana symbol, by conditioning the hiragana through reinforcement of neural patterns associated with perceptual content.

We developed a flexible framework for fMRI-based DecNef that is openly available to the scientific community, and that includes several improvements in the standard DecNef pipeline. Figure 1 illustrates the experimental procedure. First, a pattern classifier was trained to discriminate living from non-living images (i.e. dog vs scissors; Figure 1A; see also Methods). Then, several sessions of DecNef training were conducted across several days in which participants were asked to self-regulate their brain activity while exposed to an hiragana symbol with the goal of maximizing the size of a feedback disk and hence monetary reward (Figure 1B). Unbeknownst to participants, the feedback depended on the similarity between the induced patterns of brain activity in the fusiform cortex and the representation of the living concept (i.e. Dog) according to the classifier. We expected that this approach would lead to the formation of conceptual associative knowledge thereby incorporating the meaning of dog into the hiragana. The behavioural significance of the DecNef intervention was evaluated using a search task involving dogs and scissors, with the search target being pre-cued by the hiragana targeted by DecNef or by a neutral hiragana (Figure 1C; see also Methods). We hypothesized that if the hiragana was effectively associated with the perceptual meaning of 'Dogs' during DecNef, it would cue participants' attention to the matching item in the visual search task [22,23], leading to faster search performance and facilitated neural responses relative to the control hiragana. We compared the search behaviour between the DecNef group and a control group to further examine the impact of DecNef training on the representation of the target category itself. This contrast further assesses the possibility of unconscious learning operating not through associative knowledge formation, but rather influence search behaviour through the manipulation of established representations.

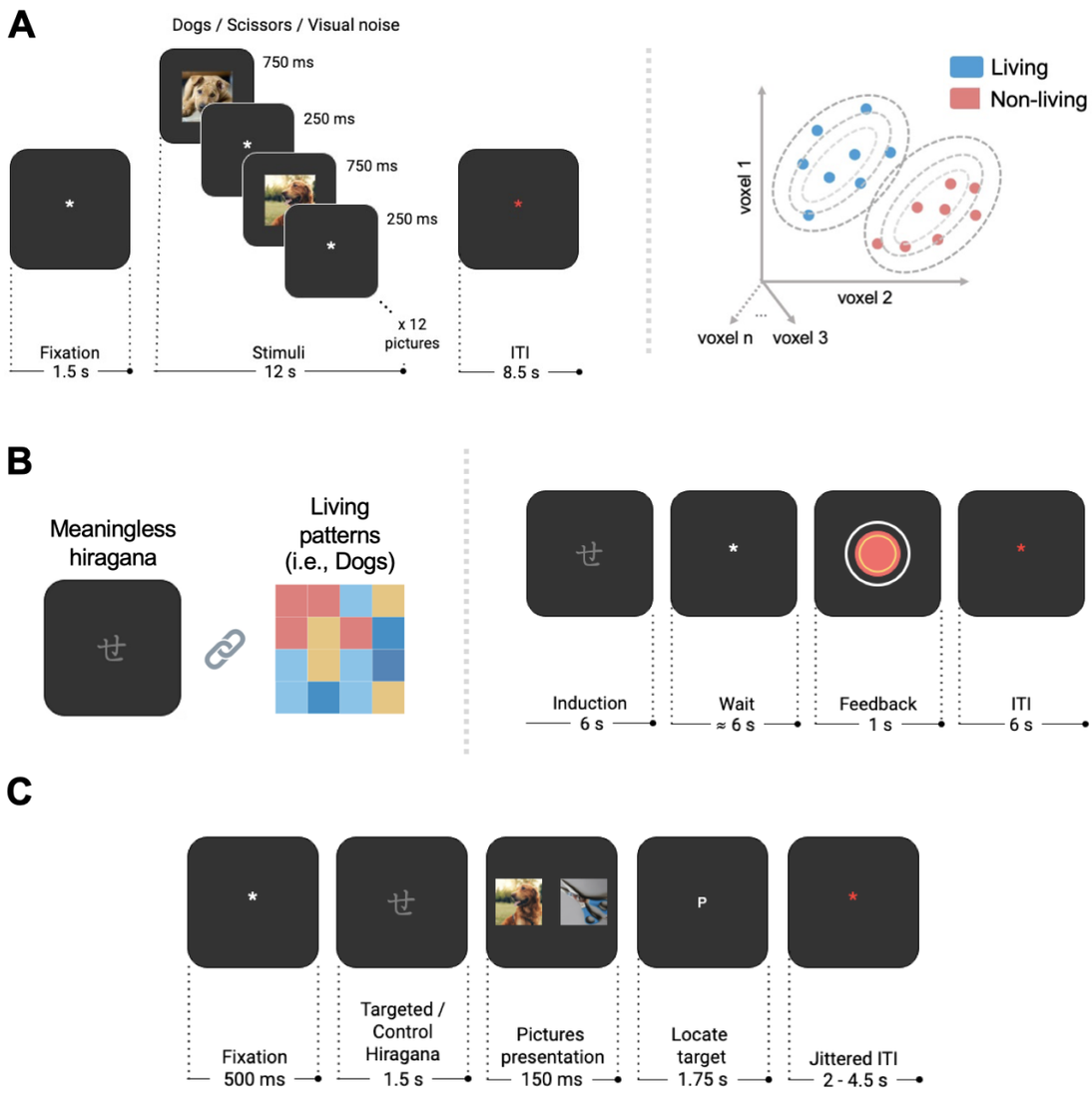


Figure 1: *Overview of the experimental design. (A) Participants were presented with images of 'Dogs' and 'Scissors'. A classifier was trained with examples in the bilateral fusiform gyrus to differentiate between the classes. (B) Across four DecNef sessions, participants self-regulated their brain activity while exposed to the hiragana. The size of the feedback disk reflected how closely the induced brain activity patterns aligned with target class according to the classifier. (C) Upon completion of the four DecNef sessions, a visual search task was administered to evaluate the behavioural significance of DecNef training. Participants searched for living or non-living targets that were pre-cued by the targeted hiragana during DecNef or by a novel hiragana.*

Results

Previous studies on DecNef (Decoded Neurofeedback) [17] employed classification models that often generated excessively binary predictions when applied to out-of-sample data. Consequently, these models hindered the utilization of nuanced neurofeedback signals to effectively facilitate participants' acquisition of skills in regulating their brain activity patterns. To resolve this issue, we implemented three strategies: (i) we developed a custom regression model using Binary Cross-Entropy Loss to provide a more fine-grained decoding of the target class while accounting for noisy representations using soft-labeling (see Methods) (ii) we standardized the fMRI scans during DecNef based on the data used for classifier training (iii) we aimed to increase the feedback variability by averaging decoding probabilities from volumes within the HRF peak of each trial.

First, we assessed the participants' ability to induce brain activity patterns corresponding to a living world referent in the bilateral fusiform, which unbeknownst to them referred to the 'Dogs' images. Notably, the out-of-distribution predictions during DecNef given by our classification pipeline were far less bimodal than previously reported in DecNef studies using logistic regression models [17] (see 'Model comparison' section in the Supplementary Results and Figures S2-S3), thereby providing participants with more informative

feedback.

A linear mixed model (LMM) analysis demonstrated significant learning effects across DecNef sessions regarding the induction of the target pattern in the fusiform cortex (Figure 2) (β = 7.833e^{-05}, SE = 3.281e^{-05}, 95% CI: 1.338e^{-05}, 1.428e^{-05}, p = .019), despite inter-individual differences in learning rate (Figure S4).

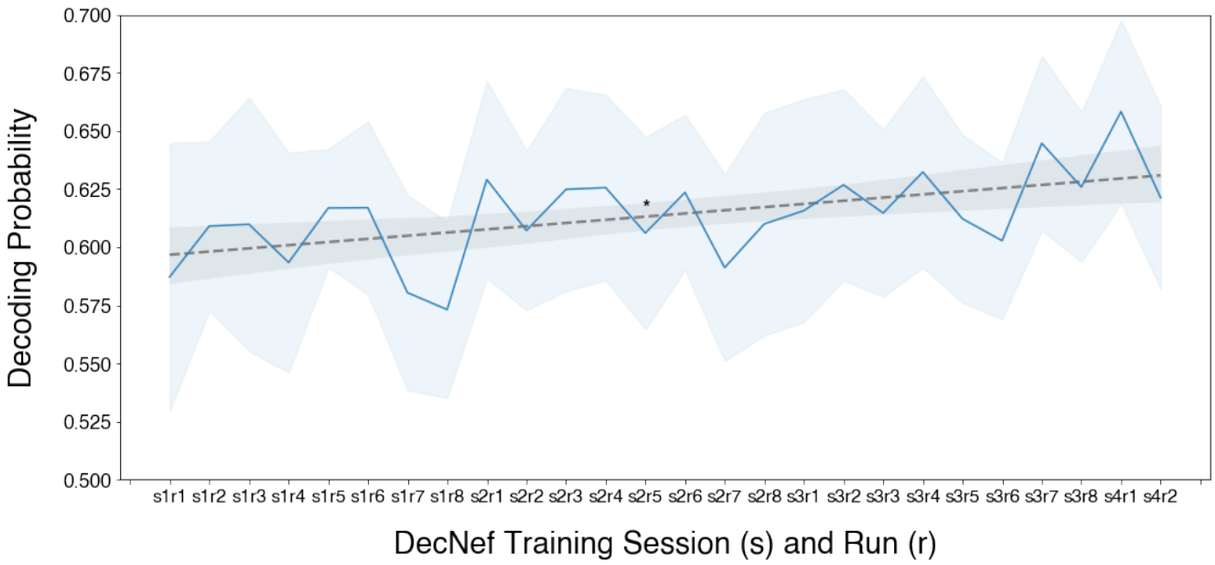


Figure 2: *Learning effects throughout DecNef training. The figure represents the average decoding probability of 'Dogs' category in the bilateral fusiform at run level across participants. * p = 0.019. Error bars represent 95% CIs according to the Cousineau–Morey method [24].*

We also explored the possibility that other brain regions beyond the fusiform were implicated in the induction of the relevant target category during DecNef. We performed an off-line searchlight decoding analysis considering all the data gathered across the DecNef sessions. The results implicated other brain regions in the induction of the target pattern, including early visual cortex (lingual and calcarine sulcus), the precuneus and posterior cingulate gyrus (see Figure 3).

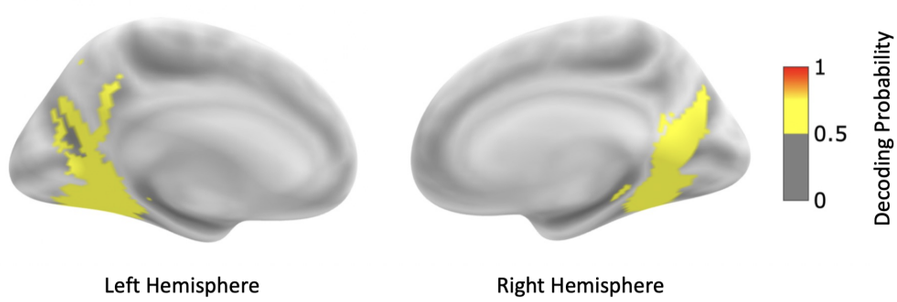


Figure 3: *Searchlight maps illustrate the brain areas in which the probability of inducing the target category across DecNef sessions exceeded the nominal 0.5 chance-level.*

Visual search task: behavioural results

To assess the impact of the DecNef training on visual search task, we conducted a Linear Mixed Effects Model on the search reaction times (RTs) of the correct responses (see Methods). On average, 2.53% of trials were removed across participants due to response errors (SD = 1.9%). We also excluded outlier RTs that fell outside of 3 standard deviations from the each participant average RT (mean: 1.17% of trials; SD = 0.71%). The results showed that search RTs were significantly longer for 'Dogs' items compared to 'Scissors' (β = 22.053, p $<$ .001; SE = 5.650; 95% CI Lower = 10.981, 95% CI Upper = 33.126). However, the main effect of the hiragana (targeted vs neutral) on search RTs was not significant (β = 4.795, p = .396; SE =

5.647; 95% CI Lower = -6.272, 95% CI Upper = 15.863). The interaction between the factors hiragana and search target was also not significant (β = -10.836, p = .175, 95% CI Lower = -26.501; 95% CI Upper = 4.829). Figure 4A depicts these results.

