Sculpting New Visual Categories into the Human Brain

Marius Cătălin Iordan[1]*, Victoria J. H. Ritvo[2], Kenneth A. Norman[2], Nicholas B. Turk-Browne[3], Jonathan D. Cohen[2]

[1]Brain and Cognitive Sciences Department & Neuroscience Department, University of Rochester, Rochester, NY 14627, USA

[2]Princeton Neuroscience Institute & Department of Psychology, Princeton University, Princeton, NJ 08544, USA

[3]Department of Psychology, Yale University, New Haven, CT 06520, USA

***Corresponding Author**

Marius Cătălin Iordan, Ph.D.

Meliora Hall, University of Rochester

Rochester, NY 14627 USA

Email: mci@rochester.edu

Twitter: @MCatalinIordan

Abstract

Learning requires changing the brain. This typically occurs through experience, study, or instruction. We report a proof-of-concept for a new way for humans to acquire visual knowledge by directly sculpting activity patterns in the human brain that mirror those expected to arise through learning. We used a non-invasive technique (closed-loop real-time functional magnetic resonance imaging neurofeedback) to create new categories of visual objects in the brain, without the participants' explicit awareness. After neural sculpting, participants exhibited behavioral and neural biases for the sculpted, but not for the control categories. The ability to sculpt new perceptual distinctions in the human brain offers a new paradigm for human fMRI research that allows for non-invasive, causal testing of the link between neural representations and behavior. As such, beyond its current application to perception, our work potentially has broad relevance to other domains of cognition such as decision-making, memory, and motor control.

Introduction

"For if someone were to mold a horse [from clay], it would be reasonable for us on seeing this to say that this previously did not exist but now does exist."

Mnesarchus of Athens, ca. 100 BCE (*1*)

Humans continuously learn through experience, both implicitly (e.g., through statistical learning; *2,3*) and explicitly (e.g., through instruction; *4,5*). Brain imaging has provided insight into the neural correlates of acquiring new knowledge (*6*) and learning new skills (*7*). As humans learn to group distinct items into a novel category, neural patterns of activity for those items become more similar to one another and, simultaneously, more distinct from patterns of other categories (*8-10*). We hypothesized that we could leverage this process using neurofeedback to devise a fundamentally new way for humans to acquire perceptual knowledge. Specifically, sculpting patterns of activity in the human brain ('molding the neural clay') that mirror those expected to arise through learning of new visual categories may lead to enhanced perception of the sculpted categories ('they now exist'), relative to similar, control categories that were not sculpted. To test this hypothesis, we implemented a closed-loop system for neurofeedback manipulation (*11-18*) using functional magnetic resonance imaging (fMRI) measurements recorded from the human brain in real time (every 2s) and used this method to create new neural categories for complex visual objects. Crucially, in contrast to prior neurofeedback studies that focused exclusively on reinforcing or suppressing existing neural representations (*11,12*), in the present work we sought to use neurofeedback to create novel categories of objects that previously did not exist in the brain; we test whether this process can be used to generate significant changes in the neural representations of complex stimuli in human cortex, and, as a result, alter perception.

Results

<u>Constructing a Normed Stimulus Set of Complex Objects for Neural Sculpting</u>

Using radial frequency components (RFCs) of an image (*19,20*), we generated a two-dimensional manifold of closed-contour shape stimuli that varied smoothly in appearance as a function of distance from a center shape (Fig. 1A). To obtain each shape, sine waves determined by the seven RFCs were added together and the resulting wave was wrapped around a circle to obtain a closed contour which was then filled in to create a shape. We chose this technique for building a complex visual stimulus space because prior work has shown that radial shapes generated using subsets of the same RFCs are perceived monotonically and that their neural representation is also monotonically related to parametric changes in the amplitude of the RFCs across multiple brain regions (*19,21*). We created a two-dimensional shape space by independently varying the amplitude of two of the seven RFCs (from 12.6 to 36.6 for the 4.94 Hz component and from -6.0 to +42.0 for the 1.11 Hz component), while holding the amplitudes of all other components constant. In this manifold, each of six equally spaced diameters (AG, BH, CI, DJ, EK, FL) defined a novel and arbitrary category boundary between the two groups of shapes on either side. We verified that the shapes were perceived similarly across categories in a psychophysical experiment with 10 participants using a self-paced two-alternative-forced-choice (2AFC) task. For each diameter through the space, we presented shapes along the diameter and participants judged which of the two "endpoint" shapes (e.g., A vs. G) more closely resembled the presented shape. We then used these categorization data to compute a psychometric function (Fig. 1B-G).

Participants did not have *a priori* biases in how they categorized items across multiple partitions (psychometric function slope, repeated measures ANOVA, factor='direction'; F(5,54)=0.29, p=0.918).

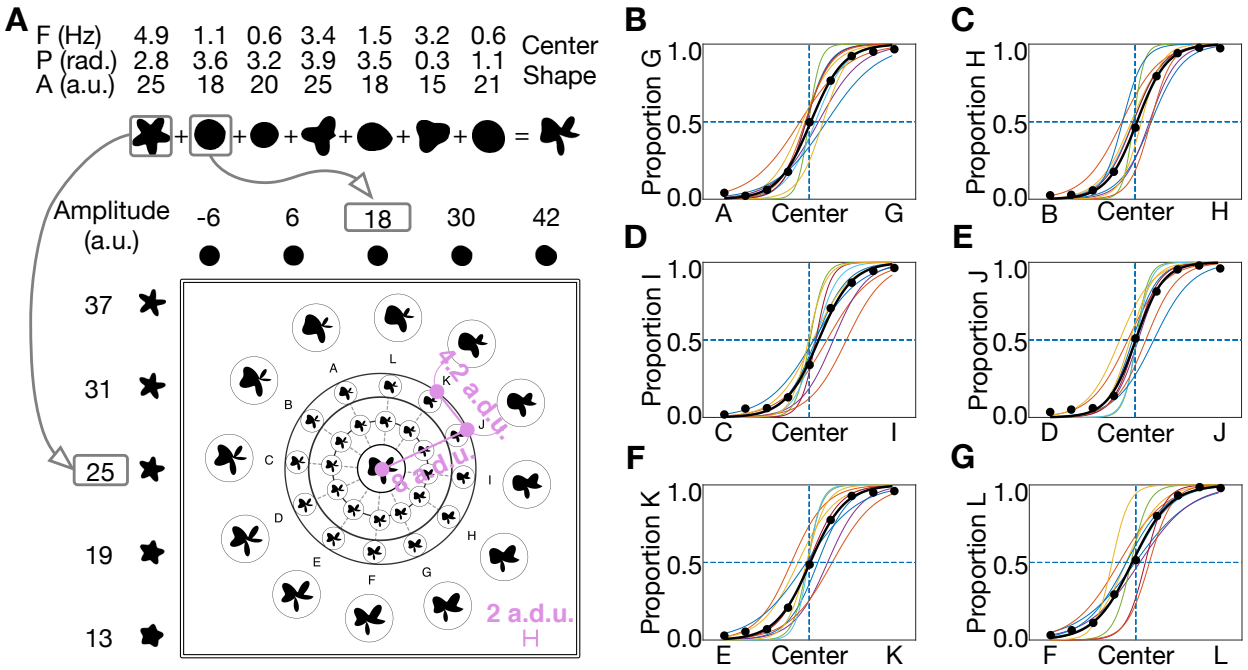


Fig. 1. Shape space stimulus set construction and perceptual norming. (A) Stimuli were constructed using combinations of seven radial frequency components (RFCs) of fixed frequency F and phase P (*19,20*). We parametrically varied the amplitude A of two of the seven RFCs independently to generate a two-dimensional manifold of complex shapes: a displacement of 1 arbitrary distance unit (a.d.u.) on the manifold corresponded to an amplitude change of 6 arbitrary units for the 1.11 Hz component and an amplitude change of 3 arbitrary units for the 4.94 Hz component (changes in component appearance shown above and to the right of the shape space). Each point on this manifold corresponded to a unique shape that could be computed using the mathematical convention described above. Example shapes at 8 a.d.u. from the center are enlarged and shown on the outer edge (e.g., A, B, etc.); they lie 4.2 a.d.u. apart from each other on the manifold. (B-G) Psychometric function estimates for the 2AFC behavioral experiment conducted for each of six diameters: AG, BH, CI, DJ, EK, and FL. Thin lines in distinct colors represent psychometric functions for each participant (n=10) and thick black lines represent participant averages for each diameter.

Using this set of complex visual objects, we trained a new group of 10 participants over 10 days each in a real-time fMRI experiment. The participant sample size was planned during the experiment design and is comparable with (or exceeds) that of most other neurofeedback studies (*11-18*) that target primarily visual regions in human cortex. Notably, our 10 participants correspond to 98 total sessions (20 behavioral, 78 in-scanner, 2h each), providing considerably

denser sampling and more data than conventional fMRI or other similar fMRI neurofeedback studies. We used a closed-loop neurofeedback procedure to sculpt a categorical distinction between the objects on either side of one diameter through shape space. By sculpting divided neural representations for these objects, we sought to change how they are perceived (Fig. 2A). On Day 1, participants performed a 2AFC behavioral task (identical to the one used to norm the stimulus space) to obtain a baseline measure (psychometric function slope) of how they categorized stimuli along each diameter in the space. We verified that the training cohort of participants did not have any biases across categorization directions (replicating the norming cohort results) and we also verified that there were no significant differences between the training and norming cohorts (psychometric function slope, repeated measures ANOVA, factors='direction', 'cohort'; interaction: $F(5,108)=0.19$, $p=0.964$; main effect of direction: $F(5,108)=0.57$, $p=0.722$; main effect of cohort: $F(1,108)=1.71$, $p=0.194$) (Fig. S1). On Days 2-3, we ran two fMRI sessions in which we mapped how each individual participant's brain represented the stimulus space, defined candidate brain regions for neurofeedback, and built a model of how the shape space was represented in participants' brains to use for real-time tracking during the training phase of the experiment. After Day 3, we randomly selected one of the diameters as the category boundary to be sculpted during training for each participant (not disclosed to them). On Days 4-9, participants underwent real-time fMRI neurofeedback training, during which they were shown wobbling versions of the shapes (Movie S1) and asked to generate a mental state that would stabilize the shape in their visual display. Unbeknownst to the participants, they were given positive visual feedback (reduced wobbling) and monetary rewards when the neural representation of the shape they were viewing resembled the neural representations of other shapes in the category selected to be sculpted during training. Finally, on Day 10, participants repeated a version of the 2AFC behavioral task, in which we evaluated changes in categorical perception across both sculpted and control category boundaries.

Using the Shape Space to Sculpt New Neural Categories of Objects in Multiple Brain Regions

We defined a region of interest (ROI) to target with neurofeedback by running two block-design fMRI localizer sessions (Days 2-3) in which we showed participants 12-15 repetitions of 49 shapes spanning the two-dimensional stimulus manifold. We constructed an ideal representational similarity matrix of a subset of 25 of these shapes (Fig. 2B) and performed a cortical searchlight analysis (*22*) to find all brain regions that represented the shape space parametrically (i.e., akin to a cognitive map (*23,24*), Fig. 2C,D). We defined the target neurofeedback ROI for each participant as the union of all their parametric representation voxels, excluding early visual cortex (since we were interested in high-level visual perception related to complex object categories, the pericalcarine region was defined anatomically using Freesurfer; (*25*)). The final neurofeedback ROIs comprised 780-2,209 voxels per participant (average 1,401±115 s.e.m. voxels; Figs. 2C & S2). A representative ROI is shown in Fig. 2D, all participants' final ROIs are shown in Figs. S3—S12, the unions of all parametric voxels for each participant (including early visual cortex) are shown in Figs. S24—S33, and the group map for the average neurofeedback ROI across all participants after alignment to common MNI space is shown in Fig. S34. We hypothesized that if any portion of these ROIs (either individually or in concert) were causally related to the categorical perception of our stimuli, then sculpting novel neural categories across all of them collectively would maximize the chances of influencing participants' perception after training.

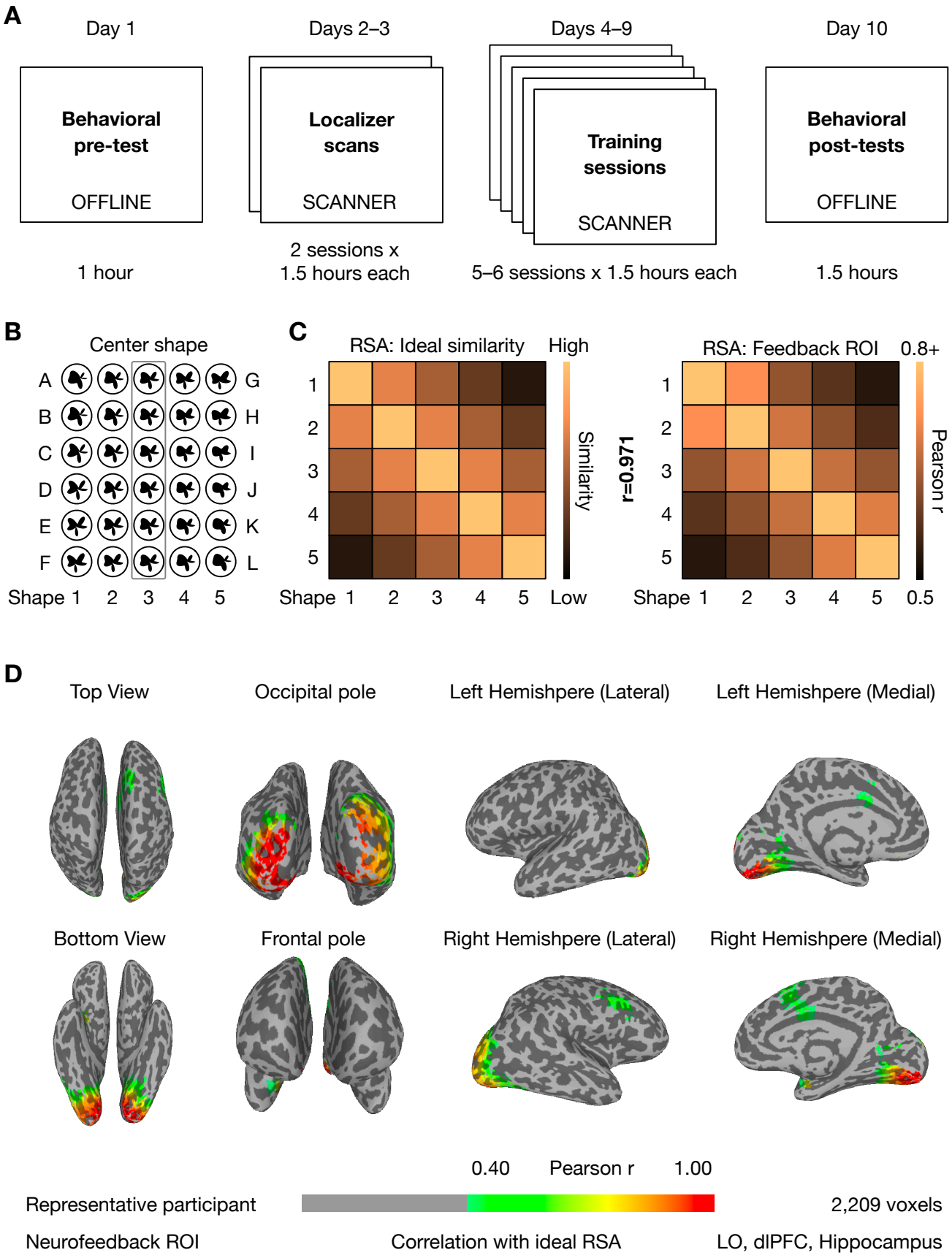

A
Day 1
Days 2–3
Days 4–9
Day 10
Behavioral pre-test
OFFLINE
Localizer scans
SCANNER
Training sessions
SCANNER
Behavioral post-tests
OFFLINE
1 hour
2 sessions x 1.5 hours each
5–6 sessions x 1.5 hours each
1.5 hours
B
Center shape
A B C D E F
G H I J K L
Shape 1 2 3 4 5
C
RSA: Ideal similarity
High
Low
Similarity
r=0.971
RSA: Feedback ROI
0.8+
0.5
Pearson r
Shape 1 2 3 4 5
D
Top View
Occipital pole
Left Hemishpere (Lateral)
Left Hemishpere (Medial)
Bottom View
Frontal pole
Right Hemishpere (Lateral)
Right Hemishpere (Medial)
0.40
Pearson r
1.00
Representative participant
2,209 voxels
Neurofeedback ROI
Correlation with ideal RSA
LO, dlPFC, Hippocampus

Fig. 2. Experimental procedure and neurofeedback ROI computation. (A) The experiment extended over 9-10 days for each participant: one behavioral pre-test (2AFC along each diameter from Fig. 1), two localizer fMRI sessions, 5-6 real-time fMRI neurofeedback training sessions (5 days: n=2, 6 days: n=8), and a final behavioral post-test (repeat of first day). (B) 25 shapes shown during the localizer scans were used to define parametric neural representations of the shape space: 5 shapes along each diameter, -8, -4, 0, +4, +8 a.d.u. from the center shape common to all diameters. (C) Ideal representational similarity matrix (RSM) for a neural parametric representation (left) and average RSM for the 25 shapes in the final neurofeedback region of interest (ROI) of the training cohort (n=10, right). The correlation between the ideal and observed RSMs was very high (r=0.971) for the selected neurofeedback ROI across all participants, suggesting strong parametric representation of the shape space in these brain regions. (D) Cortical map of the neurofeedback ROI for an example participant encompassing several brain regions, including extrastriate visual cortex, parahippocampal gyrus, hippocampus, and medial frontal gyrus (see Methods for details of the selection procedure). Cortical maps for all participants (n=10) are shown in Figs. S3-S12.