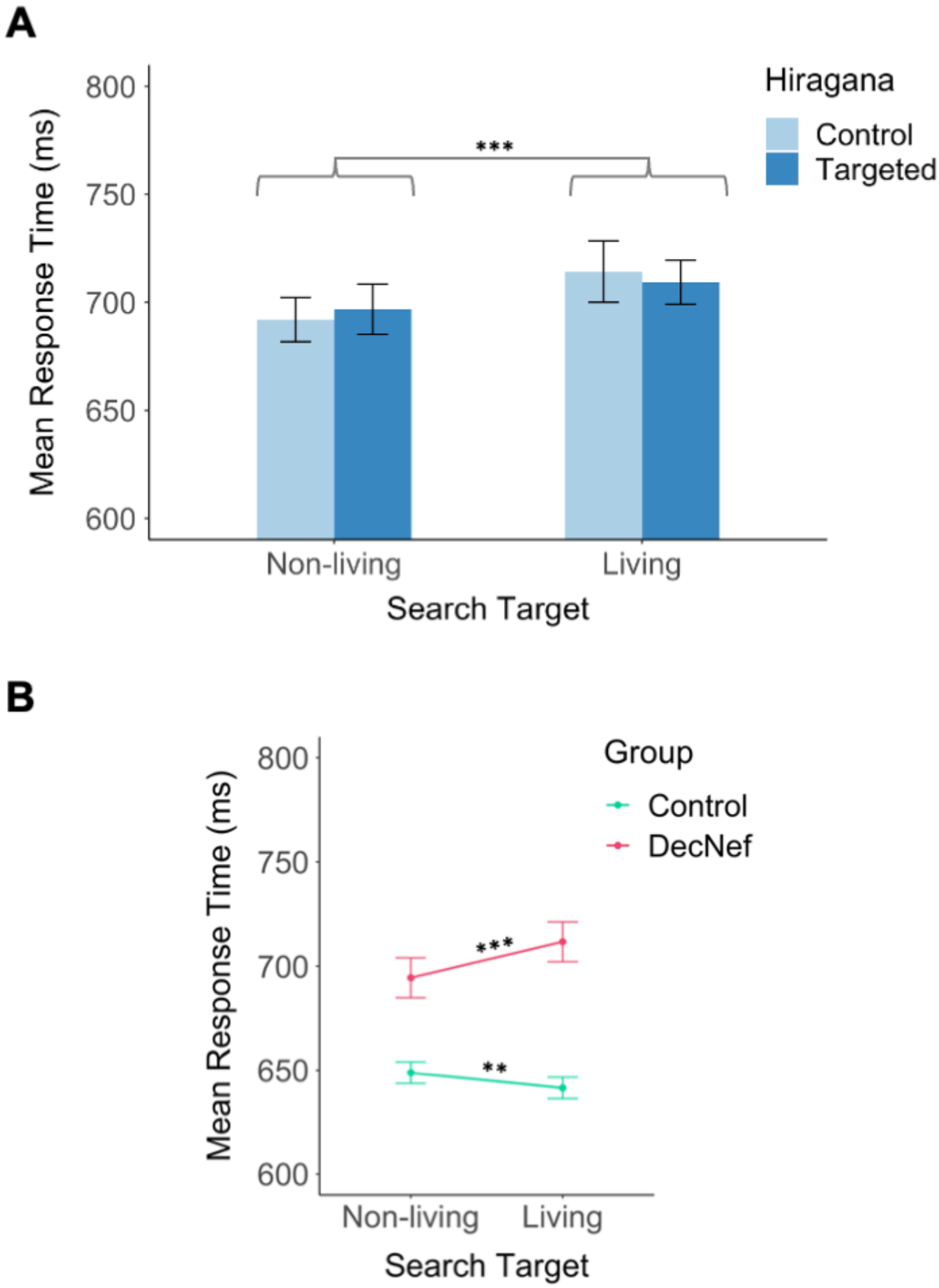


Figure 4: *Behavioural results. (A) DecNef group: Visual search reaction times (RTs) as a function of the hiragana and search target factors. (B) Search RTs as a function of the search target across DecNef and control groups. Error bars represent 95% CIs according to the Cousineau–Morey method. ** $p < 0.005$, *** $p < 0.001$.*

The slower search RTs for 'Dogs' items in DecNef participants is surprising, since previous research has demonstrated the faster recognition of living categories compared to non-living counterparts [25–28]. In line with this, a control group (N=36) that did not undergo DecNef, but performed the same task under

matching conditions, showed faster search for living compared to non-living targets. Data analyses for the control group followed the same pipeline. RTs of correct search responses were analysed (Mean error = 1.82%, SD = 1.61%). Similar to the DecNef group, outlier RTs beyond 3 standard deviations were removed (Mean = 1.23%, SD = .53%). In order to compare the search RTs across DecNef and control groups, the RT data was collapsed across the levels of the hiragana induction factor, since this was not a factor in the control group. The results of the LMM analysis showed no significant differences in search RTs across the DecNef and control groups (β = 40.029, p = .136; SE = 30.423; 95% CI Lower = -13.534, 95% CI Upper = 105.594). Crucially there was a robust interaction between group and the search target factor (β = 24.058, p $<$ .001; SE = 4.563; 95% CI Lower = 15.115, 95% CI Upper = 33.001). Simple effect analyses indicating that participants in the control group showed significantly faster RTs to pictures of 'Dogs' compared to 'Scissors' (β = 7.42, SE = 2.60, p = .004), while the DecNef group showed the opposite pattern (β = -16.64, SE = 3.75, p $<$ .001) (see Figure 4B). Additionally, we observed that the DecNef participants displayed slower search RTs for 'Dogs' compared to the control group (Contrast: Control - DecNef, β = -70.1, SE = 30.4, p $<$ .021) but no differences in search RTs for 'Scissors' (Contrast: Control - DecNef, β = -46.0, SE = 30.4, p = .130). No differences between DecNef and control groups were observed in search accuracy (see Supplementary Results).

Visual search task: Univariate fMRI results

Potential changes in brain activity driven by DecNef training were assessed by means of univariate analyses of the fMRI data collected while participants performed the visual search task. It is well established that perceptual expectations or implicit memory templates have a strong effect on neural responses during perceptual decisions and visual search tasks [22,23,29]. These effects have been linked to the phenomenon of neural repetition suppression, whereby neural responses decrease following the presentation of an expected item or following the presentation of an item that matches an implicit memory trace. Accordingly, based on prior studies assessing the interplay between memory and visual selection [22,30], here we hypothesized that if the hiragana was incepted with the perceptual representations of the 'Dogs' class, then a neural repetition suppression response would be triggered by the subsequent presentation of the 'Dogs' image in the search task, relative to when a neutral (non-incepted) hiragana preceded the search display. However, there were no significant clusters of activity consistent with our hypothesis, neither at the whole-brain level nor in a fusiform ROI in standard space (see Methods). Further, we computed the average activity in the bilateral fusiform ROI for each participant in native space and performed a repeated measures ANOVA with the following factors: hiragana (targeted vs. control) and search target ('Dogs' vs. 'Scissors'). There were no significant main effects of the factors associated with the hiragana (F(1, 17) = .181, p = .676, η_p^2 = .011) or the search target (F(1, 17) = .019, p = .893, η_p^2 = .001) and no significant interaction (F(1, 17) = .312, p = .584, η_p^2 = .018).

PCA analysis

So far, the results indicate that DecNef training did not result in associative knowledge between the hiragana and the perceptual content represented in the decoded multivoxel activity pattern in the fusiform. However, the behavioural results indicate that the representation of the target concept changed following DecNef training. We then further examined representational changes induced by DecNef training by means of a principal component analysis (PCA) of the fusiform activity patterns across DecNef runs [31]. This approach has been previously used to characterize the brain representation of both sensory [32, 33], and

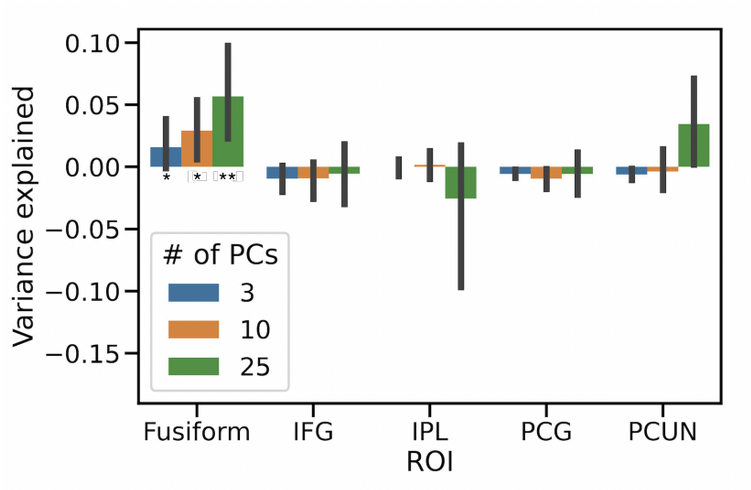


Figure 5: *Variance explained by the SVR analysis as a function of ROI and number of PCs. Error bars represent 95% CIs according to the Cousineau–Morey method. * $p < 0.05$, ** $p < 0.005$*

motor information [34, 35] in cortex.

The analysis was performed in Scikit-learn [36] on an individual basis. Specifically, principal components (PCs) were identified from the living examples used to train the decoder. Prior to fitting the PCA, trial examples were normalized by subtracting the mean and dividing by the standard deviation. Subsequently, the DecNef data were normalized and projected onto the space identified by the PCs. Then, a linear support vector regression (SVR) model was used to test for linear trends in the representation defined by the PCs across the runs of DecNef training for each participant. The linear SVR model used the PCs of the DecNef data to predict the indices of the training runs. The model was cross-validated using a 100-fold cross-validation procedure before deriving the average variance explained across folds. We run the analysis with 25 PCs based on prior analysis of DecNef induction dataset [37] and also with 10 and 3 PCs in order to estimate the robustness of the results.

The results of the PCA/SVR analysis in the fusiform ROI targeted by DecNef showed significant changes in the representation of the PCs across the runs of DecNef. This result was consistent across different analyses performed with different numbers of PCs (25 PCs: $t(17) = 2.953$, $p = 0.004$; 10 PCs: $t(17) = 2.34$, $p = 0.016$; 3 PCs: $t(17) = 1.758$, $p = 0.048$).

The same PCA/SVR analysis was performed in four control ROIs. The inferior frontal and inferior parietal ROIs were selected as control regions because their activity patterns did not significantly predict the target class in the searchlight analysis Figure 3. We also included a second pair of control regions from the searchlight analysis, namely, posterior cingulate and precuneus, that showed above chance probability of decoding the target class.

We directly compared the variance explained in the fusiform and the control ROIs by means of 5 (ROI) x 3 (number of PCs) repeated measures ANOVA using Greenhouse-Geisser for the sphericity correction correction. There was a main effect of the ROI factor $F(4, 68) = 4.235$, $p < 0.025$, $\eta_p^2 = 0.199$. There was no main effect of number of PCs ($F(2, 34) = 2.522$, $p = 0.126$, $\eta_p^2 = 0.129$) and no interaction between the number of PCs and ROIs, $F(8, 136) = 2.464$, $p = 0.105$, $\eta_p^2 = 0.127$. Figure 5 depicts the pattern of results. Additional post-hoc tests, using Holm correction for multiple comparisons, showed that the variance explained in the fusiform was significantly higher compared to the other ROIs (i.e. inferior frontal: $t = 3.421$, $p = 0.011$; inferior parietal: $t = 3.381$, $p = 0.011$; posterior cingulate: $t = 3.166$, $p = 0.019$) except the precuneus ($t = 2.054$, $p = 0.307$).

Participants' subjective reports

Participants were asked to report the strategies they used to modulate the size of the feedback disk during DecNef training. The participants used several strategies, such as mental humming, mathematical calculations, visualization of different scenarios, evoking emotions, etc. A comprehensive list of these strategies can be found in Table S1 of the Supplementary Materials. Many participants stated that they had to change and adopt a diverse set of strategies to sustain performance over time. Participants reports indicate that they were unaware of the specific research goals. Out of 18 participants, 12 correctly guessed the target category as 'Dogs' but a Chi-squared test showed the observed frequencies were consistent with a uniform distribution and therefore unreliable identification the target class ($\chi 2(1) = 2$, $p = .157$). Confidence reports revealed low levels of confidence in detecting the category overall, even among those who correctly identified it. A full report of the participants' verbal reports can be found in Table S2 of the Supplementary Materials.

Discussion

The present study aimed to manipulate conceptual representations without participant's awareness by means of DecNef training, and thereby instill perceptual meaning into a neutral hiragana. First, we observed significant learning effects across DecNef training. Participants progressively improved in their ability to induce brain activity patterns that were similar to the target concept. This observation is noteworthy because previous DecNef studies (see [17] for a review) did not reliably observe significant learning effects, despite observing DecNef effects on behavioural performance (but see [38]). Here we improved the classification model to provide a better calibrated and more fine-grained feedback to participants, compared to the logistic regression models used in previous DecNef studies that tend to give highly binomial classification probabilities as feedback [39]. The improvements in the classification model may promote participant's ability to identify more subtle variations in neural changes through feedback, thereby improving their ability for self-regulation during DecNef training. The developments presented here represent a step forward towards enhancing the precise targeting of neural representations by DecNef. However, there were inter-individual differences in learning during DecNef. This variability is typically observed in neurofeedback studies due to individual differences in attention or motivation levels [40] and metacognitive abilities [41–43].Future DecNef studies would benefit from developing personalised training protocols to enhance learning within each individual.

Despite the reliable learning effect observed during the induction blocks of DecNef which afforded strong non-conscious neural representations of the target perceptual content, the behavioural results do not support the view that the hiragana was associated with the target concept. Search performance for the target category ('Dogs') was similar when search was primed by the hiragana targeted by DecNef of by a control hiragana. Further, several univariate fMRI analyses in the visual search task did not show neural modulation driven by the matching between the targeted hiragana and the target category. While some prior studies indicated that associative learning can operate on visually masked, subliminal representations [3,4,6], mounting evidence has recently indicated otherwise [5, 10, 11] (see also [12, 13] in the case of evaluative conditioning). Previous studies showing evidence for unconscious associative learning were heavily criticised on methodological grounds, mainly related to demand characteristics of the experiment and limitations in the awareness checks [12, 13]. Our findings suggest that even though strong neural representations of the target perceptual content are induced, the formation of associative knowledge between the content represented in brain activity patterns and a neutral stimulus is likely dependent on conscious access to the

represented content within the neural patterns. However, this does not necessarily exclude that others forms of associative learning, namely, involving emotional content and fear conditioning [44] may occur outside participants' awareness (but see [45])

A previous previous study [46] reported evidence of an association between orientation and color in the early visual cortex driven by DecNef. However, the present study tested higher-order associative learning between an abstract hiragana symbol and a perceptual class, which may be more reliant on conscious access to the perceptual referent compared to the lower-order associations in Amano et al. [46]. More recently, Collin and colleagues [47] showed that contextual representations of faces or houses can be induced through DecNef in visual cortex while memorizing a different set of objects. The implicit neural context induced by DecNef (i.e., face or house) led to associative memories between the study items and the specific encoding contexts. However, participants were explicitly instructed to think about either faces or houses during DecNef training in the study by Collin and colleagues [47], thereby, likely implicating awareness of the association between the feedback and the mental context (i.e. whether the feedback was related to faces or houses).

Notably, there was a robust effect of DecNef training on the representation of the target category. Relative to a control group, selection of the target category during search was impaired. While the control group showed a visual search advantage for the living targets over the non-living counterparts, in line with previous observations in the literature (e.g. [25–28]), the DecNef group displayed the opposite pattern of results, namely, slower search for the living targets compared to the non-living counterparts. This dissociation indicates that DecNef training can lead to a reconfiguration of the template associated with the target category with consequences for search behaviour. Importantly, previous studies demonstrated that visual imagery, working memory and longer-term episodic memory have positive influences in the guidance of attention, enhancing selection and search performance for matching targets [23, 48–50]. Therefore, the effect of DecNef is qualitatively distinct (indeed opposite) to what would be expected if participants were consciously rehearsing the target representations during DecNef.

A plausible account is that DecNef triggered representational drift of the target concept [51], whereby the alignment of the neural activity patterns associated with a representation of the world changes over time within a high-dimensional space. DecNef protocols require participants to explore a search space to find the target representation by means of diverse cognitive strategies [17] that are conceptually unrelated to the target and that are further accompanied by other sources of noise (e.g. here, the presence of the hiragana). Hence, the brain is confronted with a many-to-one mapping problem that occurs in machine learning, whereby a unique target can be potentially mapped to several patterns in multidimensional space. The repeated induction of noisy neural patterns that partially overlap with the representation of the target can therefore lead to the corruption of the target representation through Hebbian learning and operant conditioning [52]. The PCA analyses demonstrated that the patterns of fusiform activity underwent changes in representation throughout the runs of DecNef training. These changes were observed in comparison to several control regions, which supports the view that the effects of DecNef training are more substantial and localized within the targeted region, in line with a targeted neural plasticity model of DecNef training effects [37].

Representational drift has been observed in the hippocampus [53] and also in cortical regions [54], including sensory cortex [32, 55] when observers are exposed to similar stimuli across sessions and days. Representational drift is higher for more complex representations associated with naturalistic stimuli relative to simple items (i.e. oriented gratings) [56]. Importantly, representational drift can occur despite sustained decoding accuracy of the target content from neural responses by a linear classifier [55], as we observed in

the present study.

Previous theoretical work proposed that representational drift may afford computational benefits, namely, for continual learning [57]. However, the causal effects of representational drift on behaviour remain to be elucidated [58]. The present results indicate that representational drift may have unwanted behavioural consequences as well. Specifically, representational drift may lead to a mismatch between internal representations held in memory and the perceptual input from the external environment, thereby influencing visual selection and search behaviour.

In summary, this study elaborates on the impact of conscious awareness on different types of learning. It indicates that higher-order associative learning is more likely to depend on conscious awareness, while lower-level forms of re-learning, modification, or plasticity in established representations can occur without conscious awareness via a mechanism of neural representational drift. Further, the work provides novel insights for the advancement of protocols for DecNef, specifically aimed at manipulating conceptual representations. These findings open up new avenues for investigating the potential of DecNef in personalized interventions, allowing for the regulation of guidance based on maladaptive representations.

Materials and Methods

Participants

The DecNef training involved a group of 18 young adults (23-36 years old, mean±SD age: 28.19±3.73 years, 6 males) with normal or corrected-to-normal vision, who were right-handed and unfamiliar with the Japanese language. A sample size of 18 is well above average in DecNef studies. The experimental procedures were conducted in accordance with the Declaration of Helsinki and approved by the BCBL Research Ethics Board. All participants provided written informed consent and at the end of the experiment, received a fixed monetary compensation of 105€ for their participation in the study. Additionally, we included a control group of 36 young participants to double the DecNef group sample size and improve the statistical power of our analysis for detecting any differences between groups (20-29 years old, mean±SD age: 23.16±2.62 years, 11 males). One of the participants had to be replaced due to a high number of errors (i.e., $> 40\%$).

MRI settings

Whole-brain data was acquired on a 3T Siemens Magnetom Prisma-fit whole-body MRI scanner with a 64-channel head-coil at the BCBL facilities. A gradient-echo echo-planar imaging sequence was used with the following parameters: time-to-repetition (TR) = 2000 ms, time-to-echo (TE) = 28 ms, flip angle (FA) = 74°, field of view (FoV) = 192 mm, 34 axial slices with no inter-slice gap and a 3*3*3.5 mm^3 voxel resolution. Five volumes were acquired as MRI heat-up at the beginning of each fMRI run to ensure steady-state magnetization. These volumes were then discarded from analyses. Additionally, one high-resolution T1-weighted structural MPRAGE scan was acquired: TR = 2530 ms, TE = 2.36 ms, FA = 7°, FoV = 256 mm, 1 mm isotropic voxel resolution, 176 slices. Visual stimuli were presented on an out-of-bore screen, which participants viewed through a mirror mounted on the head coil. To minimize head movements, participants were instructed to keep their head as still as possible, and the head-coil of the MRI scanner was cushioned with foam. Participants used padded headphones to minimize background scanner noise, to facilitate the presentation of auditory stimuli, and to enable communication with experimenters via a compatible microphone. At the end of each run, participants were invited to take a short break of 1-2 minutes.

Decoded Neurofeedback Framework

Several software packages are available for the development of neurofeedback studies: FRIEND toolbox [59], Brain-Voyant [60], Turbo-BrainVoyager [61], BrainIAK [62] and OpenNFT [63]. However, the processing pipelines available to date are often closed and rigid, require commercial licensing or have incomplete documentation. Here we developed an open-source, Python-based flexible pipeline to develop DecNef experiments.

Model construction session

The model construction session involved four perceptual runs and two mental imagery runs. First, participants completed four perceptual runs in which they were presented with pictures of 'Dogs' and 'Scissors' from the THINGS database [64]. Each perceptual run was composed of seven 'Dogs' and seven 'Scissors' trials, lasting approximately 8 minutes. A trial began with a 1500 ms white fixation at the center of a black background, followed by a sequential presentation of 12 example pictures from a specific visual category ('Dogs' or 'Scissors') for 750 ms each. The picture presentations were interleaved with 250 ms visual fixation periods. This block design was used to improve the signal-to-noise ratio for specific category representations. The same images were used for each category across trials but these were randomly presented on each mini-block of 12 images. Finally, to model the brain activity patterns associated with each category and maximize the separation of these patterns, we included a 8500 ms inter-trial interval (ITI) with a red fixation at the center of the screen. Throughout the perceptual runs, participants were instructed to focus their attention on the presented pictures and actively engage in processing their meaning. Moreover, each perceptual run included four visual noise mini-blocks. The stimuli consisted of gaussian visual noise patterns generated by randomly selecting colored pixels of different intensities from the RGB (Red, Green, Blue) channels. At each trial, 12 patterns were randomly selected from a pool of 60 gaussian visual noise examples and presented in the same way as the 'Dogs' and 'Scissors' pictures. The data collected from the visual noise trials would also be used to train the decoding model and generate a more fine grained representational gradient from 'Scissors' to 'Dogs' category. In addition, we aimed to take into account the fact that during DecNef any induced representation similar to the target representations would be internally generated or mentally evoked. Hence, following the perceptual runs, participants completed two mental imagery runs in which they were asked to vividly imagine one example picture from each of the visual categories. These samples had been freely selected based on their ease of imagining at the beginning of the perceptual runs. Each of the mental imagery runs consisted of nine imagery trials for each category, with a total trial length of 22 seconds. The trials for the mental imagery runs started with a white fixation on a black background lasting for 1500 ms, followed by a 12-seconds period in which participants were instructed to imagine the selected category picture while maintaining their fixation. To indicate which category they should imagine, auditory cues in Spanish of 'Perro' (i.e., 'Dog'; 375 ms) or 'Tijeras' (i.e., 'Scissors'; 676 ms) were presented three times during the 12-second period. Auditory cues were used to avoid any visual stimulation that could interfere with visual imagery. The volume of the auditory cues was adjusted individually for each participant. After the onset of the auditory cue, participants were required to imagine for approximately 4.5 seconds until the onset of the next auditory cue. They were instructed to mentally project a vivid and static image in the center of the screen. When the second and third auditory cues were presented, participants were instructed to re-activate the target image for that trial. This was done to improve the signal-to-noise ratio. An inter-trial interval (ITI) including a red fixation point was displayed at the center of the screen for 8500 ms. We used OpenSesame software [65, 66] to control stimulus presentation.

fMRI volumes preprocessing

The fMRI volumes for model construction were preprocessed using our pyDecNef library workflow (https://pedromargolles.github.io/pyDecNef/). This framework was designed to maximize the similarity of the data preprocessing for model construction and the real-time processing during DecNef. The raw fMRI images were preprocessed using a customized pipeline in AFNI v21.0.05 [67], which was created through the Nipype v.1.4.1 [68] interface module for Python 3.6. AFNI was preferred over other neuroimaging packages such as FSL [69] due to AFNI's faster processing speed in handling real-time fMRI data.

The first experimental volume from the model construction session obtained after the heat-up volumes was used as a functional reference for each participant. That reference DICOM file was converted to NifTi format using Dcm2Niix [70] and de-obliqued to match cardinal coordinate orientation with AFNI 3dWarp command. The volumes were then co-registered to the reference volume using AFNI 3dvolreg command using heptic (7th order) Lagrange polynomial interpolation [71] The OpenSesame task logs were used to label all volumes, indicating their stimulus category, trials onset times, and corresponding fMRI run. For each run, we linearly detrended the fMRI time series to remove drift before training the decoder model. We selected scans between 5 to 18 seconds after the first image presentation, comprising approximately 7 volumes per trial, in both perception and mental imagery runs. The selected perceptual and mental imagery samples were stacked independently. Next, the stacked volumes in the perceptual domain were z-scored by subtracting the samples' mean value at each voxel and dividing by the standard deviation. The standardization of the scans in the imagery runs was performed by using the voxels mean and standard deviation from the perceptual samples.

Regions of Interest (ROIs) selection

To enhance the reliability of the feedback provided to participants during DecNef training, we selected a region of interest (ROI) by delineating the areas of convergence between perceptual representations evoked by the pictures and internally generated representations evoked during mental imagery in the fusiform cortex [72–75]. The rationale behind this approach is that the brain activity patterns used for classifier training are mainly externally triggered through the presentation and observation of images, while the relevant brain activity patterns during DecNef training arise from internal mental simulation. By targeting voxels at the intersection of perception and imagery, we aimed to improve the cross-domain generalization of our machine learning model and promote effective DecNef training.

For each participant, a decoding searchlight analysis was conducted. The first step involved creating a binarized functional brain mask from the skull-stripped reference brain volume. Next, z-scored volumes from the perceptual and mental imagery runs were masked to exclude non-brain voxels with zero values. First, we performed a cross-domain searchlight analysis in the subject's native space using the Brain Imaging Analysis Kit (BrainIAK v. 0.9.1; http://brainiak.org; [62]. A searchlight of 9 * 9 * 9 voxels was defined by a cube cluster of 4-voxel (3 mm * 4 voxels = 12 mm) radius. The scans from the perceptual runs were labeled to train a standard logistic regression model (with L2 regularization). This classifier was then applied to the imagery runs in order to decode the content of mental imagery volumes as either 'Dogs' or 'Scissors' categories. Specifically, we utilized the perceptual data within each searchlight in the training process, excluding the data from visual noise samples from our searchlight analyses. We evaluated the performance of our model by calculating the area under the receiver operating characteristics curve (AUC), and then assigned that performance score of each searchlight to the center voxel of each sphere. This produced a comprehensive whole-brain map of cross-domain decoding outcomes (see Figure S1). At the group level,

we observed significant decoding scores in several regions, including the bilateral fusiform gyrus, calcarine cortex, precuneus, supramarginal gyrus and surrounding areas. Then, for each participant, we manually created a spherical mask in their native functional space centered around these voxels showing the highest degree of generalization from perception to mental imagery. The brain areas showing better classification performance varied to some extent across participants. While the fusiform region was our a priori target, the first participant also showed significant cross-domain decoding in the parietal cortex. Consequently, we incorporated this region alongside the fusiform into the DecNef training. The second and third participants showed higher decoding accuracy on the left compared to the right fusiform gyrus and we elected to use the left fusiform as target for DecNef. However, for subsequent participants, we opted to utilize a uniform region for DecNef training, centered in the bilateral fusiform gyrus. This region consistently exhibited a high level of decoding AUC across these three participants. The cross-domain decoding whole-brain maps were visually inspected for each participant to draw the ROI for DecNef training and we tried to use a consistent size of the ROI across participants in both left and right fusiform areas. On average, the spheres covered 561.35 voxels, with a standard deviation of ± 150.29 voxels. Spheres from the left and right fusiform regions were merged to create a single bilateral fusiform mask, which was then binarized. The stacked perceptual and mental simulation data was ultimately masked to preserve only the voxels located within the targeted bilateral ROI.

Decoder construction

Sparse logistic regression models [76] were typically used in previous DecNef training studies [19–21, 77]. In our experimental piloting, we found that logistic models generated well-calibrated results when trained and tested on the perceptual data. However, despite implementing overfitting reduction techniques, we observed that these models produced highly bimodal probability distributions and tended to make over-confident out-of-sample-predictions in the imagery runs, but also during generalization to DecNef training sessions [39]. Moreover, we note that implicit DecNef procedures require participants to explore a broad range of strategies and representational spaces beyond the target and non-target classes to maximize the size of the disk and thereby rewards. However, previous DecNef studies [17] have not directly accounted for representational noise during DecNef training. This issue may reduce the capacity of the model to discriminate target from non-target representations, constraining the feedback signal and thereby hinder participants' learning. These issues were addressed by developing a custom Logistic Regression model using Keras [78] (v2.7) and Tensorflow [79] (v2.7). This model included an input and a softmax output layer both fully-connected, and was designed to classify samples in just two categories: 'Dogs' or 'Scissors'. To prevent overfitting, we applied regularization to the kernel weights and layer outputs using an L2 kernel regularizer of 1e-1 and an L1 activity regularizer of 1e-3, respectively. We used a default Stochastic Gradient Descent optimizer with a learning rate of 1e-4. Soft labeling was used, specifically we employed Binary Cross-Entropy loss as the loss function with 'Scissors' labeled as 0, 'Dogs' as 1, and visual noise images as 0.5. By adopting this strategy, we achieved a continuous representational gradient between categories and decreased the network's susceptibility to representational noise during testing. Consequently, the network demonstrated better discrimination between relevant classes by producing moderated predicted probabilities. To evaluate the performance of our model, we utilized the AUC. We maintained a tolerance of 1e-4 and a patience of 50. Additionally, we utilized a batch size of 64 and the training was stopped when the validation loss did not improve for 10 consecutive epochs. For each participant, the model was trained using the mask of the ROI derived from the searchlight analysis noted above. The perceptual data was used as the training set

and the model's hyperparameters were fine-tuned using the mental imagery data within the same ROI as the validation dataset. This approach improved the model's generalizability to the DecNef training sessions. A comparison among the Standard, Sparse, and our Custom regression models that use BCE Loss can be found in the Supplementary Results.