Using this ROI and the fMRI data from Days 2-3, we selected an arbitrary diameter in the stimulus space as a category boundary for each participant by rolling a pink six-sided die (Fig. 3A). Participants were not informed that this was a study about visual categories, that the shapes were drawn from a continuous circular space bisected by diameters, nor about which diameter was randomly selected as the boundary. We built a model of the neural representations of the two resulting categories of shapes. Each category was modeled as a multivariate Gaussian distribution with shared covariance (Fig. 3B). The parameters of the category-specific distributions were computed using maximum-likelihood estimation for the top 100-200 principal components of the signal elicited by that category's shapes in the neurofeedback ROI (we used a grid search to determine the optimal number of components and hemodynamic lag for each participant). To verify that our model could predict the distinction between shape categories in the brain, we inverted it into a discriminative log-likelihood-ratio-based pattern classifier. Leave-one-day-out decoding accuracy was high (chance = 50%): 73%-81% in the lateral occipital region (defined using Freesurfer; (*25*)), known to represent distinctions between closed contour objects (*26*), and 71%-79% in our neurofeedback ROI (Fig. 3C). Given that our ability to induce multivariate pattern-level neural plasticity using neurofeedback is predicated on the highly accurate decoding of shape category during training, this measure was also used as a selection criterion for inviting participants back for training after Day 3 (criterion: >70% decoding accuracy for both LO and neurofeedback ROI; 7 participants excluded; Table S1). This criterion also ensured that our participants rated highly on a strong predictor of the effectiveness of neurofeedback manipulation: neural signal strength as measured by baseline task-related decoding accuracy (*34*).

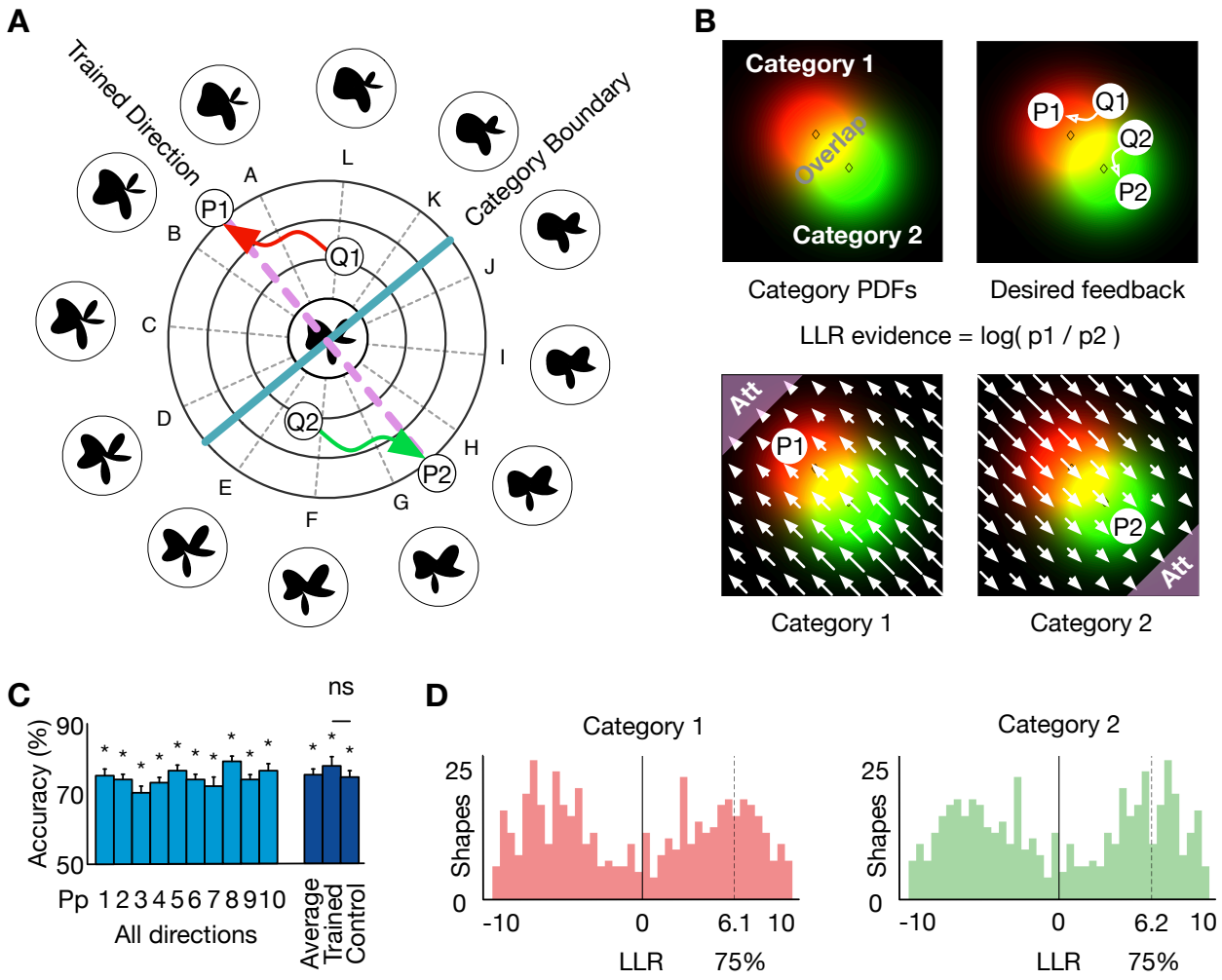


Fig. 3. Using a neural model of the shape space to guide neurofeedback. (A) Unbeknownst to participants, we chose an arbitrary category boundary (blue) and sought to sculpt separate neural representations of shape categories along the perpendicular direction (purple). For example, the neural representations of shapes Q1 and Q2 were sculpted to be increasingly distant from the category boundary, illustrated by P1 and P2, respectively. (B) Two-dimensional schematic of neural representations of shape categories (the actual neural space comprised 100-200 dimensions). We sculpted the neural representations of shapes to become more similar to the corresponding attractor regions (Att) for categories 1 and 2. Arrows indicate relative change in log-likelihood ratio (LLR) required to reach the goal. (C) Gaussian model accurately predicted shape category in all participants during localizer scans (Days 2-3, leave-one-day-out decoding 75.7%±0.854% s.e.m.; chance=50%, *p<0.001). We observed no significant differences between decoding accuracy for the trained (sculpted) and control categories (trained: 78.1%±1.46% s.e.m.; control: 75.4%±1.38% s.e.m.; trained > control $t(9)=1.12$, $p=0.293$). The graph shows only participants who exceeded the 70% threshold and subsequently underwent training (also see Table S1). (D) Graphs show histogram of LLR values for neural representations elicited by all shapes presented during the localizer scans in the neurofeedback ROI for an example participant (LLR was computed using the cross-validated, optimal Gaussian model used for neurofeedback

training). Feedback was given if neural evidence for the category of the shape on the screen (e.g., Q1) exceeded a given threshold in the distribution of LLR values for all shapes shown during the localizer scans. We note that the 75% threshold used for training may be different for each category under the same model. LLR histograms with 75% thresholds for all participants are shown in Fig. S13.

During neurofeedback training (Days 4-9), participants were told that during each trial they would see a continuously wobbling shape and had to "Generate a mental state that will make the shape wobble less or even stop!" Participants received additional instructions designed to accommodate for the hemodynamic lag inherent to the fMRI signal (e.g., that progress on the task is related to their mental state over the previous 8-10 seconds) and were incentivized with a monetary reward for each trial in which they made progress. However, participants were never told that a category boundary existed in the shape space or that feedback was based on the shape of the object, no less its relationship to the selected category boundary (see Methods for details). For each neurofeedback trial, we selected a random shape (seed) drawn from one of the two categories and generated a continuous oscillation in parameter space centered at that shape's coordinates. Visually, participants saw shapes in the center of the screen (center of mass remained stationary) that morphed smoothly and gradually between the seed shape and other nearby, similar shapes (Movie S1). The apparent magnitude of the shape morph on the screen was manipulated via neurofeedback at every fMRI timepoint (TR=2s). For each individual participant, we computed the distribution of LLR values under our estimated Gaussian distributions for all shapes shown their localizer scans (520—800 trials, Fig. 3D) and positive feedback (less oscillation) was given during training if the neural evidence for the category of the current shape exceeded a given threshold in this distribution (Figs. 3D, S13, & S14, Movie S2). The initial threshold was selected based on pilot scans indicating that the neural model's decoding performance decayed as a function of wobble intensity (Fig. S15). The threshold was continually adjusted during the experiment using an adaptive procedure designed to provide feedback on approximately 33% of the trials (Table S2). Each participant (n=10) completed 520–800 trials of closed-loop real-time neurofeedback training over the course of 5-6 daily sessions. A summary of feedback thresholds and performance for each participant, training session, and run is given Table S3.

Neural Sculpting Generated New Visual Categories in the Brain and in Perception

After finishing the training sessions, we evaluated whether neural sculpting of the novel visual categories was successful; that is, whether the closed-loop real-time neurofeedback procedure induced multivariate pattern-level neural plasticity of the shape space representation. To measure changes in the neural representations of shape categories, we computed the difference between the average LLR of each trial (the log-likelihood ratio of the neural representation elicited by each shape in the neurofeedback ROI under the Gaussian model trained to differentiate the two trained categories of shapes) during the first two days of training versus the last two days of training (since participants underwent 5 or 6 days of training, there was no clearly defined first/second half of the experiment for everyone). We compared this score between the "trained" shape categories that we expected to have separated in the brain because they were bisected by the selected diameter, versus the "control" shape categories that would have been created by the perpendicular diameter, for which no separation was expected (Fig. 4A). We found strong positive

neural sculpting effects for 6 out of 10 participants (between +0.342 and +1.204) and weak negative effects for 4 out of 10 participants (between -0.212 and -0.067) (Fig. 4D). From the first two days to last two days of training, we found a reliable difference in the LLR change between trained and control categories (Fig. 4B-D; +0.369±0.154 s.e.m; t(9)=2.27, p=0.049, Cohen's d = 1.421). Within this interaction, LLR showed an increase for the trained categories (Fig. 4B; +0.234±0.077 s.e.m., t(9)=2.87, p=0.019) and a numeric, but not significant decrease for the control categories (Fig. 4B; -0.136±0.087 s.e.m., t(9)=1.48, p=0.174). Effects of neural sculpting on neural representations for each individual participant are shown in Fig. S16A. These results were obtained using the LLR for the first timepoint of each neurofeedback trial to avoid potential repetition-suppression effects (*27*). However, similar, significant results were obtained for LLR from up to the five timepoints from each trial during which the participants potentially received feedback (Table S4), albeit with smaller effect sizes likely due to such repetition-suppression effects. Similar results were also observed using a ratio of the LLR for trained and control categories, instead of a difference score (Fig. S17). Together, these results suggest that our real-time fMRI neurofeedback neural sculpting procedure was successful in manipulating how the human brain represents complex objects from the multidimensional shape space we constructed.

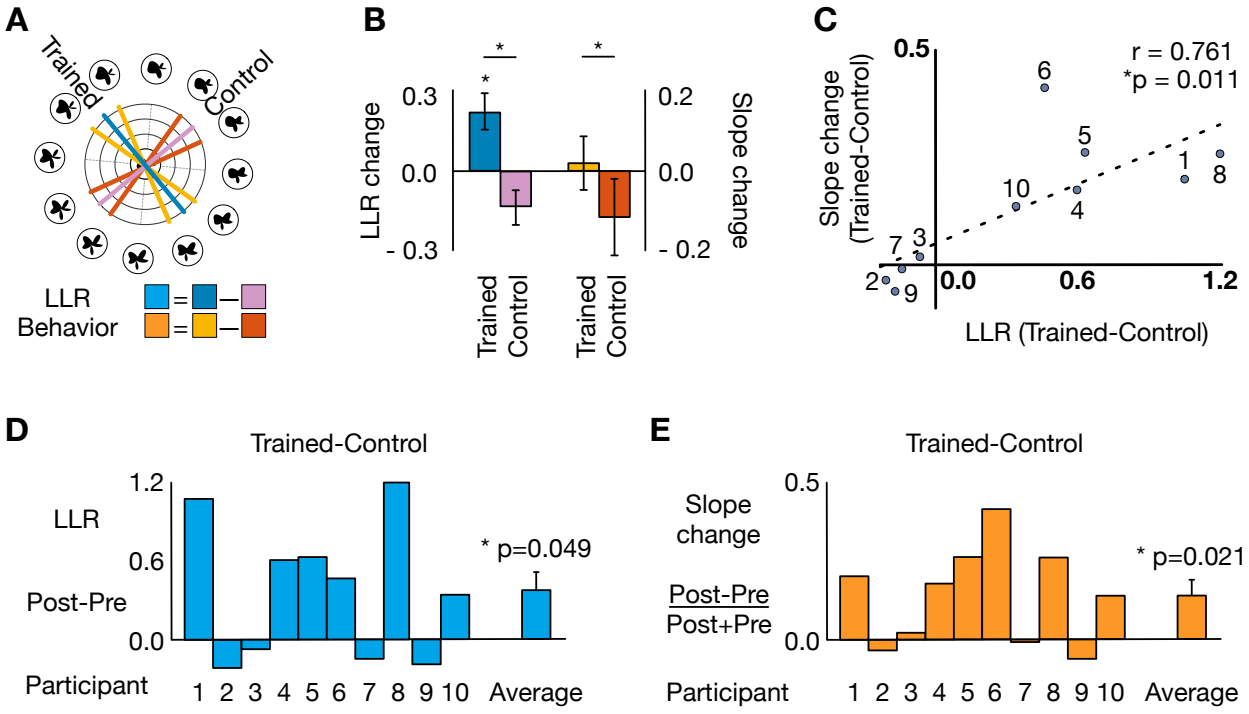


Fig. 4. Neurofeedback successfully sculpted new visual categories in the human brain and altered perception. (A) Shape space with example category boundaries. Diameters for trained category distinction: LLR = blue, psychometric function slope = average of yellow lines. Diameters for control category distinction: LLR = purple, psychometric function slope = average of red lines. (B) Effects of neural sculpting on neural representations and perception for trained and control categories: differences in LLR and psychometric function slopes (colors as in subpanel A; *p<0.05). (C) Changes in the brain due to neural sculpting predict perceptual changes. (D)

Change in LLR between the last two days and the first two days of training for individual participants. Positive values indicate stronger neural boundaries in trained vs. control categories. (E) Change in psychometric function slope for trained vs. control categories between behavioral pre- and post-tests for individual participants. Positive values indicate stronger categorical perception for trained versus control categories.

Our goal was not only to modify the brain but also to test the hypothesis that neural sculpting is sufficient to alter behavior. Namely, we predicted that training would induce categorical perception whereby shapes close to the category boundary come to be perceived as clearer category members. Thus, we hypothesized that the slope of the psychometric function running perpendicular to the sculpted boundary would become steeper, compared to the psychometric function running alongside the sculpted boundary. To measure these perceptual changes, we estimated psychometric functions for the trained and control categories from the 2AFC behavioral task conducted on Day 10. Most participants reported fatigue with the experiment (after 5-6 consecutive days of fMRI scans) and we observed a tendency for some participants to perform worse at categorization across all directions in the post-test compared to the pre-test (Table S5 & Fig. S16B), as well as higher variability overall for psychometric function slopes across all directions in the post-test compared to the pre-test (average slope pre: 1.85±0.071 s.e.m.; average slope post: 1.98±0.339 s.e.m.; Table S5 & Fig. S23). We predicted that neural sculpting was successful at altering behavior for all participants, even if fatigue potentially decreased the slope of psychometric functions across all directions for some participants. As such, we would observe a perceptual advantage for the trained direction after training, irrespective of overall categorization performance; that is, psychometric function slopes for all participants would relatively increase for the trained direction compared to the control direction, between the pre- and post-tests. To measure the relative change for the trained and control directions in the pre- and post-test measures across all participants, we calculated a normalized difference score between the psychometric function slopes from Day 10 and those on Day 1 (Fig. 4A). The discrimination slope significantly increased for the trained categories compared to the control categories (Fig. 4B,E; +0.137±0.049 s.e.m., $t(9)=2.79$, $p=0.021$, Cohen's $d = 0.505$). We observed strong positive behavioral effects for 6 out of 10 participants (0.137-0.414), a weak positive effect for one participant (0.021) and weak negative effects for 3 out of 10 participants (-0.062, -0.033, -0.008). Estimated psychometric functions for all tests, directions, and participants are shown in Fig. S16B & Table S5. Overall, across our entire cohort, positive effects of neurofeedback manipulation (both behavioral and neural) tended to be substantial, whereas negative effects were exclusively negligible. Bootstrap resampling analyses (10,000 samples) showed that our results are extremely robust at the individual level (Fig. S21; neural: $p=0.005$, behavioral: $p<0.001$; correlation $p=0.006$).

In light of this evidence, we interpret this induced categorical perception change as possibly resulting from the neural sculpting of the categories. Accordingly, the increase in behavioral discrimination slope should be related to the increase in neural separation of the category representations. Indeed, across participants there was a significant positive relationship between training-related behavioral and neural changes: the relative increase in LLR between neural representations of shapes from distinct categories in the neurofeedback ROI was correlated with the relative strengthening of perceptual categorization in the trained direction compared to the control direction (Fig. 4C; Pearson $r=0.761$, $p=0.011$); a similar result was obtained using the LLR

ratio measure (Fig. S17D; r=0.686, p=0.029). Additionally, a linear mixed effects model analysis showed that behavioral changes in our cohort were significantly predicted by neural changes induced by neurofeedback, even when accounting for the effects of number of training days, baseline decoding accuracy during localizer scans, and the random category split chosen for each participant during training (p<0.001).

The success of our neural sculpting procedure may be related to participant age, which showed a trending correlation with the strength of the observed behavioral effects (r=0.600, p=0.067). One potential explanation for this result is that participant conscientiousness or their level of engagement with the experimental task (for both of which age may be a proxy) may play a critical role in the effectiveness of the neurofeedback manipulation. Additionally, we observed that individual variation in the ability to decode category information from neural activity patterns may also be a good predictor of the eventual success of our neural and behavioral manipulation procedure. Specifically, category decoding was the basis of neurofeedback and thus better performance would lead to more precise feedback. To test this relationship, we defined a baseline neural decoding measure for each participant as the model decoding accuracy in the neurofeedback ROI during the two localizer scans (Days 2-3). As expected, baseline decoding was high for all participants (75.7%±0.854% s.e.m., chance=50%, Fig. 3C), and no significant difference was observed at baseline between trained (78.1%±1.46% s.e.m.) and control (75.3%±0.875% s.e.m.) categories (t(9)=1.12, p=0.293). However, collapsing across categories, baseline decoding was highly correlated with the neural LLR change (r=0.720, p=0.019) and the behavioral slope change (r=0.777, p=0.008) from pre- to post-training. Interestingly, these correlations persisted (LLR, r=0.710, p=0.021; slope, r=0.729, p=0.017) even after excluding the trained direction from the baseline decoding average, suggesting that the precision of shape representations (or our ability to recover them) can be considered as a general property of individuals that can be used to predict training success. Moreover, these findings provide strong post-hoc support for our decision to focus neurofeedback training on participants with high baseline neural decoding.

Our results cannot readily be explained by differences in the amount of feedback received by each participant, either overall, or for each individual category. Because of our adaptive feedback procedure, each participant received a similar amount of feedback as a proportion of total experiment trials (28.7%±0.657% s.e.m.) and for each stimulus category (Category 1: 14.9%±1.74% s.e.m.; Category 2: 13.9%±1.43% s.e.m.; t(9)=0.305, p=0.767). Additionally, the distribution of feedback values received by participants was not different between Trained and Control category boundaries (t(9)=0.82, p=0.435; Kolmogorov-Smirnov test=0.50, p=0.111) and an ideal-observer analysis showed that, across our cohort, the difference in feedback received for the Trained and Control categories was not sufficient to predict the participants' neural (r=-0.447, p=0.196) or behavioral changes (r=0.320, p=0.283) after training, suggesting that the distribution of feedback values during training does not predict neural sculpting effects or participant performance.

Furthermore, our results cannot be explained by participants explicitly learning the random category distinction enforced during the training. At the conclusion of the study (but before being fully debriefed), we informed participants that the stimuli came from an unspecified number of categories. When asked to guess freely, participants reported an average of 3.3 categories whose boundaries did not coincide with those of the categories randomly selected for each of them in the experiment (Fig. S18 & Table S6). Only when subsequently forced to split the stimulus space into 2 categories using a straight line, their performance suggested a trend toward having some crude

information about the category boundary (Figs. S18, S19 & Table S6; trained=0°, average absolute displacement=31.5°±5.89° s.e.m; control=90°, average absolute displacement=58.5°±5.89° s.e.m.; random choice=45°, t(9)=2.18, p=0.058). However, any such information about categories cannot readily be explained by implicit learning or priming (*28,29*), given that the training included an equal number of stimuli from both categories, these stimuli were presented in a random order during training, and there were no differences in the amount of feedback received for stimuli of each category. Additionally, a support vector machine-based ideal observer analysis also confirmed that the feedback values received during training were not sufficient to predict shape category (decoding accuracy = 54.2%±3.51% s.e.m., chance = 50%; accuracy > chance, t(9)=1.14, p=0.283). This further suggests that any (small) amount of information about the category boundary that may have been inferable by participants due to the feedback itself was unlikely to have significantly influenced their ability to (not be able to) discover the category boundary during the experiment.

Discussion

Neural sculpting with real-time fMRI puts forward two key advances showcased by our proof-of-concept study involving novel visual category learning. First, prior correlational studies have linked distributed patterns or fMRI activity to the perception of objects from different categories (*30-32*). By sculpting new neural representations and inducing perceptual changes, we provide causal evidence that these representations are sufficient for categorical perception. Second, prior neurofeedback studies focused exclusively on reinforcing existing neural representations of visual features or categories (*11,12*). In contrast, here we use neurofeedback for the more radical goal of sculpting categories that did not previously exist in the brain. In doing so, we demonstrate a direct causal link between multivariate pattern representations in the human brain and perception of complex visual objects that is accessible and amenable to manipulation using non-invasive functional neuroimaging. Together, our findings broaden the possibility for non-invasive causal intervention in humans with neurofeedback fMRI, including the distant possibility of sculpting more extensive knowledge or complex concepts in the human brain, bypassing experience and instruction.

Previous studies have shown that learning novel categories induces increased within-category similarity (*9,10*), as well as neural and perceptual suppression of task-irrelevant features (*33,34*). Here, we used neural sculpting to induce a change in participants' neural representations of the shapes that was intended to mimic this learning effect and that created new categories both in behavior and in the brain. Consistent with these findings, the observed neural differences between trained and control categories reflected significant increases in sensitivity for trained categories and numeric, but not significant decreases in sensitivity for control categories (LLR, Figs. 4B & S16). A numerical, but not significant, effect in this direction was also observed in the perceptual measure (psychometric function slope, Figs. 4B & S16); the discrepancy between behavioral and neural results may be, in part, explained by reported participant fatigue. Additional work is required to disambiguate the contributions of these two opposite effects of feature enhancement and suppression and their interaction with perceptual change within the context of neural sculpting. Furthermore, we sculpted visual categories within a feedback ROI that comprised multiple disparate brain regions, tailored to the individual response of each participant during independent localizer scans. Our approach was designed to maximize the possibility that our manipulation would bring about causal behavioral change but leaves open the question of which

individual brain region(s) are necessary and sufficient to influence in order to induce perceptual change. The current experiment was not designed to fully answer this question, but rather as a proof-of-concept that some brain regions exist whose manipulation via neurofeedback is not only possible, but also sufficient for influencing perception. However, several post-hoc analyses investigating pre- and post-training category decoding (Fig. S22) suggest that more posterior (e.g., posterior LO) and/or more visually selective brain regions (e.g., the occipito-temporal part of the training ROI) are potentially undergoing the most change due to neurofeedback training. As such, it is possible that these regions might be driving some of the effects we observe in behavior. However, more work is needed to fully understand where the bulk of the changes are happening in the brain due to training, a question for future research that our neural sculpting technique is uniquely suited to answer.

We found that participants who showed higher baseline levels of neural shape decoding also showed larger training effects, suggesting that differences in either the precision of the neurofeedback signal or the quality of the underlying shape space representation may affect training outcomes. Similar to prior work (*12,35,36*), we observed variable training outcomes across our cohort, with some participants being more susceptible to neurofeedback manipulation compared to others. This may be due to individual differences in factors such as attentional control (*12*), differences in the ability to access and modify neural activity under an external constraint (i.e., non-responders; *34*), or differential plasticity across brain regions (*37*). Given the open-ended nature of our instructions for the neurofeedback trials ('*Generate a mental state that will make the shape wobble less*!'), it is also possible that differences in strategies employed by the participants may have played a role in achieving success. Consistent with this, participants reported extremely varied approaches to the task, the most common of which involved naming the shapes and/or focusing on local features (e.g., an indentation) instead of the entire shape. Although no consensus strategy could be identified, nor one that could explain either the neural or behavioral outcomes (Table S6), it is possible that further work could identify such strategies.

Our study represents a proof-of-concept that neural sculpting can change neural patterns in high-level visual cortex and also induce behavioral change in participants' perception. To maximize the ability of our neurofeedback neural models to provide consistent high-quality feedback to participants throughout the training portion of the experiment, we employed a relatively stringent criterion (70% decoding accuracy for shape categories in LO and the neurofeedback ROI) for selecting participants that would undergo the second half of the study, which resulted in 7 out of 17 participants not being invited back for the full study after the two localizer scans. Similarly, since the speed with which neural changes could be induced in human cortex as a result of neural sculpting remains unclear, we designed our study to include at least 5 fMRI neurofeedback sessions across separate days. While both of these features of our study may potentially increase the difficulty of using our method going forward, they also provide evidence that with strong neural signal and decoding accuracy, as well as a sufficient amount of neurofeedback training, our type of manipulation can indeed change neural patterns in high-level visual cortex and also induce behavioral change in participants' perception. Our study was meant to provide crucial evidence of what can be achieved with current technology in the field of neural and behavioral manipulation via neurofeedback, and we strongly believe that expected advances in signal acquisition (e.g., improved multiband protocols) will help both significantly increase base levels of neural decoding across the population, and reduce the training timeline requirements for similar studies in the future.

Our findings suggest that neural sculpting may also be an effective tool for influencing and potentially enhancing the neural effects of other learning processes that involve neural differentiation. For example, in the visual domain, our technique may prove useful for enhancing neural differences associated with domain-level expertise for categorization or recognition (e.g., in the fusiform face area, FFA; (*37,38*)) or brain-based education initiatives that complement classical learning paradigms (*39*). Conversely, our method may provide novel avenues to reverse de-differentiation of neural patterns stemming from aging (*40*) or from disorders impairing the natural function of visual brain regions such as visual agnosia (*41*) or prosopagnosia (*42*). Beyond the visual domain, previous studies have shown that other forms of neurofeedback aimed at enhancing or suppressing activity or connectivity of regions of interest (*11,43*) can be used to treat various neuropsychiatric disorders, such as major depressive disorder (*44-48*) and autism spectrum disorder (*49*). Our work provides a promising avenue towards potentially enacting more complex interventions in patient populations that would attempt to sculpt specific patterns of brain activity within regions of interest, in order to neurally mimic (or increase alignment with) neural activity patterns of healthy controls. Similarly, our work invites possible future applications for novel neurorehabilitation approaches, including brain-machine interfaces (*50*) and neuroprosthetics (*51*) that rely on generating or maintaining a specific multivariate pattern of brain activity in real-time.

To conclude, the work presented here represents a proof-of-concept for a new, non-invasive approach to investigating the causal relationship between neural representations and behavior using fMRI. More than 2,100 years after Mnesarchus of Athens made his observations about the nature of human experience (*1*), we show that his philosophical insights may, in fact, adequately describe how perception arises from human neural representations: sculpt a concept in the neural clay of the brain and it may subsequently exist.

Acknowledgments

The authors would like to thank M. L. Nguyen, A. N. Hoskin, E. A. Piazza, N. Keshavarzian, E. A. McDevitt, J. W. Antony, N. Y. Kim, and A. R. Burton for help with data collection; A. C. Mennen for discussions about the infrastructure for real-time fMRI neurofeedback; N. DePinto, L. E. Nystrom, and G. McGrath for technical support; and A. Tompary for contributions to initial pilot studies. Funding**:** John Templeton Foundation, Intel Corporation, NIH Award R01 MH069456.

Author Contributions

All authors contributed equally to conceptualizing the research and the paradigm. M.C.I. and V.J.H.R. collected the data. M.C.I. performed the data curation, formal analysis, validation, and visualization of the results with input from all other authors. M.C.I. wrote the original draft of the manuscript and all authors contributed substantially to reviewing and editing the final written version of the work.

Competing interests

Authors declare no competing interests.

Materials and Methods

Building and Norming the Shape Stimulus Space

We generated complex visual shapes defined by seven radial frequency components (RFCs) (*19-20*) (Fig. 1A). To obtain each shape, sine waves determined by the seven RFCs were added together and the resulting wave was wrapped around a circle to obtain a closed contour which was then filled in to create a shape. We created a two-dimensional shape space by independently varying the amplitude of two of the seven RFCs (from 12.6 to 36.6 for the 4.94 Hz component and from -6.0 to +42.0 for the 1.11 Hz component), while holding the amplitudes of all other components constant. Distance in this two-dimensional space was computed via changes in amplitude on each axis independently: a displacement of 1 arbitrary distance unit (a.d.u.) corresponded to a change in amplitude of 6 arbitrary units for the 1.11 Hz component and a change in amplitude of 3 arbitrary units for the 4.94 Hz component. Using this mathematical convention allowed us to generate novel visual shapes for any arbitrarily chosen point within this two-dimensional manifold.

To map the stimulus space perceptually, we chose shapes that sat at ±2, ±4, ±6, and ±8 arbitrary distance units from a fixed center shape along 6 equally spaced radial directions (60 degrees; AG, BH, CI, DJ, EK, FL; Fig. 1A). This yielded 9 equally spaced shapes along each of 6 radial directions, including the common center shape (e.g., A and G are the 8 a.d.u. shapes from center for the AG direction). We then recruited 16 healthy individuals ('norming cohort') from the Princeton community with normal or corrected-to-normal vision who provided informed consent to a protocol approved by the Princeton University Institutional Review Board and were compensated ($12 per hour) to participate in a self-paced two-alternative-forced-choice behavioral experiment. For each radial direction, participants were told that the endpoints (e.g., shapes A and G, 8-distance-units away from center, shown continuously on the left and right of the screen) correspond to two different categories of shapes and they were instructed to use this information to categorize new shapes that appeared in the center of the screen. New shapes were drawn randomly from the 9 shapes chosen for that particular radial direction (see above) such that 40 repetitions of all 9 shapes were shown in each run of the experiment. To ensure that categorization effects were not influenced by the left/right placement of the endpoints on the screen, each participant performed two separate runs for each radial direction (e.g., AG vs. GA) and the order of the runs and assignment of endpoints to each run was counterbalanced across participants. All shapes subtended 5 degrees of visual angle on the screen and the experiment was run entirely using Matlab (version 2016a) and PsychToolbox software (*52*). The endpoint shapes (identical to one of the 'categories') were also included as catch-trials and we excluded from the analysis six participants who incorrectly categorized more than 10% of the endpoint shapes for any given direction (criterion: >16 out of 160; 2 participants excluded: 29/160 and 67/160) or more than 5% of the endpoint shapes overall across all directions (criterion: >48 out of 960; 4 participants excluded: 77/960, 84/960, 93/960, and 456/960).