DecNef training sessions

The DecNef training consisted of four fMRI sessions, each lasting approximately 70 minutes, performed on consecutive days at a rate of one session per day in the early morning (8am). In the first three sessions, there were eight fMRI runs, each lasting around 6.5 minutes and containing 18 trials per run. The last session only comprised two fMRI runs of the same length, followed by a behavioural evaluation in the MRI scanner. During the DecNef training, participants were presented with one of two Japanese hiraganas (counterbalanced across participants) over the four days. These were meaningless symbols since participants had no prior experience with the Japanese language. We aimed to use reinforcement learning to establish an unconscious association between perceptual brain representations of 'Dogs' and the neutral hiragana. In order to maximize the consistency of DecNef training effects on both neural and behavioural data, we elected to target DecNef on the representation of 'Dogs'. Additionally, the ROI targeted for DecNef in the bilateral fusiform gyrus is known to exhibit a stronger response to living things compared to non-living things, which could make 'Dogs' a more effective target category for inducing neural changes [80–82]. The procedure was conducted in an implicit manner. Participants only knew that they were taking part in a neurofeedback study aimed at self-regulation of brain activity, but remained completely unaware of the actual objective of the intervention. To prevent participants from associating the prior fMRI session used for decoder construction and the subsequent DecNef training, a new experimenter was assigned, and they were explicitly informed that they were taking part in a different experiment. Furthermore, participants were required to sign different consent forms and were provided with different information sheets. At the beginning of every trial, a gray hiragana was displayed against a black background for 6 seconds. During the presentation of the hiragana, participants were instructed to manipulate their brain activity to increase the size of the feedback disk and hence monetary rewards. No additional guidance or instructions were further provided. They were only encouraged to try different strategies and choose those that yielded the best performance. Participants were explicitly asked to avoid strategies that could potentially introduce artifacts to the fMRI analysis, such as manipulating their breath, moving their body etc. Following the presentation of the hiragana, participants were given a 6-second relaxation period with a white fixation point, while waiting for feedback. The feedback was then displayed as a red circle for one second. The size of the circle indicated the decoder's probability of the target class ranging from 0 to 1 (i.e. the similarity between the evoked brain representations and the brain activity pattern associated with 'Dogs' according to the previously trained decoder). We explained to participants that their monetary reward would increase as the size of the red circle displayed as feedback increased. The maximum reward per trial was indicated by a white circumference enclosing the red circle, corresponding to 100% similarity between the participant's brain activity and the model's perceptual representations of 'Dogs'. Additionally, a yellow circumference was used to set a threshold below which no reward would be given for that trial. This threshold was set at 50% similarity. If participants reached a point where they were no longer making progress or their performance had plateaued, we encouraged them to assess their current mental strategies and experiment with new ones to achieve better outcomes. Each trial concluded with a fixed inter-trial interval of 6 seconds. The final DecNef session had only two fMRI runs and participants were instructed to utilize their most effective mental strategies so far to increase the size

of the red circle. Before the final behavioural session, participants answered several questions regarding the amount of effort needed to enhance monetary rewards and their perceived ability to manipulate the circle's size. We employed a Likert scale from 1 to 10, where the lowest value signified *"minimal effort/control"* and the highest value represented *"significant effort/control"*. We also asked them about the strategies they used during DecNef sessions and any hypotheses they had concerning the experimental objectives. At the start of the DecNef training, participants were told that they would receive a fixed compensation of 85€ for participating in the experiment, as well as the possibility of earning a bonus of 25€ based on their overall performance and participants were informed of the potential to earn up to 0.1€ per trial. After each session, participants were given an estimated report of their earnings designed to maintain their motivation. However, in order to fulfill potential ethical considerations, all participants were given a fixed amount of 110€ at the end of the study, which included the 85€ compensation for their participation and the 25€ bonus reward.

fMRI decoding in real-time

Our PyDecNef framework required three high-performance computers: the computer controlling the MRI scanner; a server computer, which analyzed the data in real time; and a client computer for the presentation of stimuli and feedback to participants during MRI scanning. The three computers were connected through an ethernet local area network. Each DecNef run started with the acquisition of a whole-brain baseline of 20 volumes. FMRI volumes were transferred to a folder on the server computer through the use of the Framewise Integrated Real-time MRI Monitoring Framework [83]. A Python script on the server computer co-registered each fMRI volume with the reference volume using the same preprocessing pipeline as the model construction session. To prevent potential processing overlap between incoming volumes, it was crucial that the co-registration times were consistently shorter than the fMRI TR (in this case, co-registration was approximately 0.7 sec/volume vs. TR = 2 sec). To minimize processing times, our pyDecNef Python framework leverages the Nipype interface module to connect Python with neuroimage software written in computationally efficient programming languages such as C, including dcm2nii and AFNI. Further, to prevent potential overlap during processing of fMRI volumes, pyDecNef includes control measures that allow for independent processing of volumes in separate threads running in parallel. Following the co-registration of each volume, we applied the functional mask of the bilateral fusiform region of interest (ROI) previously extracted for each participant. Spatial smoothing was not performed. Then, the data were linearly detrended using the entire functional time series of that run, including the baseline data and the information from the last volume. Previous studies have utilized z-scoring in neurofeedback either based on run/session baselines or prior data from the current run [84]. In contrast, here we used the mean and standard deviation of the scans used for building the decoder as standardization parameters. Using a fixed baseline across the entire DecNef training has several advantages. Firstly, it enhances the interpretability of our results since all runs and sessions are standardized to the same reference data distribution. Secondly, it ensures that the test data is scaled to the same distribution as the training data, which may contribute to improve the model's performance. Volumes between 5 to 11 seconds after hiragana's appearance were selected for decoding. This timeframe matched the hiragana's Hemodynamic Response Function (HRF) peak. Approximately three volumes were decoded per trial in order to provide feedback to participants. Previous DecNef studies typically average the volumes within the HRF peaks into a single volume to improve the signal-to-noise ratio [17, 20]. However, as noted earlier, logistic regression models may produce highly bimodal category probabilities in cross-domain classification of single volumes, which could make it difficult for participants

to discern subtle changes and nuances at the representational level. To address this issue, we aimed to increase feedback variability at the trial level by computing the probability of each volume belonging to the target category, and then averaging the three probabilities. This average was sent to the client's computer as feedback for the participants. Volumes recorded after the HRF peak offset for each trial were subjected to preprocessing and added to the run time series to facilitate detrending for subsequent trials. However, these volumes were not included in any further analyses.

Visual search task

Following the last two DecNef training runs, we performed a visual search task while the participants were still inside the MRI scanner with the aim to assess the behavioural and neural effects of DecNef training regarding the creation of new conceptual associations. The task comprised five fMRI runs, each with approximately 6 minutes and 56 trials per run. Each trial started with a white fixation mark displayed on a black background for 500 ms. Subsequently, participants were randomly presented with either the targeted hiragana by DecNef or a control hiragana they had not seen before for 1.5 seconds. The hiragana was presented at fixation and subtended 1.29 degrees of visual angle. Then, participants were briefly presented with two images (a 'Dog' example and a 'Scissor' example; each subtending 8.06 degrees of visual angle), which were displayed to the left and right from the fixation point for 150 ms. The search display was followed by the presentation of a letter indicating the target image for participants' responses: 'P' for 'Dogs' (from the Spanish word *'Perros'*) and 'T' for scissors (from the Spanish word *'Tijeras'*). This letter was presented for 1750 ms, which was also the deadline for participants to respond. Participants were asked to use their dominant hand (right) to indicate the location of the picture corresponding to the target category on a MRI-compatible response pad as quickly and accurately as possible. Specifically, they were instructed to press the button on the right or left side of the pad that corresponded to the location of the picture. To align the orientation of the buttons with their left and right hand positions, they were instructed to hold the response pad horizontally. In addition, participants were instructed to use their index and middle fingers of their dominant hand to press the response pad buttons. To ensure that all variables were counterbalanced in the experiment, a total of 56 trials were conducted, with each of the eight possible combinations of the three variables Hiragana (targeted vs. novel), Picture position ('Dogs' Left - 'Scissors' Right vs. 'Scissors' Right - 'Dogs' Left), and Target category ('Dogs' vs. 'Scissors') being presented an equal number of times (7 trials each). The order of presentation for each combination was randomized across participants. A jittered interstimulus interval (ITI) ranging from 2.5 to 4.5 seconds with a pseudo-exponential distribution was used between trials, with 50% of ITIs set to 2.5 seconds, 25% to 3 seconds, 11% to 3.5 seconds, 7% to 4 seconds, and 7% to 4.5 seconds. This was done to help estimate BOLD responses across trials in univariate off-line analyses [85]. Before the experimental session began, participants completed some practice trials without any hiragana presentation, where a fixed ITI of one second was used.

The participants in the control group performed the same visual search task, but outside the fMRI scanner under matched experimental conditions regarding stimulus presentation and timing of events. Also, participants listened to scanner noise via headphones while they performed the task. Before engaging in practice trials, the control group was presented with a block-wise design where images of 'Dogs' and 'Scissors' were shown, similar to the session in the DecNef group used to build the classifier. Subsequently, they practiced the visual search task without any hiragana presentation. Finally, they completed the visual search task using the same timings and jitters as those used by the fMRI participants. Following the behavioural evaluation task, we debriefed participants into the goals of the experiment and asked them to predict the

category targeted by DecNef ('Scissors' or 'Dogs'). They were also required to express their confidence in their answer using a 4-point Likert Scale (0 - *"Chance"*; 1 - *"Not so sure"*; 2 - *"Sure"*; and 3 - *"Really sure"*).

Mixed Effects Modelling Analyses

Mixed effects modeling was used to examine participant's learning effects throughout the DecNef training, namely, whether participants could progressively enhance their ability to induce neural patterns related to the target category over the course of training. The design involved repeated measures for each participant and a hierarchical structure comprising two groups of hiraganas, participants, sessions, runs and trials. In cases of non-independence of the data, multilevel models, specifically Linear Mixed-effects Models (LMMs), have gained popularity in psychological research. LMMs offer several advantages, such as accounting for both within-subject and between-subject variability, handling missing data, and examining relationships between variables at different hierarchical levels. Traditional analysis methods, such as repeated-measures ANOVA, often aggregate data within a single condition or hierarchical level, assuming data independence, which can lead to a loss of information and statistical power. In comparison, LMMs can capture variability in the observations by leveraging all levels of data, using both fixed and random effects modeling [86, 87]. To perform the linear mixed model analysis, we used the *lme4* package [88] in the R programming language (R Core Team, 2021). We employed the lmerTest package [89] to determine the statistical significance of our results. This package estimated p-values for mixed models using Satterthwaite's method to approximate the degrees of freedom for t and F tests.

LMM analysis of DecNef induction likelihoods

In our LMM analysis, we used the decoding probability of each trial as the dependent variable and aimed to explain it by examining the impact of the global trial index from trial 1 to 468 as an independent variable and fixed effect. We also sought to account for the potential variability in the data arising from the hierarchical structure of our study. To achieve this, we introduced a random intercept for every distinct combination of the hiragana group, participant (1 - 18), session (1 - 4), and run (1 - 8) levels. We used the LMM formula '*lmer(trial_probability ∼ global_trial_index + (1 | hiragana group / participant / session / run))*' to conduct our analyses in R. The dependent variable was '*trial_probability*', the independent variable was '*global_trial_index*', and the term '*(1 | hiragana group / subject_nr / session / run)*' represented the random intercepts. Importantly, the global trial index variable was treated as a continuous numeric variable, while the hiragana group, participant, session and run variables were categorical variables with discrete levels.

LMM analysis of behavioural data

To examine the effects of the DecNef training on the visual search task, a second LMM was utilized to analyze the reaction times (RTs) of the correct responses. RTs outside of 3 standard deviations from each participant's average RT were excluded. The factors hiragana (targeted vs. control), search target ('Dogs' vs. 'Scissors') and their interaction were set as fixed effects. Additionally, we included each participant in the DecNef group as a random effect to account for inter-subject variability with the next formula : *lmer(RTs ∼ Hiragana + Search Target + Hiragana * Search Target + (1 | Participant))*. In this model, the control hiragana level was used as the reference level for the hiragana factor, while the scissors level was used as the reference level for the search target factor.

Moreover, in order to compare the reaction times (RTs) between the DecNef and Control groups, we

combined the RTs across the levels of the Hiragana factor (targeted vs. novel), as this factor did not apply to the control group. Then, we conducted a similar LMM using the following formula: *lmer(Decision RTs ∼ Search Target + Experimental Group + Search Target * Experimental Group + (1 | Participant))*. In this model, the scissors level was used as the reference level for the search target factor, while the control group level was used as the reference level for the experimental group factor.