For the remaining 10 participants, for each radial direction, we computed the probabilities (x) of categorizing each individual shape as the two shapes at the endpoints and fitted a corresponding psychometric function (estimated *slope* and *threshold*; Fig. 1B-G):

$$f(x) = \frac{1}{1 + \exp\big(slope * (threshold - x)\big)}$$

To check that the stimulus space was perceived similarly across all 6 radial directions, we performed a repeated measures ANOVA on the slopes of the resulting psychometric functions with 'direction' as a factor.

All participants in the main experiment (see below, n=10, 'training cohort') took part in an identical self-paced two-alternative-forced-choice behavioral experiment on Day 1 (see below) of the study. We performed a repeated measures ANOVA on the slopes of the resulting psychometric functions with 'direction' and 'cohort' as factors to verify that no significant differences existed between how participants in the norming and training cohorts perceived the stimulus space (Fig. S1).

fMRI Localizer Scans to Identify Cognitive Map Brain Regions

The localizer scans (which were not analyzed in real time) involved showing participants sample shapes from the stimulus space in order to (a) test whether we can decode shape category information with high precision from neural patterns (participant exclusion criterion detailed below); (b) find all brain regions that represent the stimulus space as a cognitive map ((*23-24*) prerequisite for defining our neurofeedback target region of interest (ROI), detailed below); and (c) to build a neural model of the shape space (detailed below) to be used during real-time neurofeedback training to track how participants' brains represent the stimuli presented on the screen in each experimental trial.

We performed two scan sessions per participant on separate days. To help minimize in-scanner motion and encourage participants to self-regulate this potential source of noise during the scans, we chose to avoid the use of rigid restraints (for comfort, given the length and number of sessions), but did use masking tape to both gently tether the participants' heads to the coil, and to provide physical feedback whenever they attempted to move during the experiment. Each scan comprised an anatomical scan (T1 MPRAGE acquisition), followed by 4-8 functional (echoplanar) runs (11m23s each) for a total of 12-15 total functional runs collected per participant across the two scan days. During each functional run, participants were shown 49 distinct stimuli from the shape space using a short block design: 2.5s stimulus presentation, black shapes on equiluminant gray background; followed by a 10 second inter-stimulus interval during which a countdown to the next stimulus presentation was displayed in white font on the same equiluminant gray background. During future real-time fMRI scans, z-scoring with respect to the full response of each run would not be possible during continuous acquisition. Thus, to normalize signal in each voxel and run in a consistent manner, we included a 70s countdown (white font on equiluminant gray background) at the beginning of each run, during which the participants were instructed to keep their eyes open and stay alert. After eliminating the first 6 TRs to allow T1 equilibration and accounting for hemodynamic lag, this yielded 30 TRs of functional data per run, comparable between all runs (localizer and real-time), which could be used to normalize the signal acquired at each subsequent TR (see fMRI Preprocessing for Localizer Scans). Stimuli (color: black) were displayed on a rear-projection screen using a projector (1024x768 resolution, 60 Hz refresh rate)

and subtended 5 degrees of visual angle against a uniform gray background. Participants viewed the visual display through a mirror that was mounted on the head coil. Stimulus presentation was performed entirely using Matlab (version 2016a) and PsychToolbox software.

For each localizer run, similarly to the shape space norming experiment, we chose shapes that sat at ±2, ±4, ±6, and ±8 arbitrary distance units from a fixed center shape along 6 equally spaced radial directions (60 degrees; AG, BH, CI, DJ, EK, FL; Fig. 1A). This yielded 9 equally spaced shapes along each of 6 radial directions, including the common center shape (e.g., A and G are the 8 a.d.u. shapes from center for the AG direction), for a total of 49 unique shapes. The presentation order was randomized across runs and across participants. During each unique shape presentation (3s), participants were asked to indicate using an MR-compatible button box whenever a shape 'oscillated' (orthogonal task to ensure alertness). An 'oscillation' was defined as a parametric continuous perturbation of the shape in a random direction in the parametric manifold, after which it returned it its original position with the same speed that it was perturbed (250 ms total, i.e., we picked a random point in parameter space near the original shape chosen by drawing an i.i.d. sample from the circumference of a circle of radius 2 arbitrary distance units centered at the shape, then we created an animated morph between the original shape and the one defined by the point on the circle, and back to the original shape). Each trial had either 0, 1, or 2 'oscillations', with an average of 1 'oscillation' per trial, randomized across all trials in a run. The absolute position of the center of mass of the shapes was always held constant at the center of the screen; the oscillation in parameter space simply created the visual impression that the current shape morphed into a similar shape and then reverted to the original (see Movie S1 for a demo of continuous oscillation).

fMRI Acquisition

Structural and functional MRI data were collected on a 3T Siemens Skyra scanner with a 64-channel head coil. Functional images were acquired using an echoplanar imaging sequence (TR, 2000 ms; TE, 28 ms; 36 transversal slices; voxel size, 3x3x3 mm; 64° flip angle; IPAT factor 2). This sequence produced a full brain volume for each participant. Anatomical images were acquired using a T1-weighted MPRAGE sequence, using a GeneRalized Autocalibrating Partial Parallel Acquisition (GRAPPA) acceleration factor of 2 (TR, 2530 ms; TE, 3.3 ms; voxel size, 1x1x1 mm; 176 transversal slices; 7° flip angle). Given the time constraints of real-time fMRI processing and our goal of matching acquisition parameters for classic and real-time scans, we did not correct for susceptibility-induced distortions in any of our echoplanar images.

fMRI Preprocessing for Localizer Scans

The images were preprocessed using custom AFNI (*53*), Freesurfer (*25*), and bash scripts. All analyses were performed in participants' native space and no smoothing was applied. The first six volumes of each run were discarded to allow T1 equilibration. For each run, the remaining functional images were spatially realigned to correct for head motion and registered to the participants' structural T1 image, using boundary-based registration implemented in AFNI's afni_proc.py. We then performed polynomial trend correction using AFNI's 3dDeconvolve tool and simultaneously regressed out 6 degrees of head motion (x, y, z, roll, pitch, yaw) from the functional data. We used FreeSurfer's recon-all tool to estimate the boundaries of the gray matter

for each participant, as well as anatomically defined lateral occipital (LO) and early visual cortex (EVC) regions of interest, and AFNI's 3dSurf2Vol and 3dAllineate to obtain corresponding volume masks of the gray matter and regions of interest aligned to each participant's functional data. The functional data for each localizer run within these regions was then z-scored using the mean and standard deviation of each voxel during the countdown at the beginning of each corresponding run (30 TRs after the first six volumes were discarded).

Neural Cognitive Map of the Shape Stimulus Space and Neurofeedback Region of Interest

To measure whether object-selective cortex (LO) represents the shape space as a cognitive map (i.e., similar to how it's built parametrically and how it's perceived by observers), we picked five out of the nine shapes for each direction (Fig. 2B; center, ±4 a.d.u, ±8 a.d.u.) and computed a representational similarity matrix (RSM) describing the relative relationships between neural activity elicited by the shapes in LO. We compared this RSM with an ideal RSM obtained by assuming a parametric linear distance relationship between the neural activity of the shapes (Figs. 2C & S2; e.g., the more distinct the shapes, the larger the distance between them) using Pearson correlation.

To construct our neurofeedback region of interest (ROI), we sought all brain regions in each participant's brain that represented the stimulus space parametrically (i.e., akin to a cognitive map (*23,24*)) and ran a searchlight analysis (*22*) (cube of side length 7 voxels intersected with gray matter volume, minimum 172 voxels, distance between cube centers = 2 voxels) across their entire gray matter mask. In each voxel cube, we computed the Pearson correlation with the ideal RSM (Fig. 2D). First, to increase the probability that potential sources of top-down control (e.g., prefrontal cortex, parietal cortex, etc.) would be subsumed by the neurofeedback ROI, we ensured that it must include at least one non-connected cluster of 50 or more voxels outside of extrastriate visual cortex (i.e., not connected to LO). Since the main goal of the current experiment was to be a proof-of-concept that some brain regions exist whose manipulation via neurofeedback is not only possible, but also sufficient for influencing perception, we sought to include voxels outside of visual / occipito-temporal cortex; if such voxels would be amenable to neurofeedback manipulation, and changes in these voxels were causally related to behavioral changes, then including these voxels would maximize our chances to elicit such a behavioral effect in our participants. Second, to ensure that ROI sizes are relatively similar across participants despite natural variability in signal strength, we restricted the final ROI size to between 750 and 2,250 voxels across all participants. To satisfy both of these constraints, we started by selecting all clusters of 20 or more voxels that had an r>=0.50 with the ideal RSM, excluding the EVC. If the conditions were met, this became our neurofeedback ROI. Otherwise, we either relaxed or increased the r threshold stepwise using values from the set {0.25, 0.33, 0.40, 0.50, 0.60, 0.66} until both of the conditions were met for each individual participant. Final thresholds and resulting neurofeedback ROI sizes for each participant are shown in Table S1. Surface maps of final ROIs for each participant in the full experiment are shown in Figs. S3-S12.

Neural Model of Shape Stimulus Space

To measure how the shape space was categorically represented in the neurofeedback ROI, we split the stimulus space into groups of two categories each comprising half of the shape space circle (Fig. 3A). To maximize the number of training examples available, the category boundaries

were chosen halfway between the diameters defined in Fig. 1A, e.g., halfway between lines AG and BH, halfway between lines BH and CI, etc. This yielded 6 categorical partitions of the shape space, each associated with 24 out of 49 unique trials per localizer experiment run (24 shapes shown belonged to one category partition, 24 belonged to the other partition, and the center shape was on the dividing category boundary line, and thus was not used for estimating a category model). For each trial in each functional localizer run, we selected two TRs following each short block onset (accommodating for optimal activation given hemodynamic lag, see below) and considered them as independent training points in order to maximize the amount of training data available. This procedure yielded 48 training points per category per functional run, with a total of 576—720 training points per category per run for each participant and each of the 6 categorical partitions (i.e., ABCDEF vs. GHIJKL, BCDEFG vs. HIJKLA, CDEFGHI vs. JKLABC, etc.). To build a neural model, we modeled each categorical partition as a multivariate Gaussian distribution. The parameters of these category-specific distributions were computed using maximum likelihood estimation using the localizer scan data associated with each category partition (schematic shown in Fig. 3B). Given the high dimensionality of the training data (750—2,250 voxels per participant) and the relatively small number of training examples (n=576—720 per category per participant), we used a dimensionality reduction procedure to project the training data (voxel activations) onto its first 100, 150, and 200 principal components ($ProjectedVoxelActivity_k=\{projectedVoxelActivity_i\},\ k \in \{1,2\}$) and estimated a shared covariance matrix (Σ_S) for both shape categories in each partition:

$$\mu_1 = \frac{1}{n} \sum_{i \in Category1}^{n} projectedVoxelActivity_i$$

$$\mu_2 = \frac{1}{n} \sum_{i \in Category2}^{n} projectedVoxelActivity_i$$

$$\Sigma_1 = \frac{1}{n} ProjectedVoxelActivity_1^T * ProjectedVoxelActivity_1 - \mu_1^T * \mu_1$$

$$\Sigma_2 = \frac{1}{n} ProjectedVoxelActivity_2^T * ProjectedVoxelActivity_2 - \mu_2^T * \mu_2$$

$$\Sigma_S = \frac{1}{2}(\Sigma_1 + \Sigma_2)$$

Additionally, given potential variability in hemodynamic lag between brain regions and across participants (*54*), we also selected two potential lags: the training data comprised either the 3rd and 4th TRs following short block onset (4s lag) or the 4th and 5th TRs following short block

onset (6s lag). We built neural category models for each direction in each participant using all six combinations of these two parameters (PC dimensions = 100, 150, 200 x hemodynamic lag = 4s, 6s) in both LO and the neurofeedback ROI. To verify that our model could correctly predict the distinction between shape categories in the brain, we used it as a linearly discriminative log-likelihood-ratio-based pattern classifier by computing the probability of a new point x (n-dimensional neural representation of a new shape) under the two estimated distributions:

$$p_1(x) = \frac{1}{\sqrt{|\Sigma_S|(2\pi)^n}} \exp\left(-\frac{1}{2}(x-\mu_1){\Sigma_S}^{-1}(x-\mu_1)^T\right)$$

$$p_2(x) = \frac{1}{\sqrt{|\Sigma_S|(2\pi)^n}} \exp\left(-\frac{1}{2}(x-\mu_2){\Sigma_S}^{-1}(x-\mu_2)^T\right)$$

$$If\ log\frac{p_1(x)}{p_2(x)} > 0, \qquad then\ predict\ Category\ 1, \qquad else\ predict\ Category\ 2$$

Using our classifier, participants whose average leave-one-run-out cross-validated decoding accuracy across all 6 category partitions was below 70% (chance = 50%) in either LO or the neurofeedback ROI using all possible parameter combinations were excluded from participating in the training portion of the experiment. This stringent exclusion criterion was employed because (1) neurofeedback neural models require high specificity to be able to provide correct feedback during training; and (2) we sought to maximally select for the strongest known predictor of the effectiveness of neurofeedback manipulation across our cohort: neural signal strength as measured by baseline visual category decoding accuracy). As such, participants whose reliability of decoding was below this threshold (70%) were considered to have too low signal to afford reliable real-time neurofeedback training. For all participants that met the threshold, we selected the parameter combination (PC dimensions x hemodynamic lag) that gave the highest average decoding accuracy across all 6 categorical partitions in the neurofeedback ROI (Fig. 3C), regardless of performance in LO. Localizer sessions decoding results for LO and neurofeedback ROIs, as well as final parameter combination choices for each trained participant are shown in Table S1.

Neurofeedback fMRI Experiment Design

The full neurofeedback experiment comprised ten total sessions per participant (Fig. 2A):

Session 1: Behavioral 2AFC pre-test

Sessions 2-3: Neural localizer scans / Model training data acquisition

Sessions 4-9: Real-time fMRI neurofeedback training

Session 10: Behavioral 2AFC post-test (identical to pre-test, counterbalanced presentation order)

We designed our experiment with a planned sample size of 10 participants and we recruited and trained participants sequentially until 10 of them both met the initial decoding thresholds listed

in section Neural Cognitive Map of the Shape Stimulus Space and finished the training and behavioral post-test portion of the study in its entirety (i.e., without dropping out of the study voluntarily). As such, we recruited a total of twenty-five healthy participants over the course of 17 months (May 2018-October 2019; 13 female, 23 right-handed, ages 18-35, mean age 22.4) from the Princeton community with normal or corrected-to-normal vision who provided informed consent to a protocol approved by the Princeton University Institutional Review Board and who were compensated for their participation ($12/h for behavioral sessions, $20/h for fMRI sessions, retention incentives: $5/session cumulative bonus for each fMRI session beyond the first, $0.10 per successful neurofeedback trial, $100 for completing the full experiment). The behavioral pre-test and the neural localizer scans were run within the span of 1-3 weeks for each participant. If the participant had satisfactory performance on the behavioral test (using criteria listed in section Building and Norming the Shape Stimulus Space) and if a neural model with high decoding accuracy could be built using the neural data from the two localizer scans (using criteria listed in section Neural Cognitive Map of the Shape Stimulus Space), then the participant was invited to take part in the rest of the experiment. Of our initial set of twenty-five participants, two were rejected for not meeting MRI safety criteria after participating in the behavioral pre-test, eight were rejected for poor behavioral and/or neural decoding performance during sessions 1-3, and five withdrew their participation after 1-3 sessions. Ten participants (7 female, all right-handed, ages 20-35, mean age 24.0) took part in the full experiment, two of which were trained for 5 days and eight of which were trained for 6 days. Participant performance during the localizer sessions and reasons for exclusion are shown in Table S1.

Each session of the experiment was run on a separate day. The real-time training portion and subsequent behavioral post-test took place 2-40 days after the localizer scans. To minimize fatigue and afford potential benefits from overnight memory consolidation (*55*), we sought to have participants come into the lab for at most 2 hours each day on consecutive days during the training portion of the experiment. Participants performed a variable number of trials per day, depending on the amount of time the scanner setup and participant pre-scan procedures took (minimum 20 trials, maximum 140 trials per day). We ensured that participants performed at least 500 total neurofeedback training trials before the final behavioral post-test. To minimize potential effects of over-training and to guarantee a comparable amount of training across our entire cohort, we stopped the training procedure once a participant performed 800 trials (two participants). Participants were allowed to skip at most one day for emergencies and/or personal reasons (ensuring that no more than 48h elapsed between sessions when this happened) and three participants chose to avail themselves of this opportunity.

Real-Time fMRI Neurofeedback Procedure

For each participant, after choosing the optimal combination of parameters for our neural model (see Neural Model of Shape Stimulus Space), we randomly choose one of the 6 radial category partitions (by rolling the same pink six-sided die for all participants) to become their training category boundary (information hidden from participants), ensuring that neural decoding accuracy was at least 70% for this specific direction and that it hadn't been chosen before for any previous participant (after we trained 6 participants on all 6 distinct partitions, we reset this latter constraint).

We performed five to six neurofeedback training scan sessions per participant on separate days (see Table S3 for session details). Each scan comprised an anatomical scan (T1 MPRAGE acquisition), followed by one functional (echoplanar) localizer run (11m37s, identical to Sessions 2-3), and 1-7 functional (echoplanar) neurofeedback training runs per day (10m32s each) for a total of 26-40 total functional neurofeedback runs collected per participant (active in-scanner training time: 4.6h-7.0h per participant). The training runs used a block design with twenty trials each (16s=8TRs stimulus presentation, followed by 12s=6TRs ITI) and a 72s countdown at the beginning of each localizer run, analogous to the localizer scans (see fMRI Localizer Scans to Identify Cognitive Map Brain Regions), during which the participants were instructed to keep their eyes open and stay alert. Stimuli (color: black) were displayed on a rear-projection screen using a projector (1024x768 resolution, 60 Hz refresh rate) and subtended 5 degrees of visual angle against a uniform gray background. Participants viewed the visual display through a mirror that was mounted on the head coil. Stimulus presentation was performed entirely using Matlab (version 2016a) and PsychToolbox software. For each trial, a random shape was chosen from the stimulus space (i.i.d. from the shape space circle, real number parameters distinct from localizer scan shapes), making sure that equal numbers of shapes from both categories are shown in each training run (ten each, randomized order). Since preliminary tests showed that the decoding performance of the neural shape model deteriorated for shapes close to the category boundary (Fig. S15), we ensured that chosen shapes were at least one arbitrary distance unit away from the category boundary for each participant. To ensure that this constraint did not introduce an artificial perceptual cue about the category boundary, we also excluded shapes that were within one arbitrary distance unit from the perpendicular direction to the category boundary (i.e., the direction of maximal separation). For each trial, we generated a continuous oscillation centered at each trial-specific shape, determined by drawing points i.i.d. from a uniform distribution on the unit circle of a fixed radius around the shape (Fig. S14). Similarly to the localizer scans alertness task (except with no stops in movement), the shape traveled to the point on the circle via the corresponding radius, then returned to the original shape, then traveled to the next randomly selected point on the circle, etc. Each oscillation (back and forth to the circle) took 500ms, with 4 independent oscillation for each 2s TR (Movie S1).

The radius of the circle that determined each cluster of 4 oscillations in each TR was the main method by which neurofeedback was given to study participants. They were told that during each trial they would see a continuously oscillating shape on the screen and given the following set of instructions:

(1) "Generate a mental state that's going to make each shape oscillate / wobble less or even stop!";

(2) “Different shapes may require different strategies.”; and

(3) “When you are successful in slowing down the shapes, it has to do with your mental state over the past 8-10 seconds; progress is not instantaneous.”

The first instruction encouraged participants to generate mental state variability that may help shift representations in the neurofeedback ROI. The second instruction addressed the possibility that shifting representations in neural space in different (opposing) directions of a cognitive map may require different types of top-down influences; while it may have alerted participants to the possibility that multiple categories of shapes exist, the high variability of the shape space was designed to make boundaries between such categories difficult to intuit. The final instruction was necessary due to hemodynamic lag inherent to the fMRI signal, which together

with its temporal autocorrelation properties that would invariably influence the visual (shape-related) feedback received during the experiment. As participants attempted to generate such a target mental state while looking at the shape on the screen, during each TR (2s), we preprocessed their functional data (see fMRI Preprocessing for Real-Time Scans), and then we used our neural model to identify how far from the neural category boundary the shape was represented in the participants' neurofeedback ROI; if the shape was strongly represented as a member of its ground truth category (known to the experimenter, but not to the participant), then positive feedback was given to the participant by shrinking the radius of the circle from which the oscillation parameters were chosen (causing the oscillation to look less extreme, Movie S1, trial 2), otherwise no feedback was given and the participant saw the oscillation continue with the same apparent amplitude (see Neurofeedback fMRI Data Analysis for details). Each oscillating shape was presented for a total of 8 TRs. During the first 3 TRs of the block, no feedback was given since hemodynamic lag prevented it. Afterwards, for the remaining 5 TRs of the block, participants received appropriate (or no) feedback on a TR-by-TR basis. Feedback was cumulative, such that if the participant maintained a mental state for the shape that took it away from the category boundary for 3 TRs out of the 5 TRs in which feedback is given, then the shape actually stopped oscillating altogether (radius of oscillation decreased to zero). In practice, this was extremely rare, but did occur sometimes, especially towards the end of the experiment. We limited each training trial block to 16s (8 TRs) to eschew potential repetition suppression effects that might diminish our ability to apply our neural model and/or decode the category of the shape on the screen (*26*). Participants were also instructed that they would receive a monetary reward of $0.10 every time they shrank the radius of the oscillation (positive feedback), together with a bonus of $0.25 if they managed to stop the oscillation completely (positive feedback for 3 out of a total of 5 possible feedback TRs per trial).

fMRI Preprocessing for Real-Time Scans

Structural and functional MRI data were collected in an identical fashion to the localizer scans (coil, equipment, parameters, etc.). The images were preprocessed using custom AFNI (*53*), Matlab (version 2016a), C++, and bash scripts. All analyses were performed in participants' native space and no smoothing was applied. The first six volumes of each run were discarded to allow T1 equilibration. We used AFNI's afni_proc.py script to preprocess the localizer run from each training day scan (analogous to fMRI Preprocessing for Localizer Scans), as well as align this functional run to the anatomical scan matching the Sessions 2-3 data used to generate the offline neural shape model. This allowed us to generate a template image (the 50th volume of the localizer EPI run) to which we would be able to align new volumes acquired during the neurofeedback training runs in a reasonable amount of time (aligning to an anatomical scan or to a full functional run usually requires tens of seconds and is not useable in a real-time setting, but aligning a single volume to another single volume requires <500ms on average). This template volume was also aligned to the neurofeedback ROI mask, which allowed fast recovery of the relevant functional data from each real-time acquired volume after fast single-volume-to-single-volume alignment, such that the neural model could subsequently be evaluated in real time.

The real-time functional scans were transferred via a fast network connection from the scanning console (Siemens) memory directly to the hard drive of a high-powered server on which all analyses were performed. Each real-time volume was aligned to the template volume (see above) using AFNI's 3dvolreg and the neurofeedback ROI mask was used to subsequently extract

the relevant functional data. Given the constraints of the real-time environment, classical preprocessing techniques and normalization procedures could not be implemented (e.g., polynomial trend regression, motion parameter regression, z-scoring with respect to the entire timecourse). Instead, the functional data acquired up to a given TR were temporally high-pass filtered (e.g., on TR 45, the first 45 TRs were filtered together; on TR 46, the first 46 TRs were filtered together; etc.) with a 52s (26 TRs) period cutoff (the length of two consecutive trials, including ITIs) using a fast custom script written in C++. The functional data was then z-scored using the mean and standard deviation of each voxel during the countdown at the beginning of each corresponding run (30 TRs after the first six volumes were discarded to allow T1 equilibration).

Neurofeedback Online fMRI Data Analysis

The functional data from the neurofeedback ROI was projected into the dimensionality-reduced space defined during the procedure used to select the optimal number of PC dimensions and hemodynamic lag for shape category decoding (see Neural Model of Shape Stimulus Space). The multivariate Gaussian neural model was then used to estimate the log-likelihood ratio (LLR) of the current data point (representing the neural activity elicited by the current stimulus being shown on the screen in the scanner) for each of the two categories of shapes (e.g., ABCDEF vs. GHIJKL) selected for training for the current participant. At the beginning of each trial, the default oscillation radius (see Real-Time fMRI Neurofeedback Procedure) was set to 1.875 arbitrary distance units in the parameter space based on a trade-off between ensuring that the perturbation is visually salient and pilot results regarding how much the wobble amplitude degraded stimulus category decoding (Fig. S15). If the computed LLR for a given TR was above a particular threshold (described below), then the amplitude of the oscillation was reduced during the subsequent TR by 0.625 arbitrary distance units (for all four oscillation iterations shown during that 2s TR, 500ms each), otherwise the radius was kept unchanged. Feedback was cumulative, allowing the radius of the oscillation to become zero (thus causing the shape to become static on the screen) after 3 cumulative (but not necessarily consecutive) TRs of positive feedback during a trial (see Fig. S14 for a diagram of this process). The LLR threshold for the first training run was initially set to the 70th percentile of the distribution of all data points (1,152—1,440 per participant) collected during the localizer scans for that particular participant under the optimal offline Gaussian neural model. After the first training run, the threshold was potentially adjusted using an adaptive procedure for two reasons. First, since the only way in which participants can learn to shift their neural representations is through receiving feedback via the shape oscillation radius, we sought to converge on giving feedback for approximately 1/3 of the twenty trials in each run. Second, given the high variability of conditions elicited by fMRI scanning across multiple days in terms of signal strength and sources of external noise, we expected our ability to decode using a static neural model to vary across the training week, as well as across training runs. As such, an adaptable threshold can recover from being too stringent (and never providing feedback) or too lax (providing feedback on almost every trial), both of which would be undesirable outcomes of the training procedure. Thus, the threshold was adjusted given participant performance (how many trials generated positive feedback) on all previous runs since the beginning of the training day as detailed in Table S2. By design, positive feedback could only be given on trials where the model correctly guessed the ground truth category of the shape being shown and, to avoid potentially random feedback being given due to classifier uncertainty near the category boundary (Fig. S15),

the threshold was never lowered below 60% of the localizer scan LLR regardless of participant performance. The threshold was also kept unchanged between the last session of a training day and the first session of a subsequent day, regardless of performance on the final run of the previous day. The LLR distribution for a representative participant, together with a sample threshold (75%) are shown in Fig. 3D. Distributions and sample thresholds for all participants are shown in Fig. S13.

Post-Experiment fMRI Data Analysis

To evaluate our ability to influence neural representations using neurofeedback, we computed the model LLR of all stimuli shown to each participant across all their neurofeedback runs during their first two days of training (pre) and their last two days of training (post), for the trained category boundary (Trained, e.g., ABCDEF vs. GHIJKL) and for the orthogonal category boundary (Control, e.g., DEFGHI vs. JKLABC) (Fig. 4A). To maximize sensitivity of this measure given potential repetition suppression effects (*27*), we focused on LLR values from the first TR of each trial only as the most reliable measure of LLR across the neurofeedback training runs. Results using the first 2, 3, 4, and 5 TRs of each trial (participants received at most 5 instances of feedback per trial) were highly similar and shown in Table S4.

To measure relative neural change as a consequence of training, we computed a participant-level summary statistic of the effect of neurofeedback training by averaging the LLR across the first two days (pre) and the last two days (post) and used two measures. First, we used a difference score to compute the LLR change separately for the trained and control categories:

$$Trained_diff(pre, post) = trained_{post} - trained_{pre}$$

$$Untrained_diff(pre, post) = untrained_{post} - untrained_{pre}$$

We then used a paired t-test to investigate whether training induced different changes for the trained and control category distinctions and computed Cohen's d to measure the effect size of this change between trained and control categories.

Second, we also computed the ratio between trained and control LLR for the pre and post conditions:

$$Pre_diff_{ratio}(trained, control) = \frac{trained_{pre}}{control_{pre}}$$

$$Post_diff_{ratio}(trained, control) = \frac{trained_{post}}{control_{post}}$$

We then used a paired t-test to investigate whether the LLR ratio changed as a consequence of training between the first two days and the last two days of the experiment (Fig. S17).

To measure potential changes in neural category patterns across our experiment, we computed Pearson correlation with the ideal parametric similarity matrix (Figs. 2C & S22) for patterns of activity elicited in LO and in the neurofeedback ROI during the single functional localizer run from the beginning of each in each neurofeedback fMRI training session (see Real-Time fMRI Neurofeedback Procedure) for the first two vs. last two days of training (1 run per participant per day), which is similar in design (and hence shapes presented) to the runs of the localizer scans (Days 2—3, 12-15 runs per participant; see fMRI Localizer Scans to Identify Cognitive Map Brain Regions). To measure whether a potentially stronger effect of categorical change in neural patterns would manifest within a different ideal correlation pattern, we also built a categorical ideal matrix (Fig. S22) and replicated this analysis for the LO and neurofeedback ROI.

To measure the strength of category representations within sub-regions of our neurofeedback ROIs, we: (a) split LO into equal sized anterior (aLO) and posterior (pLO) halves using the Z coordinate of the Freesurfer map from which the ROI was defined; and (b) split the neurofeedback ROI into "visual" (vis-ROI, all voxels within occipito-temporal cortex) and "high-level" (high-ROI, all remaining ROI voxels outside occipito-temporal cortex) components. We then trained new category decoding models within these pairs of sub-ROIs for the runs collected during the two localizer scans (Days 2—3, 12-15 runs per participant; see fMRI Localizer Scans to Identify Cognitive Map Brain Regions).

Behavioral Tests, Debrief, and Post-Experiment Data Analysis

To evaluate the influence of our neurofeedback training on participants' perceptual experience, we conducted a 2AFC experiment (procedure identical to that described in Building and Norming the Shape Stimulus Space) during the first and last sessions of the full experiment. Analogously to the norming experiment, we estimated psychometric function slopes for each of the 6 radial directions in our stimulus space separately for the behavioral pre-test (before training, first session), and for the behavioral post-test (after training, last session). To compare the strength of category separation between the pre- and post-tests, we computed the average slope for the two directions most perpendicular to the training category boundary in each participant (Trained, e.g., for categories ABCDEF and GHIJKL, the average slope of lines CI and DJ), as well as the average slope for the two directions most parallel to the training category boundary in each participant (Control, e.g., for categories ABCDEF and GHIJKL, the average slope of lines AG and FL) (Fig. 4A). As a control analysis, in addition to the change in psychometric slope for the trained and Control directions in the stimulus space for each participant, we also computed this change for the average of the two directions that were neither parallel, nor perpendicular to the trained category boundary (Neutral, e.g., for categories ABCDEF and GHIJKL, the average slope of lines BH and EK) (Fig. S20).We used a normalized difference score to compute the relative change in psychometric function slope before and after training for the trained and Control directions:

$$Trained_diff_{norm}(pre, post) = \frac{trained_{post} - trained_{pre}}{trained_{post} + trained_{pre}}$$

$$Control_diff_{norm}(pre, post) = \frac{control_{post} - control_{pre}}{control_{post} + control_{pre}}$$

Finally, we computed a two-sided paired t-test between Trained and Control slopes to evaluate whether our neurofeedback training procedure induced a significant behavioral change consistent with our hypothesis (e.g., higher relative psychometric function sharpening in the Trained direction, compared to the Control direction) and we computed Cohen's d to measure the effect size of this behavioral change. Estimated psychometric functions for all tests, directions, and participants are shown in Table S5.

To measure whether a relationship exists between neural change and perception, we computed the Pearson correlation between the LLR change (computed using both difference score and ratio) and the behavioral change, i.e., how well does LLR change in the neurofeedback ROI predict psychometric function change for the Trained vs. Control direction in behavior.

To measure whether our results are robust in individuals and at the group level with our sample size (n=10), we performed a bootstrap resampling analysis for the neural change elicited by the neurofeedback manipulation between the first two days of training and the last two days of training, as well as for the behavioral change observed between Day 1 and Day 10 of the experiment, and the correlation between neural change and behavioral change. The analysis employed a sample-with-replacement procedure that was used to generate 10,000 random draws of 10 samples (i.e., 10 participants) from the distributions of observed neural and behavioral effects (i.e., LLR difference values and psychometric function normalized difference scores reported in Fig. 4), and the subsequent computation of the Pearson correlation of each matched sample pair. Histograms of the sample values for the neural, behavioral, and correlation effects are shown in Fig. S21.