Whole-brain searchlight analysis

We used searchlight analyses to examine the probability of detecting the target category throughout the DecNef sessions. This analysis aimed to verify the effectiveness of our approach to target the bilateral fusiform gyrus with DecNef and also explore the involvement of other brain regions in the representation of the target category. The analysis employed a searchlight method, where a classifier trained on the perceptual data (previously used for decoder construction) was then applied to compute the probability of the living category for all DecNef training volumes within trial-related HRF peaks (i.e., from 5 to 11 seconds after hiragana's onset). For each subject, we used a similar approach to the searchlight analysis used for ROI delineation. This perception-to-DecNef cross-domain searchlight analysis was conducted by using BrainIAK [62]. In order to guarantee that the searchlight analysis results were highly comparable to the real-time DecNef outcomes, it would have been ideal to use our custom logistic regression model with Binary Cross-Entropy loss for the searchlight analysis. However, the significant training time required for each searchlight prevented us from using it. Instead, we trained a Standard Logistic Regression model (L2 regularization) using the perceptual volumes as the training data. For every searchlight, we utilized voxel data to forecast the probability of the 'Dogs' category for each of the DecNef training volumes. The probability obtained from each searchlight was assigned to the corresponding target voxel, producing a whole-brain map of probability scores for each volume. To obtain a single probability map per subject, the probability maps for all DecNef training volumes were averaged. For each subject, we adjusted the whole-brain maps by subtracting 0.5 from every probability score, effectively setting the chance-level performance to zero. These maps were then registered from the functional to the participant's T1-weighted structural space using *epi_reg* FSL tool [69], and from the structural space to the Montreal Neurological Institute (MNI) standard space (MNI-152 template, 2 * 2 * 2 mm^3) using FMRIB's Linear Image Registration Tool [69, 90]. A non-parametric approach was used to perform statistical contrasts. Participants' standardized 3D masks were concatenated together using fslmerge, a tool from the FSL software package, to create a single 4D image file. A non-parametric one-sample t-test was used to identify these brain regions where the average probabilities were significantly greater than zero (i.e., above chance level). This t-test contrast was calculated using FSL Randomise with Threshold-Free Cluster Enhancement (TFCE cluster correction) and 10,000 random permutations across subjects [91]. To increase the statistical power of the analyses, a spatial FWHM Gaussian Kernel of 5 mm was used for variance smoothing. The Family-wise Error (FWE) correction was applied to adjust for multiple comparisons. After performing t-test contrasts using Randomise, a 1-p corrected statistical image was obtained. Subsequently, that image was thresholded to display only these statistically significant clusters with p-values less than or equal to 0.05. For better visualization, a global average mask was generated by averaging the 4D image file that comprised the masks of all participants. To restore the chance level to 0.5, we added 0.5 to every probability value in the average mask. Finally, the average map was projected onto previously recognized significant clusters, displaying only the average values within those particular brain regions.

Univariate analyses performed on the visual search task

The FSL suite tools were used for conducting univariate analyses (FMRIB Software Library v6.0 available at fsl.fmrib.ox.ac.uk/fsl/fslwiki/FSL) [90, 92]. First, the T1-weighted structural image was cropped by employing the FSL Robust Field of View (FOV) method to remove the neck and lower head slices. The resulting image was then skull-stripped using the Brain Extraction Tool. Next, RAW functional data in NifTi format was preprocessed using fMRI Expert Analysis Tool (FEAT) [93, 94]. To ensure signal stabilization, the first five volumes of each functional run were discarded and the MCFLIRT algorithm was applied to correct for head-motion [90]. Non-brain voxels were removed through functional brain extraction, and the data was smoothed with a 5 mm full-width-half maximum (FWHM) gaussian kernel. Additionally, high-pass temporal filtering was performed with sigma = 60s. Finally, the functional data was linearly registered to the high-resolution structural brain image using a Boundary Based Registration (BBR) cost function. Then, linear affine transformations with 12 parameters were used to transform the data into the standard MNI-152 space (2*2*2 mm^3).

Following the preprocessing of the data, we carried out hierarchical general linear modeling using FMRIB's Improved Linear Modelling (FILM). For each visual search run, the first level analysis modeled the time-series of each voxel by incorporating four regressors [93]: Targeted Hiragana - 'Dogs' Target, Targeted Hiragana - 'Scissors' Target, Control Hiragana - 'Dogs' Target, and Control Hiragana - 'Scissors' Target. Explanatory variables in the GLM were specified based on the onset of hiragana in each trial across the different experimental conditions. These onset times were adjusted to account for the 5 warm-up volumes that were discarded in the preprocessing stage. A double-gamma canonical hemodynamic response function was used to model each trial in the GLM beginning with the onset of the hiragana. To address motion-related confounding factors, the standard and extended motion parameters estimated by MCFLIRT were included as covariates. Additionally, to account for temporal autocorrelation in the voxel time-series, prewhitening was performed on the data. To improve the model, we incorporated temporal derivatives and filtering to the design matrix, which helped to reduce the effects of lags in the onset of the H.R.F and minimize the presence of low-frequency noise over time. Four regressors were used: Control Hiragana - 'Dogs' target, Control Hiragana - 'Scissors' target, Targeted Hiragana - 'Dogs' target, Targeted Hiragana - 'Scissors' target. For each run, we calculated 4 contrasts of parameter estimates (COPES) including Control Hiragana > Targeted Hiragana (i.e., 1, 1, -1, -1); Control Hiragana < Targeted Hiragana (i.e., -1, -1, 1, 1); Control Hiragana - 'Dogs' > Targeted Hiragana - 'Dogs' & Control Hiragana - 'Scissors' < Targeted Hiragana - 'Scissors' (i.e., 1, -1, -1, 1); Control Hiragana - 'Dogs' < Targeted Hiragana - 'Dogs' & Control Hiragana - 'Scissors' > Targeted Hiragana - 'Scissors' (i.e., -1, 1, 1, -1). Then, we combined the first-level COPEs files into a second-level fixed-effects model. Finally, group-based, third-level analyses [94, 95] were performed using the second-level COPEs files obtained in the previous step for each individual. To address inter-subject variability within each group, a mixed-effects analysis was conducted using the FMRIB Local Analysis of Mixed Effects (FLAME) 1 + 2 algorithm. Although the bilateral fusiform gyrus was our a priori ROI, DecNef inception-related neural modulation could hypothetically occur in any region across the semantic network or ventral visual pathway. Therefore, we conducted first whole-brain analyses to examine group effects in the visual search task without constraining the analysis to any particular brain region. This allowed us to investigate the possibility of conceptual representational changes in brain areas beyond the bilateral fusiform gyrus. Group-based analyses used cluster based thresholding, with a voxel-wise threshold of Z = 3.1 and a cluster significance threshold of P = 0.05, corrected for multiple comparisons using Gaussian Random Field Theory. The uncorrected contrast of parameter estimates and zstats for the different contrasts are

available at: https://pedromargolles.github.io/pyDecNef/. Following the whole-brain analyses to examine group effects, we focused on the bilateral fusiform gyrus as our region of interest (ROI) by running the GLM on binary masks of the fusiform in MNI space. The initial space of the ROI was selected based on an MVPA decoding analysis with the objective of identifying these voxels around the bilateral fusiform gyrus that accurately discriminated between perceived living and non-living stimuli among participants. We accomplished this by performing a whole-brain decoding searchlight analysis in MNI space, where a classifier was trained iteratively using functional perceptual data from 17 participants to predict the picture classes observed by a participant left out of the analysis. To perform that analysis, the first step involved transforming the perceptual volumes to MNI space for every subject. This was achieved by first registering their functional reference volumes to the structural image, and then to MNI space. Then, all detrended and z-scored perceptual volumes for each participant were mapped to the MNI standard template using the same transformation matrices, as the preprocessed volumes had previously been co-registered to reference volumes. Lastly, the MNI-registered labeled data of all participants was concatenated. To prevent data leaks to the test set, a leave-one-subject-out cross-validation procedure was performed by generating 18 training-testing sets of whole-brain data. For each of the 18 participants, their data was left out as the test set, while the data from the remaining 17 participants' was standardized together using Z-scoring and stacked to create the training set. The test set, consisting of the stacked data from the left-out participant, was standardized using z-scoring and scaled based on the mean and standard deviation of the corresponding training set. We conducted 18 separate decoding searchlight analyses, each using the data from one participant as the test set and the perceptual volumes from the remaining 17 participants to train the model, using previously reported searchlight sizes. The searchlight analyses were limited to a broad mask around bilateral fusiform regions in the MNI space to reduce computational time costs. The resulting decoding searchlight maps from all participants were combined and averaged. A binarized mask was created manually covering these voxels in the bilateral fusiform with higher average AUC among participants. Our second approach to selecting the region of interest was based on the identification of regions with higher probabilities of the 'Dogs' category in the previous searchlight analysis from perception to DecNef volumes. These regions were hypothesized to be the areas where inception might have occurred across participants, and therefore, could potentially exhibit activation differences during the visual search task. In our final analysis, we aimed to examine the average voxel activation within the bilateral fusiform masks utilized for each participant during DecNef training. To do so, we repeated our first-level analyses by registering the functional data from the visual search task to the reference volume used during DecNef training sessions with FEAT. Then, we utilized the outputs from the second-level fixed-effects in the functional native space. Specifically, we computed the average activation within the ROI masks using the FSL utility fslmeants, over the 4 COPEs corresponding to each condition (i.e., Control Hiragana | 'Dogs' Target; Control Hiragana | 'Scissors' Target; Targeted Hiragana | 'Dogs' Target; and Targeted Hiragana | 'Scissors' Target). We obtained a mean value for each subject and COPE at that ROI, which were then compared across participants through a repeated measures ANOVA analysis.

Open Science Statement

The scripts and documentation of the Python library for decoded neurofeedback (PyDecNef) can be found at: https://pedromargolles.github.io/pyDecNef/. The experimental data will be uploaded to OpenNeuro upon acceptance for publication.

Acknowledgements

P.M acknowledges support from an FPI grant from the Spanish Ministry of Economy and Competitiveness. P.E. acknowledges support from the Basque Government PREDOC grant. D.S. acknowledges support from the Basque Government through the BERC 2022-2025 program, from the Spanish Ministry of Economy and Competitiveness, through the 'Severo Ochoa' Programme for Centres/Units of Excellence (CEX2020-001010-S) and also from project grants PID2019-105494GB-I00. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing interests statement

The authors declare no competing interests.

References

[1] M. D. Smith, "The acquisition of word meaning: An introduction," *Child Development*, pp. 950–952, 1978.

[2] P. Blewitt, "Word meaning acquisition in young children: A review of theory and research," *Advances in child development and behavior*, vol. 17, pp. 139–195, 1982.

[3] A. R. Seitz, D. Kim, and T. Watanabe, "Rewards evoke learning of unconsciously processed visual stimuli in adult humans," *Neuron*, vol. 61, no. 5, pp. 700–707, 2009.

[4] M. Pessiglione, P. Petrovic, J. Daunizeau, S. Palminteri, R. J. Dolan, and C. D. Frith, "Subliminal instrumental conditioning demonstrated in the human brain," *Neuron*, vol. 59, no. 4, pp. 561–567, 2008.

[5] L. I. Skora, M. R. Yeomans, H. S. Crombag, and R. B. Scott, "Evidence that instrumental conditioning requires conscious awareness in humans," *Cognition*, vol. 208, p. 104546, 2021.

[6] C. M. Raio, D. Carmel, M. Carrasco, and E. A. Phelps, "Nonconscious fear is quickly acquired but swiftly forgotten," *Current Biology*, vol. 22, no. 12, pp. R477–R479, 2012.

[7] W. Hofmann, J. De Houwer, M. Perugini, F. Baeyens, and G. Crombez, "Evaluative conditioning in humans: a meta-analysis.," *Psychological bulletin*, vol. 136, no. 3, p. 390, 2010.

[8] A. W. Staats, C. K. Staats, W. G. Heard, and L. P. Nims, "Meaning established by classical conditioning.," *Journal of Experimental Psychology*, vol. 57, no. 1, p. 64, 1959.

[9] W. W. Tryon and S. D. Cicero, "Classical conditioning of meaning—i. a replication and higher-order extension," *Journal of behavior therapy and experimental psychiatry*, vol. 20, no. 2, pp. 137–142, 1989.

[10] D. R. Shanks and M. F. S. John, "Characteristics of dissociable human learning systems," *Behavioral and brain sciences*, vol. 17, no. 3, pp. 367–395, 1994.

[11] G. Mertens and I. M. Engelhard, "A systematic review and meta-analysis of the evidence for unaware fear conditioning," *Neuroscience & Biobehavioral Reviews*, vol. 108, pp. 254–268, 2020.

[12] A. P. Field, "I like it, but i'm not sure why: Can evaluative conditioning occur without conscious awareness?," *Consciousness and cognition*, vol. 9, no. 1, pp. 13–36, 2000.

[13] D. R. Shanks and A. Dickinson, "Contingency awareness in evaluative conditioning: A comment on baeyens, eelen, and van den bergh," *Cognition and Emotion*, vol. 4, no. 1, pp. 19–30, 1990.

[14] M. M. Page, "Demand characteristics and the classical conditioning of attitudes experiment.," *Journal of Personality and Social Psychology*, vol. 30, no. 4, p. 468, 1974.

[15] C. Stahl, J. Haaf, and O. Corneille, "Subliminal evaluative conditioning? above-chance cs identification may be necessary and insufficient for attitude learning.," *Journal of Experimental Psychology: General*, vol. 145, no. 9, p. 1107, 2016.

[16] H. Lau, *In consciousness we trust: The cognitive neuroscience of subjective experience*. Oxford University Press, 2022.

[17] T. Watanabe, Y. Sasaki, K. Shibata, and M. Kawato, "Advances in fmri real-time neurofeedback," *Trends in cognitive sciences*, vol. 21, no. 12, pp. 997–1010, 2017.

[18] K. Shibata, T. Watanabe, M. Kawato, and Y. Sasaki, "Differential activation patterns in the same brain region led to opposite emotional states," *PLoS biology*, vol. 14, no. 9, p. e1002546, 2016.

[19] A. Koizumi, K. Amano, A. Cortese, K. Shibata, W. Yoshida, B. Seymour, M. Kawato, and H. Lau, "Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure," *Nature human behaviour*, vol. 1, no. 1, p. 0006, 2016.

[20] V. Taschereau-Dumouchel, A. Cortese, T. Chiba, J. Knotts, M. Kawato, and H. Lau, "Towards an unconscious neural reinforcement intervention for common fears," *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. 3470–3475, 2018.

[21] A. Cortese, K. Amano, A. Koizumi, M. Kawato, and H. Lau, "Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance," *Nature communications*, vol. 7, no. 1, p. 13669, 2016.

[22] D. Soto, G. W. Humphreys, and P. Rotshtein, "Dissociating the neural mechanisms of memory-based guidance of visual selection," *Proceedings of the National Academy of Sciences*, vol. 104, no. 43, pp. 17186–17191, 2007.

[23] D. Soto, J. Hodsoll, P. Rotshtein, and G. W. Humphreys, "Automatic guidance of attention from working memory," *Trends in cognitive sciences*, vol. 12, no. 9, pp. 342–348, 2008.