To investigate if participant training outcomes may be related to individual differences in access to neural information via computational analysis of fMRI data, we defined the baseline neural decoding measure for each participant as the Gaussian classifier decoding accuracy within the training ROI during the two localizer scans (Days 2 & 3), averaged across all six directions of the stimulus space we scanned (Fig. 3C). We then computed the correlations between baseline decoding for each participant and their LLR neural change measures, as well as their behavioral outcomes. Additionally, we tested whether the effects of our neural sculpting procedure can be predicted by participant age.

To measure the strength of category representations within sub-regions of our ROIs, we (1) split LO into equal sized anterior (aLO) and posterior (pLO) halves using the Z coordinate of the Freesurfer map from which the ROI was defined; and (2) split the neurofeedback ROI into "visual" (vis-ROI, all voxels within occipito-temporal cortex) and "high-level" (high-ROI, all remaining ROI voxels outside occipito-temporal cortex) components. We then trained new category decoding models within these pairs of sub-ROIs for the runs collected during the two localizer scans (Days 2—3, 12-15 runs per participant). To quantify changes in decoding accuracy

due to training, we computed the decoding accuracy of these models for the trained and control directions, across the calibration runs collected during first two days and the last two days of training. We also computed the difference of differences for this latter quantity (Fig. S22).

We also performed a linear mixed effects model analysis (Matlab function: fitlme) to measure whether behavioral changes observed in our cohort were significantly predicted by neural changes when accounting for the effect of number of training days (5 or 6), baseline decoding during localizer scans (cf. above), and the random category split assigned to each participant during training.

After the two-alternative forced choice behavioral post-test was completed (but before full debrief as to the goals of the study), we asked our participants to fill out a short questionnaire about their experience, including asking them about whether they suspected there were multiple categories of shapes, if so, how many and which ones, as well as commenting on strategies they used to perform the task during neurofeedback. The questionnaire and a summary of the answers are shown in Table S6. The free guess category boundaries, as well as the 2-category forced choice category boundaries given by each participant are shown in Figs. S18 & S19.

To measure whether the distribution of positive feedback trials throughout the experiment could explain our results, we also performed ideal observer analyses to measure whether the feedback values received during training would be sufficient to predict the participants' neural or behavioral changes or to explain the participants' forced choice guesses about the category boundary at the end of the experiment. First, we computed the proportion of feedback points that fell into the two sets of putative categories defined by the Trained boundary used in the experiment, as well as by the perpendicular, Control, boundary, and we normalized the amount of feedback to a proportion for each category in each condition. Then, we computed D1 as the difference in the proportion of positive feedback received on the two sides of the Trained boundary and D2 as the difference in the proportion of positive feedback received on the two sides of the Control (perpendicular) boundary. Using a two-sided t-test and a Kolmogorov-Smirnov test, we measured whether the proportion of feedback differs between the Trained and Control boundaries (D1 vs. D2). We also computed the Pearson correlation between the difference of differences (D3=D2-D1) and the neural and behavioral changes observed after training. Second, to further test whether an ideal observer model could have used the feedback received during training to guess the correct shape category boundary, we trained a linear support vector machine classifier (SVM) and employed a ten-fold cross-validation procedure to predict shape category based on the feedback values received during training.

Data and Materials Availability

Raw fMRI data for all 98 experimental sessions of the study is publicly available in BIDS format in the following NIH repository: https://nda.nih.gov/study.html?id=1098. Before publication, code and derivatives for the study are available upon reasonable request from the corresponding author (M.C.I.). Upon publication, all code and derivatives will be made publicly available through the Open Science Framework (OSF) data sharing platform, on as well as on the corresponding author's website (www.MariusCatalinIordan.com).

Supplementary Materials

Figures S1-S34

Tables S1-S6

Movies S1-S2

References:

1. A. A. Long, D. N. Sedley, *The Hellenistic Philosophers*, Vol. 2 (Cambridge University Press, Cambridge, UK, 1987).
2. S. P. Thompson, E. L. Newport, Statistical learning of syntax: the role of transitional probability. *Lang. Learn. And Devel.* **3**, 1-42 (2007).
3. N. B. Turk-Browne, P. J. Isola, B. J. Scholl, T. A. Treat, Multidimensional visual statistical learning. *J Exp Psychol: Learn. Mem. Cogn.* **34**, 399-407 (2008).
4. T. D. Wachs, *The Nature of Nurture* (SAGE, London, UK, 1992).
5. C. Thomas, C. I. Baker, Teaching and adult brain new tricks: a critical review of evidence for training-dependent structural plasticity in humans. *Neuroimage* **73**, 225-236 (2013).
6. P. J. Reber, D. R. Gitelman, T. B. Parrish, M. M. Mesulam, Dissociating explicit and implicit category knowledge with fMRI. *J Cogn. Neurosci.* **15**, 574-583 (2006).
7. B. Calvo-Merino, D. E. Glasser, J. Grèzes, R. E. Passingham, P. Haggard, Action observation and acquired motor skills: an fMRI study with expert dancers. *Cereb. Cortex* **15**, 1243-1249 (2005).
8. J. R. Folstein, T. J. Palmeri, I. Gauthier, Perceptual advantage for category-relevant perceptual dimensions: the case of shape and motion. *Front. Psychol.* 2014.01394 (2014).
9. A. Clarke, P. J. Pell, C. Ranganath, L. K. Tyler, Learning warps object representations in the ventral temporal cortex. *J. Cogn. Neurosci.***28**, 1010-1023 (2016).
10. R. Hammer, V. Sloutsky, Visual category learning results in rapid changes in brain activation reflecting sensitivity to the category relation between perceived objects and to decision correctness. *J. Cogn. Neurosci.* **28**, 1804-1819 (2016).
11. K. Shibata, T. Watanabe, Y. Sasaki, M. Kawato, Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* **334**, 1413-1415 (2011).
12. M. T. deBettencourt, J. D. Cohen, R. F. Lee, K. A. Norman, N. B. Turk-Browne, Closed-loop training of attention with real-time brain imaging. *Nat. Neurosci.* **18**, 470-475 (2015).
13. T. Ros, M. A. M. Munneke, L. A. Parkinson, J. H. Gruzelier, Neurofeedback facilitation of implicit motor learning. *Biol. Psychol.* **95**, 54-58 (2014).
14. T. Chiba, T. Kanazawa, A. Koizumi, K. Ide, V. Taschereau-Dumouchel, S. Boku, A. Hishimoto, M. Shirakawa, I. Sora, H. Law, H. Yoneda, M. Kawato, Current status of neuro-feedback for post-traumatic stress disorder: a systemic review and the possibility of decoded neurofeedback. *Front. Hum. Neurosci.* **13**, 233 (2019).
15. A. Koizumi, K. Amano, A. Cortese, K. Shibata, W. Yoshida, B. Seymour, M. Kawato, H. Lau, Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nat. Hum. Behav.* **1**, s41562-016-0006 (2016).
16. V. Taschereau-Dumouchel, A. Cortese, T. Chiba, J. D. Knotts, M. Kawato, Towards an unconscious neural reinforcement intervention for common fears. *Proc. Natl. Acad. Sci.* **115**, 3470-3475 (2018).

17. D. E. J. Linden, D. L. Turner, Real-time functional magnetic resonance imaging neurofeedback in motor neurorehabilitation. *Curr. Opin. Neurol.* **29**, 412-418 (2016).

18. K. E. MacDuffie, J. MacInnes, K. C. Dickerson, K. M. Eddington, T. J. Strauman, R. A. Adcock, Single session real-time fMRI neurofeedback has a lasting impact on cognitive behavioral therapy strategies. *Neuroimage: Clin.* **19**, 868-875 (2018).

19. H. Op de Beeck, J. Wagemans, R. Vogels, Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nat. Neurosci.* **4**, 1244-1252 (2001).

20. D. M. Drucker, G. K. Aguirre, Different spatial scales o shape similarity representation in lateral and ventral LOC. *Cereb. Cortex* **19**, 2269-2280 (2009).

21. N. Kriegeskorte, R. Goebel, P. Bandettini, Information-based functional brain mapping. *Proc. Natl. Acad. Sci.* **103**, 3863-3868 (2006).

22. Y. Shrager, C. B. Kirwan, L. R. Squire, Neural basis of the cognitive map: path integration does not require hippocampus or entorhinal cortex. *Proc. Natl. Acad. Sci.* **105**, 12034-12038 (2008).

23. R. C. Wilson, Y. K. Takahashi, G. Schoenbaum, Y. Niv, Orbitofrontal cortex as a cognitive map of task space. *Neuron* **81**, 267-279 (2014).

24. M. Reuter, N. J. Schmansky, H. D. Rosas, B. Fischl, Within-subject template estimation for u biased longitudinal image analysis. *Neuroimage* **57**, 19-21 (2012).

25. K. Grill-Spector, Z. Kourtzi, N. Kanwisher, The lateral occipital complex and its role in object recognition. *Vis. Res.* **41**, 1409-1422 (2001).

26. K. Grill-Spector, R. Malach, fMR-adaptation: a tool for studying the functional properties of human cortical neurons. *Acta Psychol.* **107**, 293-321 (2001).

27. M. M. Chun, Y. Jiang, Top-down attentional guidance based on implicit learning of visual covariation. *Psych. Sci.* **10**, 360-365 (1999).

28. D. L. Schacter, L. A. Cooper, S. M. Delaney, M. A. Peterson, M. Tharan, Implicit memory for possible and impossible objects: Constraints on the construction of structural descriptions. *J. Exp. Psych.: Gen.,* **17**, 3-19 (1991).

29. J. W. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, P. Pietrini, Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* **293**, 2425-2430 (2001).

30. N. Kriegeskorte, M. Mur, D. A. Ruff, R. Kiani, J. Bodurka, H. Esteky, K. Tanaka, P. A. Bandettini, Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126-1141 (2008).

31. M. C. Iordan, M. R. Greene, D. M. Beck, L. Fei-Fei, Basic level category structure emerges gradually across human ventral visual cortex. *J. Cogn. Neurosci.* **27**, 1427-1446 (2015).

32. Y. Sasaki, J. E. Nanez, T, Watanabe, Advances in visual perceptual learning and plasticity. *Nat Rev. Neurosci.* **11**, 53-60 (2010).

33. M. Failing, T. Feldmann-Wüstefeld, B. Wang, C. Olivers, J. Theeuwes, Statistical regularities induce spatial as well as feature-specific suppression. *J. Exp. Psychol: Hum. Percep. Perf.* **45**, 1291-1303 (2019).

34. E. F. Oblak, J. S. Sulzer, J. A. Lewis-Peacock, A simulation-based approach to improve decoded neurofeedback performance. *Neuroimage* **195**, 300-310 (2019).

35. E. F. Oblak et al., http://biorxiv.org/content/10.1101/2020.03.02.973586v2 (2020).

36. M. J. Tarr, I. Gauthier, FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nat. Neurosci.* **3**, 764-769 (2000).

37. N. Kanwisher, G. Yovel, The fusiform face area: a cortical region specialized for the perception of faces. *Phil. Trans. R. Soc. B* **361**, 2109-2128 (2006).

38. U. Goswami, Neuroscience and education: from research to practice? *Nat. Rev. Neurosci.* **7**, 406-413 (2006).

39. J. A. Bernard, R. D. Seidler, Evidence for motor cortex dedifferentiation in older adults. *Neurobiol. Aging* **33**, 1890-1899 (2012).

40. C. Cavina-Pratesi, R. W. Kentridge, C. A. Heywood, A. D. Milner, Separate processing of texture and form in the ventral stream: evidence from fMRI and visual agnosia. *Cereb. Cortex* **20**, 433-446 (2009).

41. G. Avidan, M. Behrmann, Functional MRI reveals compromised neural integrity of the face processing network in congenital prosopagnosia. *Curr. Biol.* **19**, 146-1150 (2009).

42. K. D. Young, G. J. Siegle, M. Misaki, V. Zotev, R. Phillips, et al., Altered task-based and resting-state amygdala functional connectivity following real-time fMRI amygdala neurofeedback training in major depressive disorder. *Neuroimage: Clin.* **17**, 691-703 (2018).

43. S. M. LaConte, Decoding brain states in real-time. *Neuroimage* **56**, 440-454 (2011).

44. D. E. J. Linden, I. Habes, S. J. Johnston, S. Linden, R. Tatineni, L. Subramanian, B. Sorger, D. Healy, R. Goebel, Real-time self-regulation of emotion networks in patients with depression. *Plos One* 0038115 (2012).

45. K. D. Young, V. Zotev, R. Phillips, M. Misaki, H, Yuan, W. C. Drevets, J, Bodurka, Real-time fMRI neurofeedback training of amygdala activity in patients with major depressive disorder. *Plos One* 0088785 (2014).

46. K. D. Young, G. J. Siegle, V. Zotev, R Phillips, M. Misaki, et al., Randomized clinical trial of real-time fMRI amygdala neurofeedback for major depressive disorder: effects on symptoms and autobiographical memory recall. *Am. J. Psychiatry* **174**, 748-755 (2017).

47. D. M. Schnyer, C. G. Beevers, M. T. deBettencourt, S. M. Sherman, J. D. Cohen, K. A. Norman, N. B. Turk-Browne, Neurocognitive therapeutics: from concept to application in the treatment of negative attention bias. *Biol. Mood Anx. Disord.* **5**, 1 (2015).

48. M. Ramot, S. Kimmich, J. Gonzales-Castillo, V. Roopchansingh, H. Popal, E. White, S. J. Gotts, A. Martin, Direct modulation of aberrant brain network connectivity through real-time neurofeedback. *eLife* **6**, e28974 (2017).

49. A. Ramos-Murguialday, D. Broetz, M. Rea, L. Läer, O. Yilmaz, et al., Brain-machine interface in stroke rehabilitation: a controlled study. *Ann. Neurol.* **74**, 100-108 (2013).

50. D. Borton, S. Micera, J. R. Millán, G. Courtine, Personalized neuroprosthetics. *Science Transl. Med.* **5**, 210rv2210rv2 (2013).

51. P. Kok, N. B. Turk-Browne, Associative prediction of visual shape in the Hippocampus. *J. Neurosci.* **38**, 6888-6899 (2018).

52. D. H. Brainard, The psychophysics toolbox. *Spat. Vis.* **10**, 433-436 (1997).

53. R. W. Cox, AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Comput. Biomed. Res.* **29**, 162-173 (1996).

54. D. A. Handwerker, J. Gonzales-Castillo, M. D'Esposito, P.A. Bandettini, The continuing challenge of understanding and modeling hemodynamic variation in fMRI. *Neuroimage* **62**, 1017-1023 (2012).

55. M. P. Walker, R. Stickgold, Sleep-dependent learning and memory consolidation. *Neuron* **44**, 121-133 (2004).

Supplementary Information for

Sculpting New Visual Categories into the Human Brain

Marius Cătălin Iordan, Victoria J. H. Ritvo, Kenneth A. Norman, Nicholas B. Turk-Browne, Jonathan D. Cohen

Correspondence to: mci@rochester.edu

Contents:

Supplementary Figures S1-S34

Supplementary Tables S1-S6

Supplementary Movies S1-S2

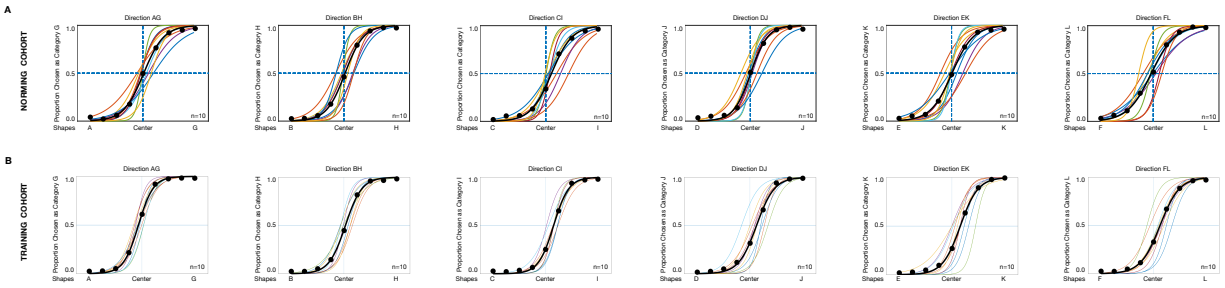


Fig. S1. Stimulus space was perceived identically by norming and training cohorts of participants.

Psychometric functions estimated for each of the six directions in the stimulus space (AG, BH, CI, DJ, EK, FL) for the norming cohort (A, n=10, replicated from Fig. 1B-G for convenience) and the training cohort (B, n=10, bottom). Color lines indicate individual participants and bolded line indicates average. No significant differences were observed between the two cohorts or between the six directions (psychometric function slope, repeated measures ANOVA, factors=‘direction’, ‘cohort’; interaction: $F(5,108)=0.19$, $p=0.964$; main effect of direction: $F(5,108)=0.57$, $p=0.722$; main effect of cohort: $F(1,108)=1.71$, $p=0.194$).

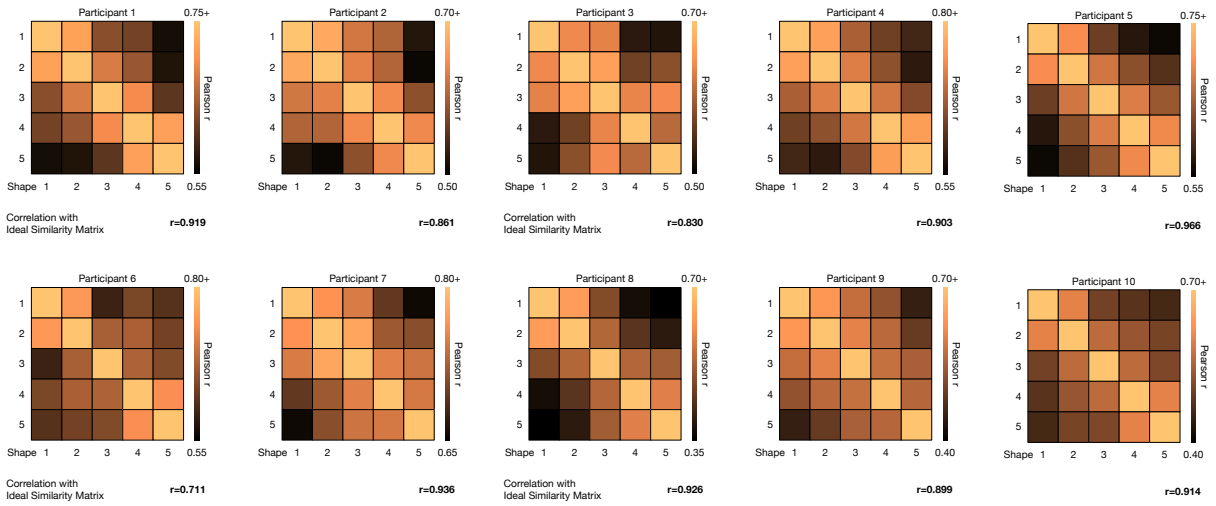


Fig. S2. Shape space is represented parametrically in neurofeedback ROI.

Representational similarity matrices (RSM) obtained by computing the Pearson correlation between neural activity elicited by pairs of shapes spanning the stimulus space (Fig. 2B) in the neurofeedback ROI of each participant during the localizer scans (Days 2-3). Correlation with the ideal representational similarity matrix RSM (Fig. 2C) is shown below each participant RSM.

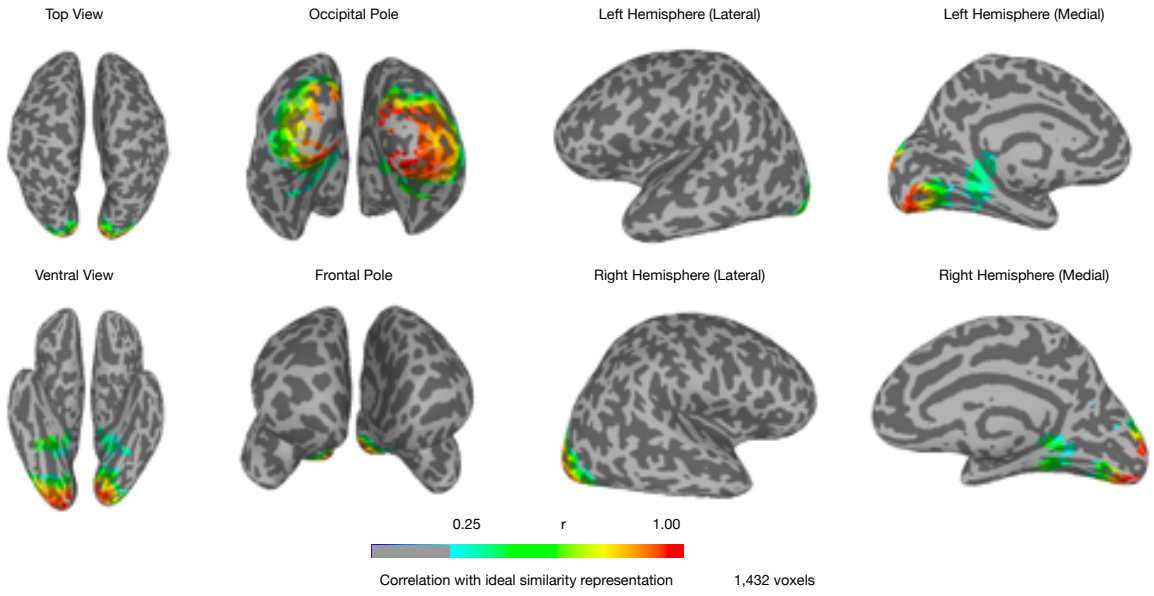


Fig. S3. Neurofeedback ROI for participant #1.

Searchlight analysis indicating all of participant #1's brain regions that represented the shape space parametrically, including extrastriate visual cortex, parahippocampal gyrus, and hippocampus. See Methods for details of the searchlight procedure and threshold selection.

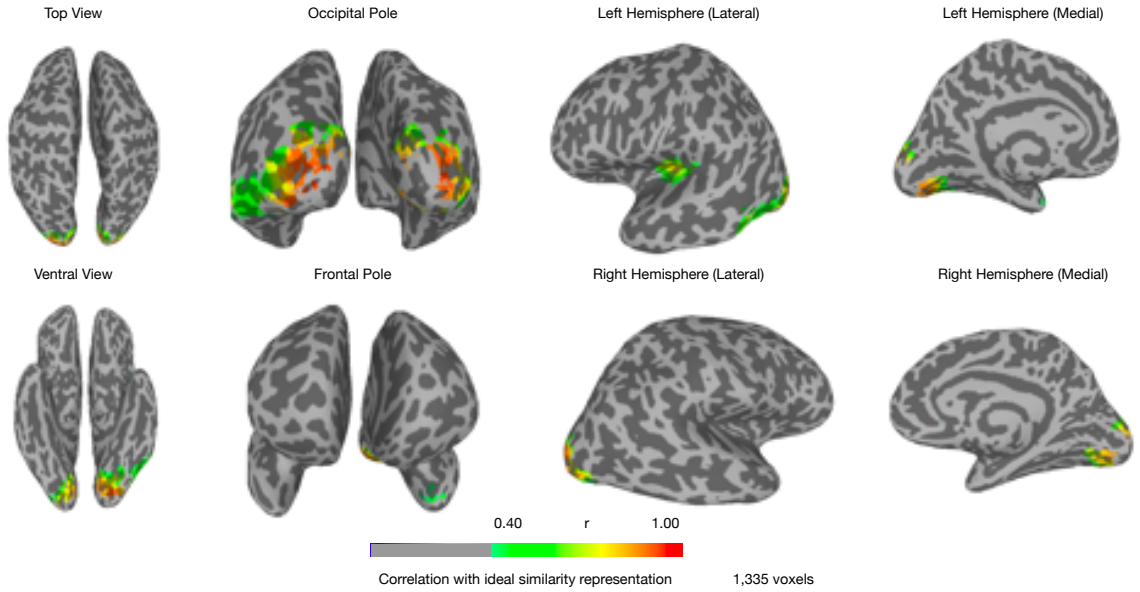


Fig. S4. Neurofeedback ROI for participant #2.

Searchlight analysis indicating all of participant #2's brain regions that represented the shape space parametrically, including extrastriate visual cortex, inferior temporal cortex, anterior temporal cortex, and superior temporal sulcus. See Methods for details of the searchlight procedure and threshold selection.

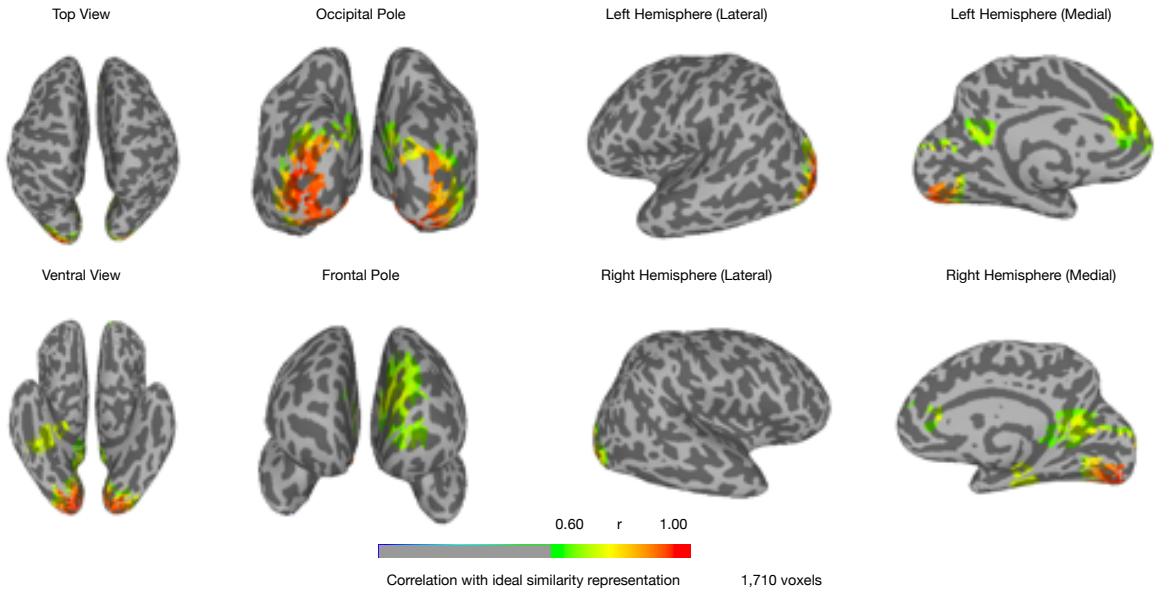


Fig. S5. Neurofeedback ROI for participant #3.

Searchlight analysis indicating all of participant #3's brain regions that represented the shape space parametrically, including extrastriate visual cortex, parahippocampal gyrus, hippocampus, and medial frontal gyrus. See Methods for details of the searchlight procedure and threshold selection.

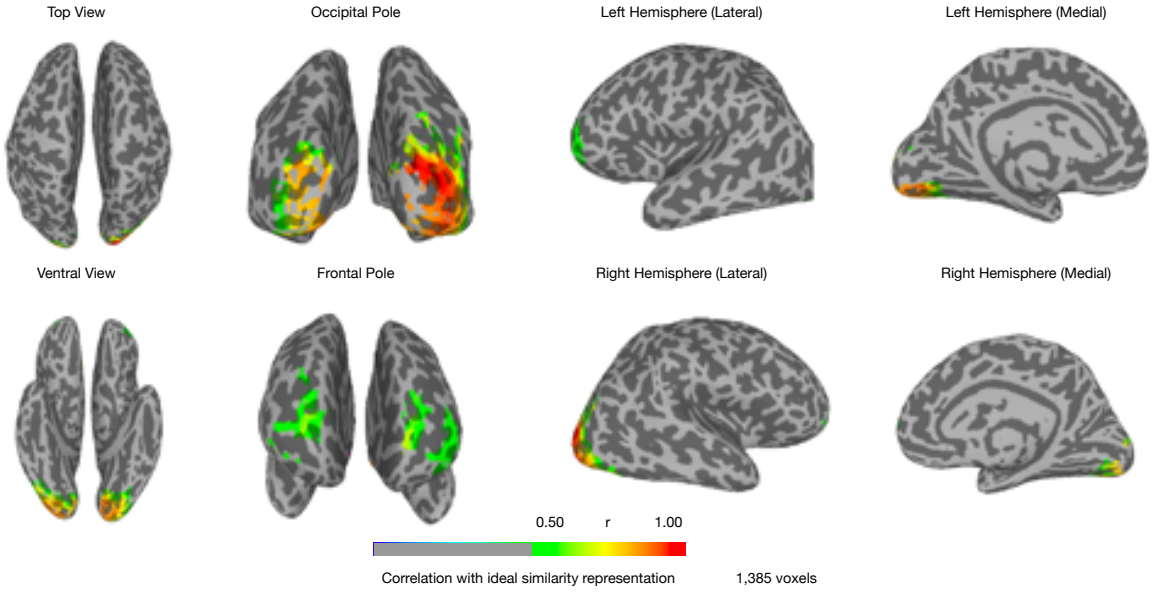


Fig. S6. Neurofeedback ROI for participant #4.

Searchlight analysis indicating all of participant #4's brain regions that represented the shape space parametrically, including extrastriate visual cortex, transverse occipital sulcus, and dorsolateral prefrontal cortex. See Methods for details of the searchlight procedure and threshold selection.

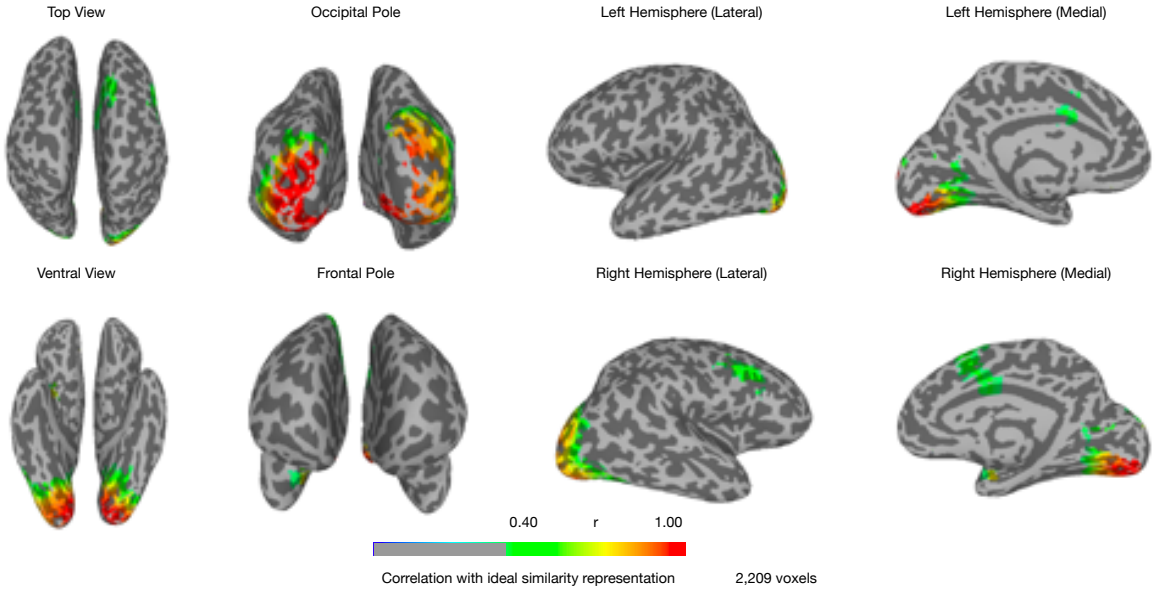


Fig. S7. Neurofeedback ROI for participant #5.

Searchlight analysis indicating all of participant #5's brain regions that represented the shape space parametrically, including extrastriate visual cortex, inferior temporal cortex, parahippocampal gyrus, and dorsolateral prefrontal cortex. See Methods for details of the searchlight procedure and threshold selection.