[24] D. Cousineau and F. O'Brien, "Error bars in within-subject designs: a comment on baguley (2012)," *Behavior Research Methods*, vol. 46, pp. 1149–1151, 2014.

[25] S. Wang, N. Tsuchiya, J. New, R. Hurlemann, and R. Adolphs, "Preferential attention to animals and people is independent of the amygdala," *Social Cognitive and Affective Neuroscience*, vol. 10, no. 3, pp. 371–380, 2015.

[26] J. New, L. Cosmides, and J. Tooby, "Category-specific attention for animals reflects ancestral priorities, not expertise," *Proceedings of the National Academy of Sciences*, vol. 104, no. 42, pp. 16598–16603, 2007.

[27] J. Tipples, A. W. Young, P. Quinlan, P. Broks, and A. W. Ellis, "Searching for threat," *The quarterly journal of experimental psychology section A*, vol. 55, no. 3, pp. 1007–1026, 2002.

[28] C. He and O. S. Cheung, "Category selectivity for animals and man-made objects: beyond low-and mid-level visual features," *Journal of vision*, vol. 19, no. 12, pp. 22–22, 2019.

[29] C. Summerfield, E. H. Trittschuh, J. M. Monti, M.-M. Mesulam, and T. Egner, "Neural repetition suppression reflects fulfilled perceptual expectations," *Nature neuroscience*, vol. 11, no. 9, pp. 1004–1006, 2008.

[30] D. Soto, C. M. Greene, A. Chaudhary, and P. Rotshtein, "Competition in working memory reduces frontal guidance of visual selection," *Cerebral cortex*, vol. 22, no. 5, pp. 1159–1169, 2012.

[31] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 61, no. 3, pp. 611–622, 1999.

[32] D. Deitch, A. Rubin, and Y. Ziv, "Representational drift in the mouse visual cortex," *Current biology*, vol. 31, no. 19, pp. 4327–4339, 2021.

[33] O. Mazor and G. Laurent, "Transient dynamics versus fixed points in odor representations by locust antennal lobe projection neurons," *Neuron*, vol. 48, no. 4, pp. 661–673, 2005.

[34] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, S. I. Ryu, and K. V. Shenoy, "Cortical preparatory activity: representation of movement or first cog in a dynamical machine?," *Neuron*, vol. 68, no. 3, pp. 387–400, 2010.

[35] M. B. Ahrens, J. M. Li, M. B. Orger, D. N. Robson, A. F. Schier, F. Engert, and R. Portugues, "Brain-wide neuronal dynamics during motor adaptation in zebrafish," *Nature*, vol. 485, no. 7399, pp. 471–477, 2012.

[36] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[37] K. Shibata, G. Lisi, A. Cortese, T. Watanabe, Y. Sasaki, and M. Kawato, "Toward a comprehensive understanding of the neural mechanisms of decoded neurofeedback," *NeuroImage*, vol. 188, pp. 539–556, 2019.

[38] K. Shibata, T. Watanabe, Y. Sasaki, and M. Kawato, "Perceptual learning incepted by decoded fmri neurofeedback without stimulus presentation," *science*, vol. 334, no. 6061, pp. 1413–1415, 2011.

[39] Y. Bai, S. Mei, H. Wang, and C. Xiong, "Don't just blame over-parametrization for over-confidence: Theoretical analysis of calibration in binary classification," in *International Conference on Machine Learning*, pp. 566–576, PMLR, 2021.

[40] K. C. Kadosh and G. Staunton, "A systematic review of the psychological factors that influence neurofeedback learning outcomes," *Neuroimage*, vol. 185, pp. 545–555, 2019.

[41] M. Stirner, G. Gurevitch, N. Lubianiker, T. Hendler, C. Schmahl, and C. Paret, "An investigation of awareness and metacognition in neurofeedback with the amygdala electrical fingerprint," *Consciousness and Cognition*, vol. 98, p. 103264, 2022.

[42] N. Lubianiker, C. Paret, P. Dayan, and T. Hendler, "Neurofeedback through the lens of reinforcement learning," *Trends in Neurosciences*, 2022.

[43] S. Muñoz-Moldes and A. Cleeremans, "Delineating implicit and explicit processes in neurofeedback learning," *Neuroscience & Biobehavioral Reviews*, vol. 118, pp. 681–688, 2020.

[44] J. E. LeDoux and D. S. Pine, "Using neuroscience to help understand fear and anxiety: a two-system framework," *American journal of psychiatry*, 2016.

[45] P. F. Lovibond and D. R. Shanks, "The role of awareness in pavlovian conditioning: empirical evidence and theoretical implications.," *Journal of Experimental Psychology: Animal Behavior Processes*, vol. 28, no. 1, p. 3, 2002.

[46] K. Amano, K. Shibata, M. Kawato, Y. Sasaki, and T. Watanabe, "Learning to associate orientation with color in early visual areas by associative decoded fmri neurofeedback," *Current Biology*, vol. 26, no. 14, pp. 1861–1866, 2016.

[47] S. H. Collin, P. L. van den Broek, T. van Mourik, P. Desain, and C. F. Doeller, "Inducing a mental context for associative memory formation with real-time fmri neurofeedback," *Scientific Reports*, vol. 12, no. 1, p. 21226, 2022.

[48] R. M. Reinhart, L. J. McClenahan, and G. F. Woodman, "Visualizing trumps vision in training attention," *Psychological science*, vol. 26, no. 7, pp. 1114–1122, 2015.

[49] A. C. Nobre and M. G. Stokes, "Premembering experience: A hierarchy of time-scales for proactive attention," *Neuron*, vol. 104, no. 1, pp. 132–146, 2019.

[50] E. M. Tartaglia, L. Bamert, F. W. Mast, and M. H. Herzog, "Human perceptual learning by mental imagery," *Current Biology*, vol. 19, no. 24, pp. 2081–2085, 2009.

[51] M. E. Rule, T. O'Leary, and C. D. Harvey, "Causes and consequences of representational drift," *Current opinion in neurobiology*, vol. 58, pp. 141–147, 2019.

[52] J. L. McClelland, A. G. Thomas, B. D. McCandliss, and J. A. Fiez, "Understanding failures of learning: Hebbian learning, competition for representational space, and some preliminary experimental data," *Progress in brain research*, vol. 121, pp. 75–80, 1999.

[53] Y. Ziv, L. D. Burns, E. D. Cocker, E. O. Hamel, K. K. Ghosh, L. J. Kitch, A. E. Gamal, and M. J. Schnitzer, "Long-term dynamics of ca1 hippocampal place codes," *Nature neuroscience*, vol. 16, no. 3, pp. 264–266, 2013.

[54] L. N. Driscoll, N. L. Pettit, M. Minderer, S. N. Chettih, and C. D. Harvey, "Dynamic reorganization of neuronal activity patterns in parietal cortex," *Cell*, vol. 170, no. 5, pp. 986–999, 2017.

[55] K. Aitken, M. Garrett, S. Olsen, and S. Mihalas, "The geometry of representational drift in natural and artificial neural networks," *PLOS Computational Biology*, vol. 18, no. 11, p. e1010716, 2022.

[56] T. D. Marks and M. J. Goard, "Stimulus-dependent representational drift in primary visual cortex," *Nature communications*, vol. 12, no. 1, p. 5169, 2021.

[57] L. N. Driscoll, L. Duncker, and C. D. Harvey, "Representational drift: Emerging theories for continual learning and experimental future directions," *Current Opinion in Neurobiology*, vol. 76, p. 102609, 2022.

[58] S. Sadeh and C. Clopath, "Contribution of behavioural variability to representational drift," *Elife*, vol. 11, p. e77907, 2022.

[59] J. R. Sato, R. Basilio, F. F. Paiva, G. J. Garrido, I. E. Bramati, P. Bado, F. Tovar-Moll, R. Zahn, and J. Moll, "Real-time fmri pattern decoding and neurofeedback using friend: an fsl-integrated bci toolbox," *PLoS One*, vol. 8, no. 12, p. e81658, 2013.

[60] O. Cohen, R. Malach, M. Koppel, and D. Friedman, "Brain-voyant: A general purpose machine-learning tool for real-time fmri whole-brain pattern classification," in *2018 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2018.

[61] R. Goebel, "Brainvoyager—past, present, future," *Neuroimage*, vol. 62, no. 2, pp. 748–756, 2012.

[62] M. Kumar, M. J. Anderson, J. W. Antony, C. Baldassano, P. P. Brooks, M. B. Cai, P.-H. C. Chen, C. T. Ellis, G. Henselman-Petrusek, D. Huberdeau, *et al.*, "Brainiak: The brain imaging analysis kit," *Aperture neuro*, vol. 1, no. 4, 2021.

[63] Y. Koush, J. Ashburner, E. Prilepin, R. Sladky, P. Zeidman, S. Bibikov, F. Scharnowski, A. Nikonorov, and D. Van De Ville, "Opennft: An open-source python/matlab framework for real-time fmri neurofeedback training based on activity, connectivity and multivariate pattern analysis," *NeuroImage*, vol. 156, pp. 489–503, 2017.

[64] M. N. Hebart, A. H. Dickter, A. Kidder, W. Y. Kwok, A. Corriveau, C. Van Wicklin, and C. I. Baker, "Things: A database of 1,854 object concepts and more than 26,000 naturalistic object images," *PloS one*, vol. 14, no. 10, p. e0223792, 2019.

[65] S. Mathôt, D. Schreij, and J. Theeuwes, "Opensesame: An open-source, graphical experiment builder for the social sciences," *Behavior research methods*, vol. 44, pp. 314–324, 2012.

[66] S. Mathôt and J. March, "Conducting linguistic experiments online with opensesame and osweb," *Language Learning*, vol. 72, no. 4, pp. 1017–1048, 2022.

[67] R. Cox, "Afni: Programvare for analyse og visualisering av funksjonell magnetisk resonans neuroimages," *Comput Biomed Res*, vol. 29, pp. 162–173, 1996.

[68] K. Gorgolewski, C. D. Burns, C. Madison, D. Clark, Y. O. Halchenko, M. L. Waskom, and S. S. Ghosh, "Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python," *Frontiers in neuroinformatics*, p. 13, 2011.

[69] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[70] X. Li, P. S. Morgan, J. Ashburner, J. Smith, and C. Rorden, "The first step for neuroimaging data analysis: Dicom to nifti conversion," *Journal of neuroscience methods*, vol. 264, pp. 47–56, 2016.

[71] T. R. Oakes, T. Johnstone, K. O. Walsh, L. L. Greischar, A. L. Alexander, A. S. Fox, and R. J. Davidson, "Comparison of fmri motion correction software tools," *Neuroimage*, vol. 28, no. 3, pp. 529–543, 2005.

[72] N. Dijkstra, S. van Gaal, L. Geerligs, S. E. Bosch, and M. A. van Gerven, "No evidence for neural overlap between unconsciously processed and imagined stimuli," *Eneuro*, vol. 8, no. 5, 2021.

[73] M. Stokes, R. Thompson, R. Cusack, and J. Duncan, "Top-down activation of shape-specific population codes in visual cortex during mental imagery," *Journal of Neuroscience*, vol. 29, no. 5, pp. 1565–1572, 2009.

[74] L. Reddy, N. Tsuchiya, and T. Serre, "Reading the mind's eye: decoding category information during mental imagery," *Neuroimage*, vol. 50, no. 2, pp. 818–825, 2010.

[75] S.-H. Lee, D. J. Kravitz, and C. I. Baker, "Disentangling visual imagery and perception of real-world objects," *Neuroimage*, vol. 59, no. 4, pp. 4064–4073, 2012.

[76] O. Yamashita, M.-a. Sato, T. Yoshioka, F. Tong, and Y. Kamitani, "Sparse estimation automatically selects voxels relevant for the decoding of fmri activity patterns," *NeuroImage*, vol. 42, no. 4, pp. 1414–1429, 2008.

[77] A. Cortese, K. Amano, A. Koizumi, H. Lau, and M. Kawato, "Decoded fmri neurofeedback can induce bidirectional confidence changes within single participants," *NeuroImage*, vol. 149, pp. 323–337, 2017.

[78] F. Chollet, "Keras.[online] available at: https://github. com/fchollet/keras," *Accessed*, vol. 12, no. 01, p. 2021, 2015.

[79] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[80] P. E. Downing, A.-Y. Chan, M. Peelen, C. Dodds, and N. Kanwisher, "Domain specificity in visual cortex," *Cerebral cortex*, vol. 16, no. 10, pp. 1453–1461, 2006.

[81] A. C. Connolly, J. S. Guntupalli, J. Gors, M. Hanke, Y. O. Halchenko, Y.-C. Wu, H. Abdi, and J. V. Haxby, "The representation of biological classes in the human brain," *Journal of Neuroscience*, vol. 32, no. 8, pp. 2608–2618, 2012.

[82] L. Sha, J. V. Haxby, H. Abdi, J. S. Guntupalli, N. N. Oosterhof, Y. O. Halchenko, and A. C. Connolly, "The animacy continuum in the human ventral vision pathway," *Journal of cognitive neuroscience*, vol. 27, no. 4, pp. 665–678, 2015.

[83] N. U. Dosenbach, J. M. Koller, E. A. Earl, O. Miranda-Dominguez, R. L. Klein, A. N. Van, A. Z. Snyder, B. J. Nagel, J. T. Nigg, A. L. Nguyen, *et al.*, "Real-time motion analytics during brain mri improve data quality and reduce costs," *Neuroimage*, vol. 161, pp. 80–93, 2017.

[84] E. F. Oblak, J. S. Sulzer, and J. A. Lewis-Peacock, "A simulation-based approach to improve decoded neurofeedback performance," *NeuroImage*, vol. 195, pp. 300–310, 2019.

[85] J. Ollinger, M. Corbetta, and G. Shulman, "Separating processes within a trial in event-related functional mri: Ii. analysis," *Neuroimage*, vol. 13, no. 1, pp. 218–229, 2001.

[86] H. Singmann and D. Kellen, "An introduction to mixed models for experimental psychology," in *New methods in cognitive psychology*, pp. 4–31, Routledge, 2019.

[87] L. Meteyard and R. A. Davies, "Best practice guidance for linear mixed-effects models in psychological science," *Journal of Memory and Language*, vol. 112, p. 104092, 2020.

[88] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *arXiv preprint arXiv:1406.5823*, 2014.

[89] A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen, *et al.*, "Package 'lmertest'," *R package version*, vol. 2, no. 0, p. 734, 2015.

[90] M. Jenkinson, C. Beckmann, T. Behrens, and M. Woolrich, "Fsl," *Neuroimage*, vol. 62, pp. 782–790.

[91] S. M. Smith and T. E. Nichols, "Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference," *Neuroimage*, vol. 44, no. 1, pp. 83–98, 2009.

[92] S. M. Smith, M. Jenkinson, M. W. Woolrich, C. F. Beckmann, T. E. Behrens, H. Johansen-Berg, P. R. Bannister, M. De Luca, I. Drobnjak, D. E. Flitney, *et al.*, "Advances in functional and structural mr image analysis and implementation as fsl," *Neuroimage*, vol. 23, pp. S208–S219, 2004.

[93] M. W. Woolrich, B. D. Ripley, M. Brady, and S. M. Smith, "Temporal autocorrelation in univariate linear modeling of fmri data," *Neuroimage*, vol. 14, no. 6, pp. 1370–1386, 2001.

[94] M. W. Woolrich, T. E. Behrens, C. F. Beckmann, M. Jenkinson, and S. M. Smith, "Multilevel linear modelling for fmri group analysis using bayesian inference," *Neuroimage*, vol. 21, no. 4, pp. 1732–1747, 2004.

[95] M. W. Woolrich, T. E. Behrens, and S. M. Smith, "Constrained linear basis sets for hrf modelling using variational bayes," *NeuroImage*, vol. 21, no. 4, pp. 1748–1761, 2004.

[96] Y. Miyawaki, H. Uchida, O. Yamashita, M.-a. Sato, Y. Morito, H. C. Tanabe, N. Sadato, and Y. Kamitani, "Visual image reconstruction from human brain activity using a combination of multiscale local image decoders," *Neuron*, vol. 60, no. 5, pp. 915–929, 2008.

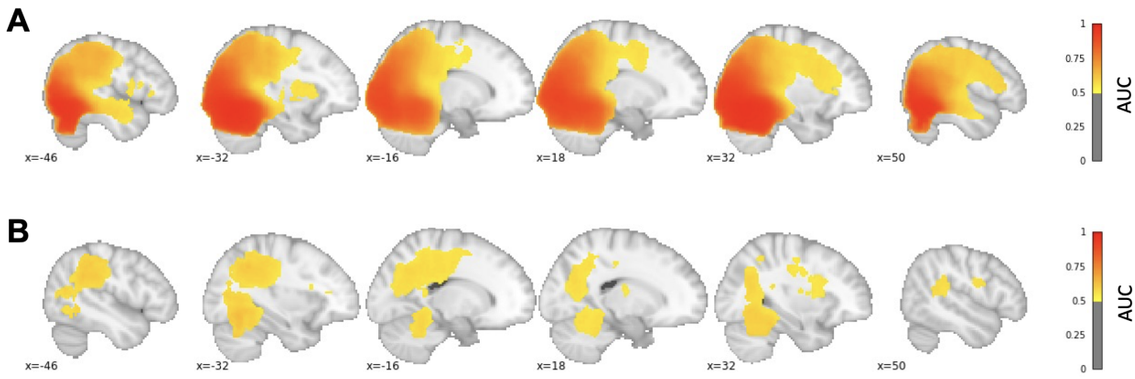
[97] K. Majima, "Python implementation of sparse multinomial logistic regression (smlr)." `https://github.com/KamitaniLab/smlr`, 2015.

[98] R Core Team, *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2021.

[99] JASP Team, "JASP (Version 0.17.1)[Computer software]," 2023.

Supplementary Materials

Decoding construction: Searchlight analysis



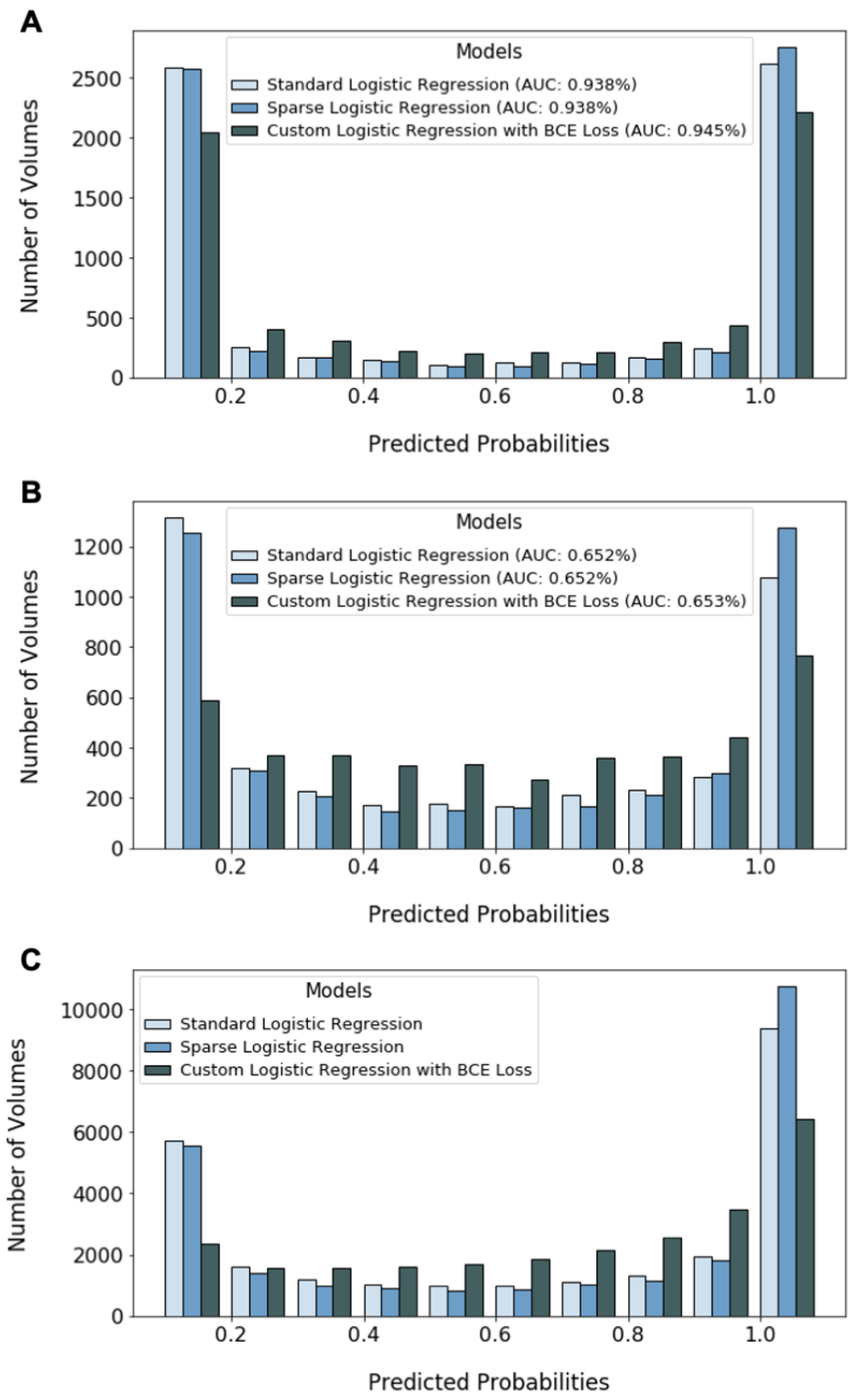
Supplementary Figure 1: *Decoder construction. Within-domain (perception) and cross-domain (from perception to imagery) searchlight decoding analyses were performed at the group level in standard space to verify the consistency of the generalization between perception and imagery within our region of interest in the fusiform.*

Models comparison

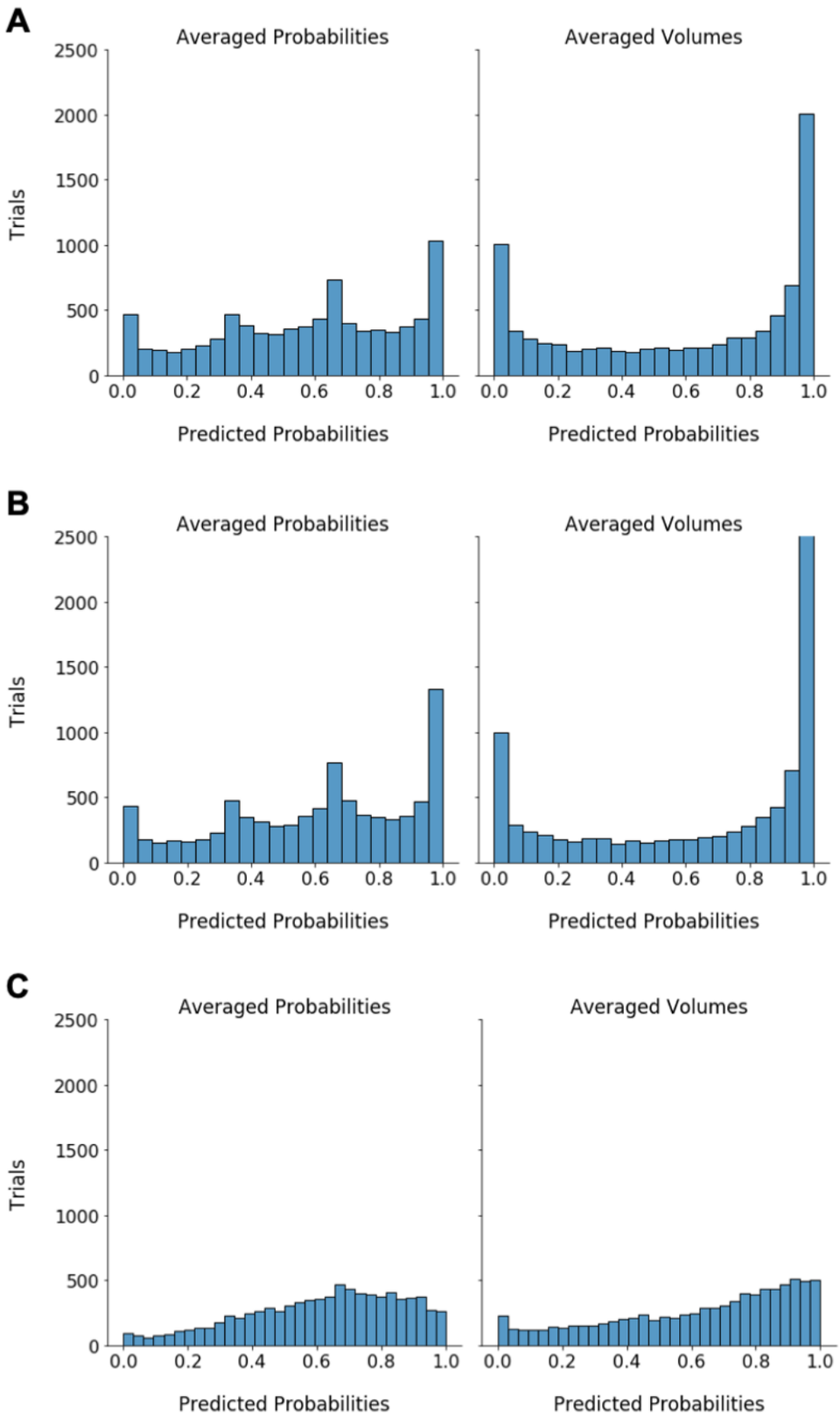
We assessed how well the custom logistic regression model with BCE Loss performed compared to the standard and sparse logistic regression commonly used in DecNef research [76, 96]. We evaluated the AUC and prediction probability distributions of the models under three scenarios across participants: perception data, cross-domain generalization from perception to mental imagery data, and cross-domain generalization from perception to DecNef data. In the perception condition, we implemented a Leave-One-Run-Out Cross-Validation approach to estimate performance of machine learning algorithms. Here the models were iteratively trained using data that corresponded to volumes within trials' HRF peaks from 3 perceptual runs, and were then tested on volumes of interest from the remaining run. For the standard logistic regression model, the maximum number of iterations was set to 5000. The L2 regularization penalty was applied to the logistic regression coefficients to prevent overfitting. The inverse of the regularization strength, known as C, was set to 1 to balance model complexity and fit to the data. Lastly, the lbfgs solver was used to optimize the logistic regression objective function, since it is well-suited for small to medium-sized datasets. For the sparse logistic regression model [76], we used the Sparse Multinomial Logistic Regression (SMLR) Python implementation [97] available through this link: https://github.com/KamitaniLab/smlr. For comparability purposes we used default parameters. To be specific, we set a maximum number of iterations of 1000 for model training and used 1e-5 as the tolerance value of the stopping criteria. Visual noise trials were not used neither for training-testing standard nor sparse logistic regression models. Finally, to maintain consistency with our experimental approach, we trained our custom logistic regression model with BCE Loss using the same parameters as mentioned in the Decoder training section. During the training process, the data was labeled with visual noise samples assigned to 0.5 during the training process. Training the model with Tensorflow required a validation set and accordingly we used a cross-validation approach in which the left-out run was divided into two halves to create a validation set and a test set. To obtain a more comprehensive assessment of the model's performance, we repeated this cross-validation procedure by swapping the validation and test sets, ensuring that all data was used as a test set. In the cross-domain generalization from perception to mental imagery data, the standard and sparse logistic regression models were trained using all data from perception sessions (again, excluding visual noise trials) and then tested over the HRF peak volumes from the two mental imagery runs. The training data for our model with BCE

loss consisted of volumes from 'Dogs', 'Scissors', and Visual noise trials from the first three perceptual runs. The validation set included 'Dogs' and 'Scissors' volumes from the 4th perceptual run, while the mental imagery trial's served as the test set. Lastly, in cross-domain generalization from perception to DecNef data, the standard and sparse logistic regression models were trained using 'Dogs' and 'Scissors' data from all perception runs and evaluated on all the HRF peak volumes within each DecNef training trial. To evaluate our custom logistic regression model, we utilized the pre-trained models used in our experiment for each participant. Since the ground truth is unknown in DecNef training, we compared the models by assessing the distribution of the predicted probabilities.

Aggregated predicted probabilities for the three models across subjects can be visualized in Figure S2. We used the Kolmogorov-Smirnov test to compare pairs of model distributions, which include standard vs. sparse logistic regression, standard vs. custom logistic regression, and sparse vs. custom logistic regression. In all scenarios, the model distributions were found to be significantly different from each other. In the perception condition (Standard vs. Sparse - KS statistic: .056; $p < .001$, Standard vs. Custom - KS statistic: .149, $p < .001$:, Sparse vs. Custom - KS statistic: .19, $p < .001$); from perception to mental imagery (Standard vs. Sparse - KS statistic: .053; $p < .001$, Standard vs. Custom - KS statistic: .175, $p < .001$:, Sparse vs. Custom - KS statistic: .161, $p < .001$); and from perception to neurofeedback (Standard vs. Sparse - KS statistic: .059; $p < .001$, Standard vs. Custom - KS statistic: .156, $p < .001$:, Sparse vs. Custom - KS statistic: .212, $p < .001$). The probability distributions generated by the standard and sparse logistic regression models exhibited higher overconfidence and greater similarity. These characteristics are particularly noticeable when examining cross-domain generalization scenarios. While maintaining performance, our customized model helped to mitigate the binomial overconfident predictions in all scenarios. To further tackle this issue of the binomial predictions, we computed the average of the predicted probability across each DecNef volume, instead of using the more conventional approach of averaging all DecNef volumes prior to decoding. Figure S3 compares the predicted probabilities for each DecNef trial obtained through both approaches for the standard, sparse and our custom logistic regression model with BCE Loss. The approach used here also seemed to help to moderate the overconfidence in the classifier, thereby providing participants with finer-grained and more informative feedback.



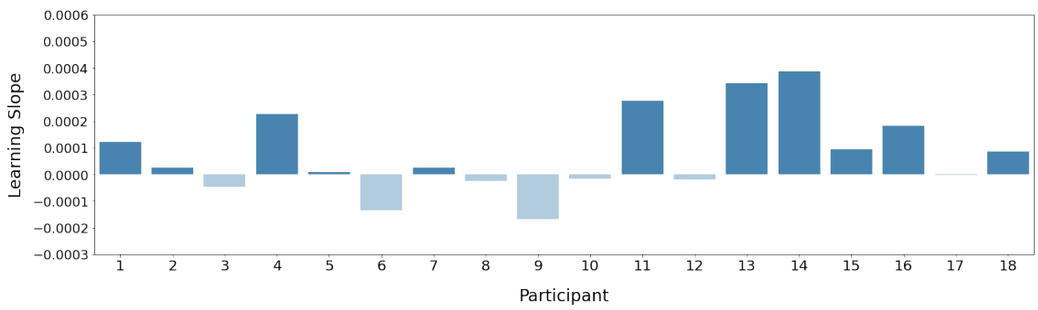
Supplementary Figure 2: *Histogram depicting the frequency (i.e. number of volumes across participants) of the predicted probabilities across the different models, namely, standard logistic regression, sparse logistic regression, and our custom logistic regression model with Binary Cross-Entropy (BCE) Loss across the different classification problems, namely: (A) Perception: aggregated probabilities across the four cross-validation folds. (B) Visual imagery: aggregated cross-domain generalization probabilities based on a classifier trained on perception. (C) Aggregated cross-domain generalization probabilities during DecNef based on the decoder trained on perception and validated on imagery.*



Supplementary Figure 3: *Histogram depicting the frequency (i.e. number of trials across participants) of the predicted probabilities across the different models and averaging procedures. (A) Standard logistic regression. (B) Sparse logistic regression. (C) Our custom logistic regression model with BCE Loss. 'Averaged Probabilities' (on the left) shows the results when the volumes within each trial are decoded individually and the different probabilities are averaged; 'Averaged Volumes' (on the right) shows the results when the different volumes within a trial are first averaged, and then the probability of decoding the target class is computed.*

Individual differences in learning to control brain activation patterns during DecNef training

To provide further examination of the observed global learning effects, we investigated individual differences in this learning ability. We created an index of representational control for each participant by means of a linear regression analysis in R [98]. Specifically, we used the decoding probability of the target on each trial as the dependent variable and the fMRI trial index along the DecNef training (as numeric variable) to estimate the beta coefficients. The beta coefficient for the independent variable in the linear regression model represents the rate of change in the decoding probability as a function of the trial index, which can be interpreted as a measure of the participants' learning progress over the course of the DecNef training. More than half of the participants (11 out of 18) exhibited positive learning effects, as indicated by their beta values being greater than 0 (Figure S4). This suggests that DecNef training had a positive impact on the learning process for a majority of the participants. A one-sample t-test showed that the beta values across participants were significantly greater than zero (t(17) = 2.076, p = .026), providing further evidence of the learning effect.



Supplementary Figure 4: *Individual differences in learning effects.*

Analysis of the error rate of the search responses

We also analyzed the number of trials in which the participants in the DecNef group failed to identify the search target, to determine if there were any variations in accuracy due to the hiragana and search target factors. To do this, we performed a repeated measures ANOVA analysis in JASP [99] using participants' error rates for each condition. No significant effects of DecNef were observed on the error rate of the search responses, namely for the hiragana (F(1, 17) = .029, p = .868; η_p^2 = .002) or the search target factor (F(1, 17) = .022, p = .883; η_p^2 = .001). There was no significant interaction between the two factors (F(1, 17) = .471, p = .502, η_p^2 = .027). A separate paired t-test was conducted in the control group to examine the impact of the Search Target factor on the error rate, but no significant differences were found (t = -0.695, p = .492). Finally, to compare the error rates between the two experimental groups, another ANOVA was conducted with search target as a within-subjects factor and Experimental Group as a between-subjects factor. No significant effects were observed for the search target (F(1, 52) = .265, p = .609, η_p^2 = .005), Experimental Group (F(1, 52) = 1.972, p = .166, η_p^2 = .037), or the interaction between the two factors (F(1, 52) = .061, p = .805, η_p^2 = .001).

Participants subjective reports following the DecNef and the visual search fMRI experiments

Participants provided subjective reports of the level of effort required to increase the size of the disk during, and the perceived level of control over the neurofeedback signal. Two distinct Likert scales were

used, with values ranging from 1 to 10, where the lowest score represented *"very little effort/control"* and the highest score represented *"a great deal of effort/control"*. The responses were grouped into quintiles, which were determined based on the possible range of values on the scale. The quintiles were defined as follows: Quintile 1 (Q1): 1-2; Quintile 2 (Q2): 3-4; Quintile 3 (Q3): 5-6; Quintile 4 (Q4): 7-8; Quintile 5 (Q5): 9-10. From the distribution of responses concerning perceived effort, the majority of participants (77.8%) reported medium to high levels of effort (Quintile 3 or Quintile 4). A smaller proportion of participants (11.1%) reported low effort, falling in Quintile 2, or high effort, falling in Quintile 5. Regarding the perceived control on the feedback signal, the distribution of responses revealed that most of participants (55.6%) felt they had a high level of control over the feedback during the DecNef training sessions, with the majority of responses falling in Quintile 4. A smaller proportion of participants (44.4%) reported moderate levels of control, with all responses falling in Quintile 3. No responses were recorded in either Quintile 1 or Quintile 5. To evaluate whether participants were aware of the experimental purpose of the DecNef training, we queried them on two occasions. The first was immediately after the completion of the DecNef training and before the visual search task, and the second was after completing the visual search task. The reports indicated that participants were unaware of the specific research goals. Many participants believed that the study was oriented towards investigating attentional control, stress, or motivation, rather than the manipulation of conceptual representations. Upon completion of the visual search task, some participants realized that one of the two hiragana symbols used in the DecNef training sessions had also appeared during the visual search task. Nevertheless, the participants remained unaware of the specific experimental aims of the study. Additional information is available in Table S2. Finally, we examined whether participants could correctly identify the stimulus class that was the target of the DecNef training sessions. Out of the 18 participants, 12 were able to correctly report the 'Dogs' category. However, a multinomial chi-squared test showed that the observed frequencies were consistent with a uniform distribution ($\chi2(1) = 2$, $p = .157$), indicating that participants were not able to detect the ground truth. We also evaluated the participants' confidence in their responses using a 4-point Likert scale, on which 0 indicated "Guess", 1 indicated "Little confidence", 2 indicated "Fairly confident" and 3 indicated "Sure". Overall, the distribution of confidence reports revealed that the majority of participants reported very low levels of confidence regarding their ability to detect the target category, with only a small percentage reporting seemingly high levels of confidence. Specifically, 6 participants (33.3%) reported that they were guessing, 7 participants (38.9%) reported little confidence, 2 participants (11.1%) reported feeling fairly confident, and 3 participants (16.7%) reported feeling sure. Importantly, among the 12 participants who correctly identified the subliminally presented 'Dogs' category, 4 participants (33.3%) reported that they were guessing , 4 participants (33.3%) reported little confidence, 2 participants (16.7%) reported feeling fairly confidence, and 2 participants (16.7%) reported feeling sure. This indicates that even among those who correctly detected the subliminal category, there was still some uncertainty about their response. However, some caution needs to be taken with interpreting these reports. For instance, higher levels of confidence could be associated with subjective preferences for dogs relative to scissors, as some participants reported.

Supplementary Table 1: ***Participants' reports about the strategies during DecNef training.***

Participant	Strategies
1	Mental humming of music lyrics worked fine.
2	Action thoughts (i.e. karate) and food visualization. Best strategy seemed to imagine movements.
3	Mental simulation of people, social events and situations. I thought the circle increased its size as a function of excitement and emotion in memories.
4	Mathematical calculations (polynomials); crosswords; playing *Tetris* and *Connect four* mentally. I thought it worked better when mentally listening to classical music.
5	Face imagery, remembering things and self talk. I thought it worked best when visualizing and attending to movie characters by remembering their names.
6	Mental painting; mathematical calculations; mental humming; self-talking in Italian. I think maths was the most useful strategy.
7	Simulating strong positive and negative emotions. Negative and vivid emotions seemed to work the best, although happy travels worked at the beginning but then lost effectiveness.
8	Thinking about smiley faces, clowns, girlfriend and family, nasty and rotten things, mutilations. The strategy that seemed to work the best was imagining a comedian dancing with background music.
9	Imagery of squeezing the circle, extending it manually, and filling it with red. I often had to change her strategy as it stopped working and varied across different days.
10	Simulation of positive emotions and euphoric states, playing sports, surfing, dancing, situations with friends. The auditory modality was not always important and the visual modality seemed to work better. Not thinking about anything seemed to increased the size of the circle sometimes. The strategies tended to stop working if they were used continuously and I had to keep changing the strategies.
11	Audiovisual replay of TV shows. Music did not appear to work as well as visual thoughts. Then, I started imagining smiling faces and feeling strong emotions. When I focused on faces, the higher the vividness, the larger the circle was. Also, strong emotions also triggered and increase in the size of the circle.
12	Thinking about daily tasks, such as the paper I was writing or packaging things. I also thought about songs and books and naming animals but had to changed the strategies a lot they stopped working the more I used them.
13	Imagining a balloon, happy things and faces, having superpowers, retrieving words that start with specific letters, playing mentally with Playdough. The best strategy seemed to imagining basketballs with high detail.
14	Counting, thinking about statistics, dancing or situations with my cat, or what I would do next. The strategies that worked better involved movement such as dancing or running.
15	Simulating feelings and emotions, putting on clothes, cold and heat sensations, imagery of animals, flowers and situations. Imagining things worked better. The best strategies were simulating emotions, putting on socks, and imagining colors.
16	Simulating positive and negative emotions, feelings related to playing hockey, thinking about travels and situations with family. The best strategies involved hockey related feelings, which seemed to work better than visual memories. When I repeated the same strategy, it stopped working.
17	Mental humming, random memories, emotional thoughts, hypothetical situations. Mental humming, mind-wandering and stopping conscious thought were the best strategies.
18	Simulation of activities which implied movement such as sports, mental planning and hypothetical situations, vividly remembering old dreams and imagining alternative endings for known movies. The best strategies involved questioning the why of the things and not thinking about anything.

Supplementary Table 2: ***Participants' beliefs about the experimental aims before and after completing the behavioural search task.***

Participant	**Reported experimental aim after DecNef training**	**Reported experimental aim after the behavioural evaluation task (i.e., the visual search task).**
1	No idea.	Attentional control and stress.
2	Something related to mental visualization	No idea.
3	No idea. At the beginning I thought the aim was related to decoding one's mental processes, but then she discarded that idea.	Control of brain activity related to the symbols.
4	Mental inhibition, cognitive flexibility or trying to make a specific brain region more active.	I recognized that symbols were from DecNef training sessions and dogs and scissors belong to a previous experimental session, but no idea of connection between both.
5	No idea.	No idea.
6	Test whether I could learn to control the size of the circle across sessions.	Something to do with imagery and control of brain activation .
7	Understand attentional control and mental imagery vividness.	Something related with the new hiragana that was not presented before.
8	Investigate neurofeedback. At the beginning I thought perceptual sessions and neurofeedback training sessions were related, but then I discarded this possibility.	Investigating how different semantic categories are processed.
9	No idea.	I thought the aim was to examine reaction times and how we get used to the presented images.
10	Examining activation in a brain region. Something related to emotions, relaxation and excitement states. Also related to motivational changes.	To examine whether the symbol associated with happiness improves attention in the visual search task.
11	Examining attentional control and focus.	No idea.
12	No idea. Experiment was related to anxiety, stress or attention.	No idea.
13	How the brain reacts if you get disappointed when trying to make more money.	Examination of decision reaction times.
14	No idea.	No idea.
15	A study related to retrieval of old memories. The experiment implied increasing brain activation at a specific brain region.	No idea.
16	Increasing the activation of specific brain regions.	No idea.
17	Training the brain to create a specific a brain activity pattern. Increasing the strength of thoughts.	Attentional control and concentration.
18	Examining attentional control over a specific brain region.	Something related with the symbols and their impact on search reaction times.