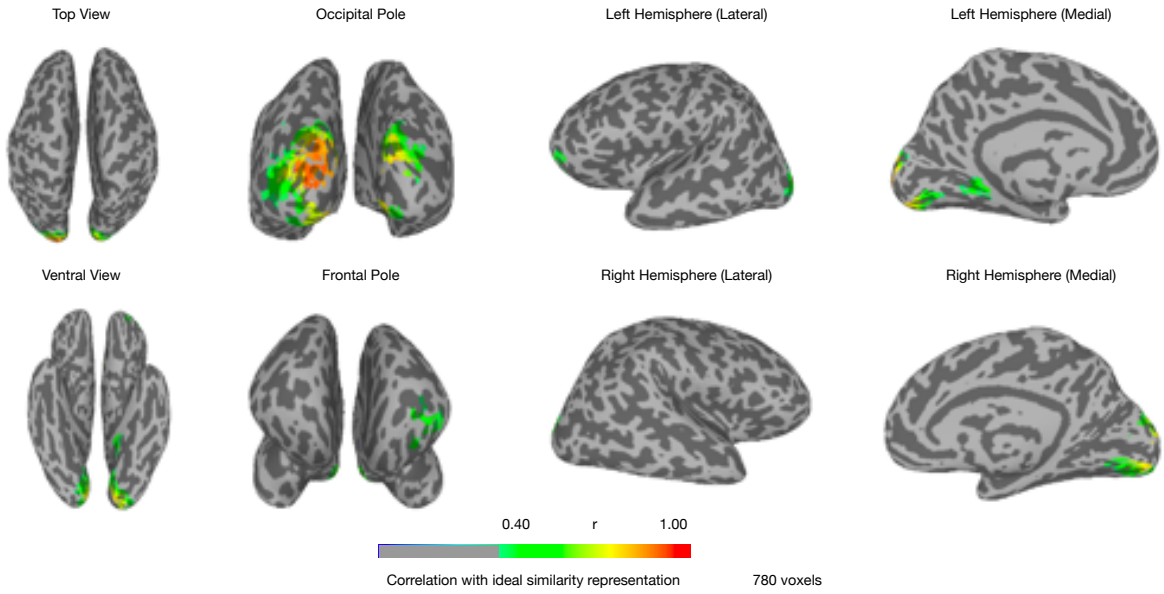


Fig. S8. Neurofeedback ROI for participant #6.

Searchlight analysis indicating all of participant #6's brain regions that represented the shape space parametrically, including extrastriate visual cortex and orbitofrontal cortex. See Methods for details of the searchlight procedure and threshold selection.

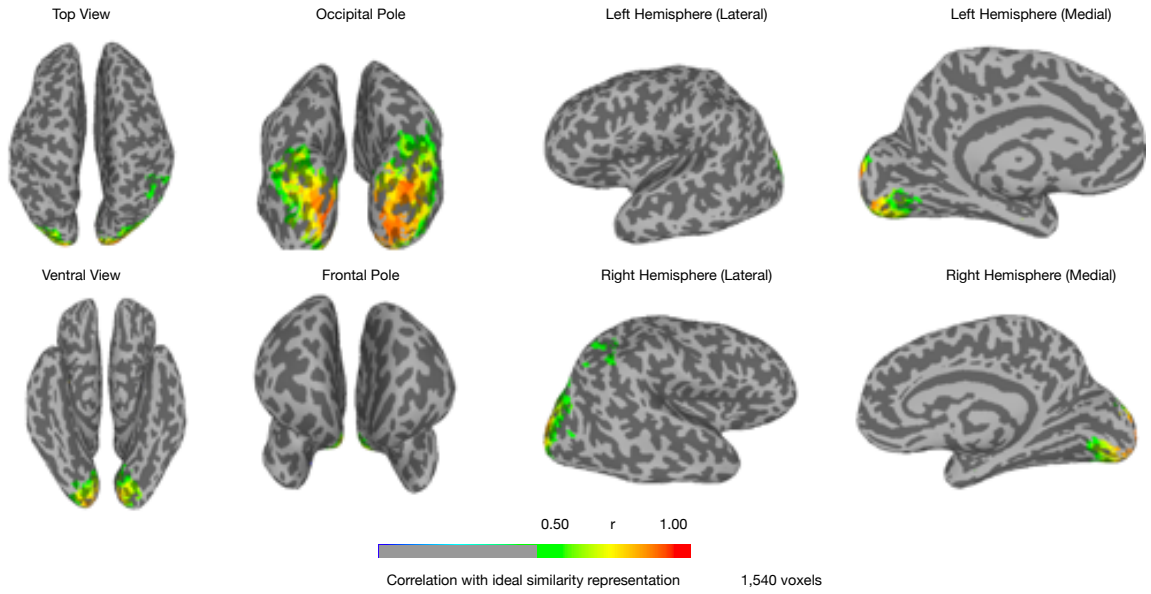


Fig. S9. Neurofeedback ROI for participant #7.

Searchlight analysis indicating all of participant #7's brain regions that represented the shape space parametrically, including extrastriate visual cortex and parietal cortex. See Methods for details of the searchlight procedure and threshold selection.

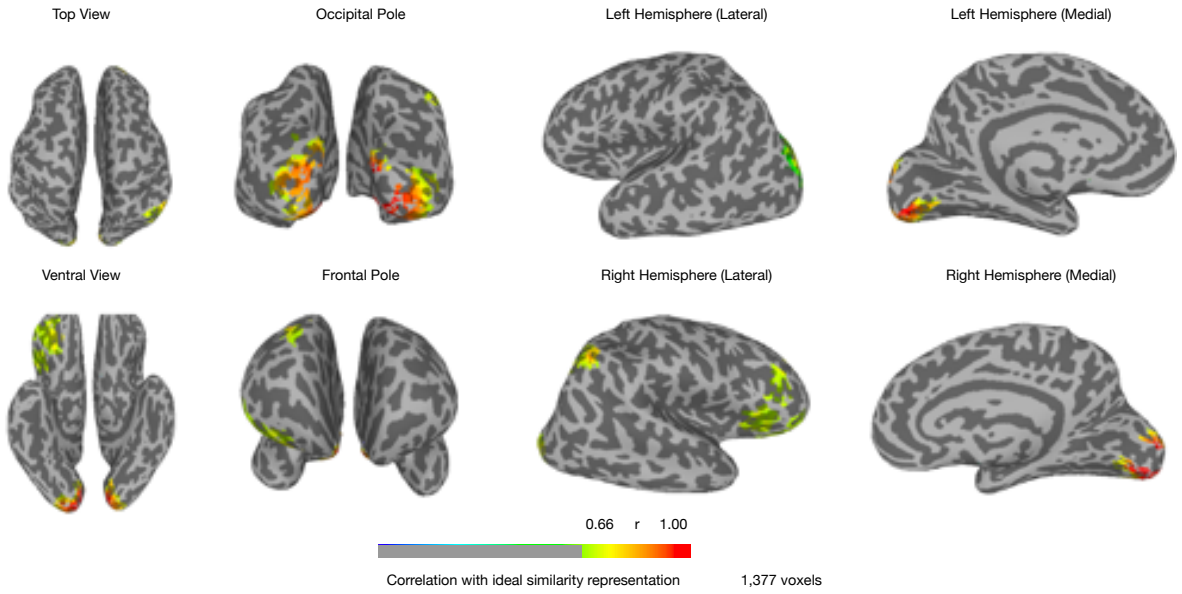


Fig. S10. Neurofeedback ROI for participant #8.

Searchlight analysis indicating all of participant #8's brain regions that represented the shape space parametrically, including extrastriate visual cortex, parietal cortex, and dorsolateral prefrontal cortex. See Methods for details of the searchlight procedure and threshold selection.

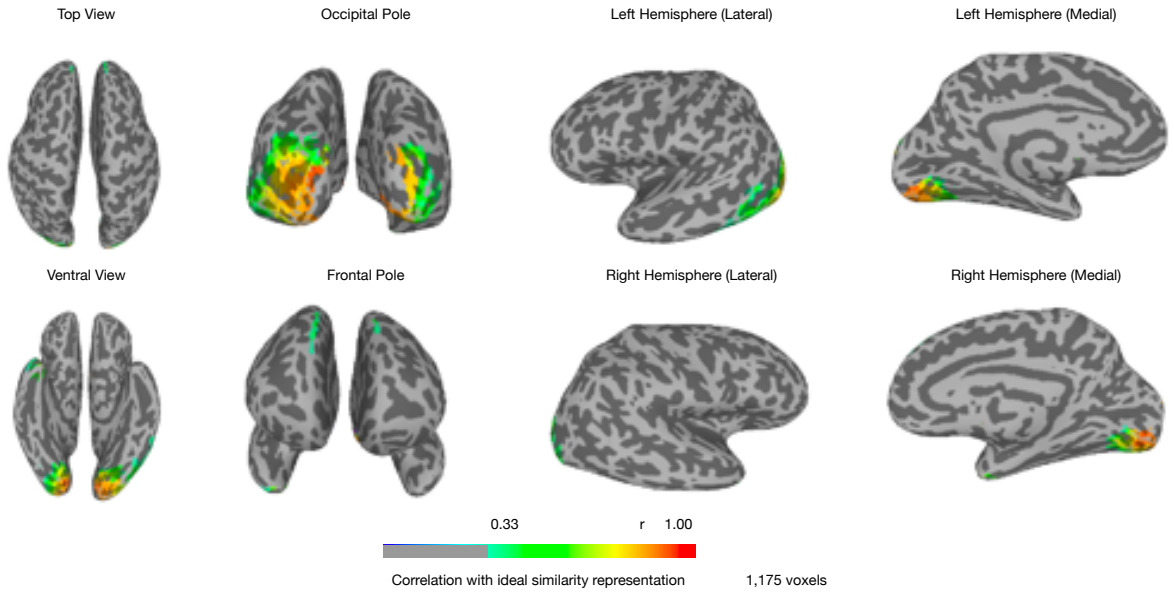


Fig. S11. Neurofeedback ROI for participant #9.

Searchlight analysis indicating all of participant #9's brain regions that represented the shape space parametrically, including extrastriate visual cortex, transverse occipital sulcus, inferior parietal cortex, and dorsolateral prefrontal cortex. See Methods for details of the searchlight procedure and threshold selection.

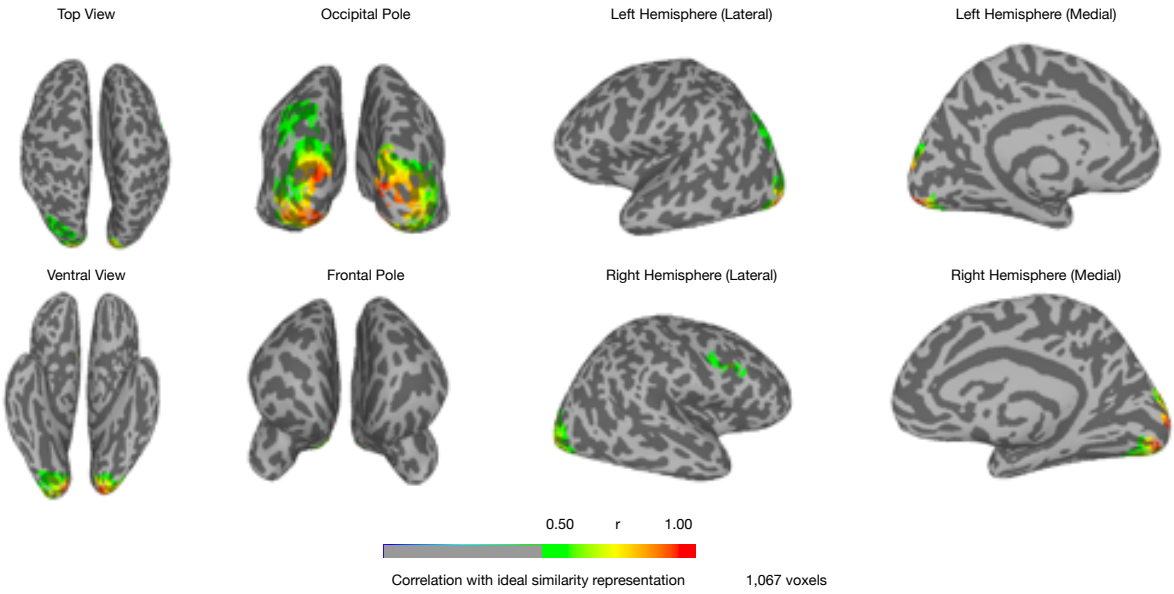


Fig. S12. Neurofeedback ROI for participant #10.

Searchlight analysis indicating all of participant #10's brain regions that represented the shape space parametrically, including extrastriate visual cortex, inferior temporal cortex, anterior temporal cortex, and dorsolateral prefrontal cortex. See Methods for details of the searchlight procedure and threshold selection.

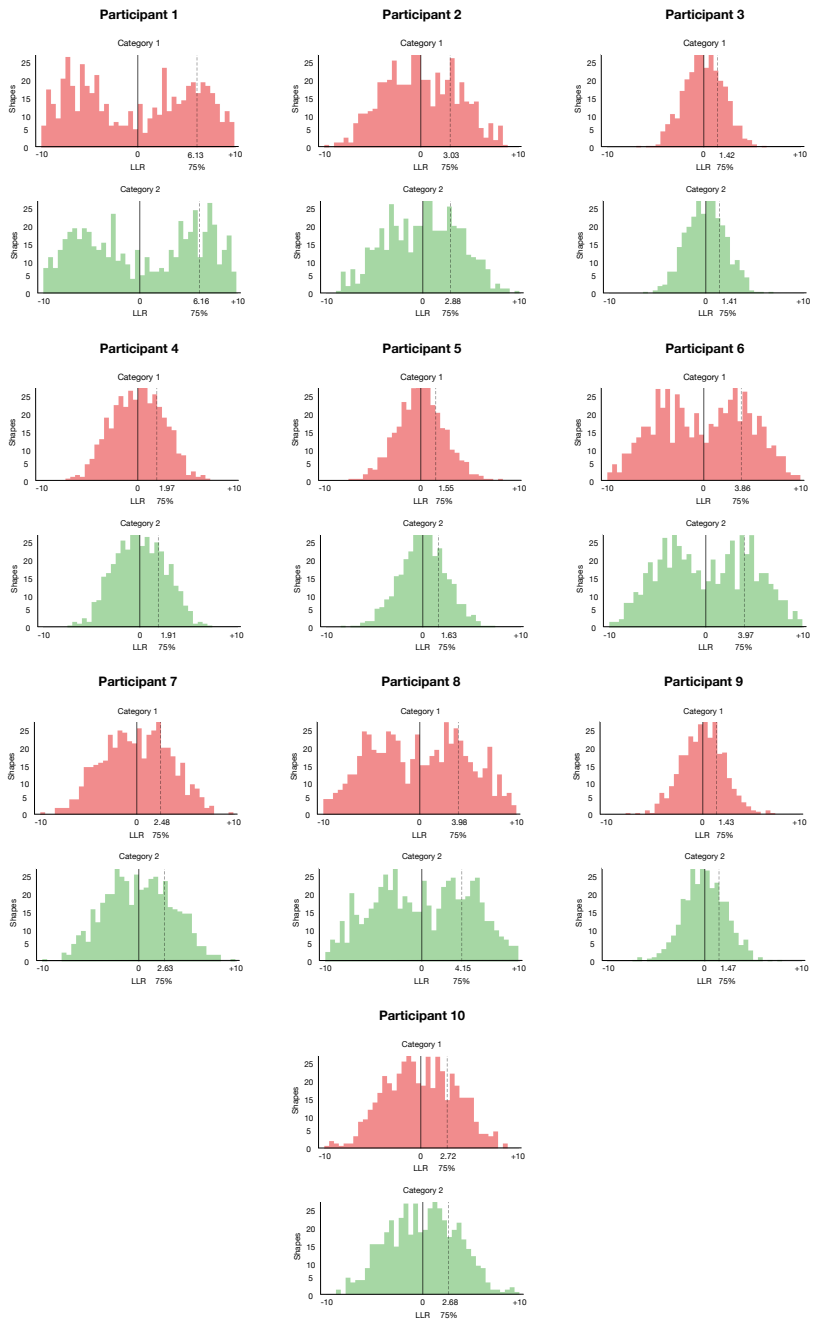


Fig. S13. LLR histograms for shapes shown during localizer scans (days 2-3).

We computed a distribution of log-likelihood ratios (LLR) using the estimated neural model Gaussian distributions for 48 out of the 49 shapes shown to participants (all except the ambiguous center shape) during each fMRI run scanned during the Days 2-3 fMRI sessions (12-15 runs per participant, 576-720 examples per category per participant). The full histogram of the LLR distributions are shown for each participant and each category (by definition, LLR1 = - LLR2), together with a representative threshold of 75% of the distribution used during the neurofeedback trials to assess performance and give feedback to the participants for each category.

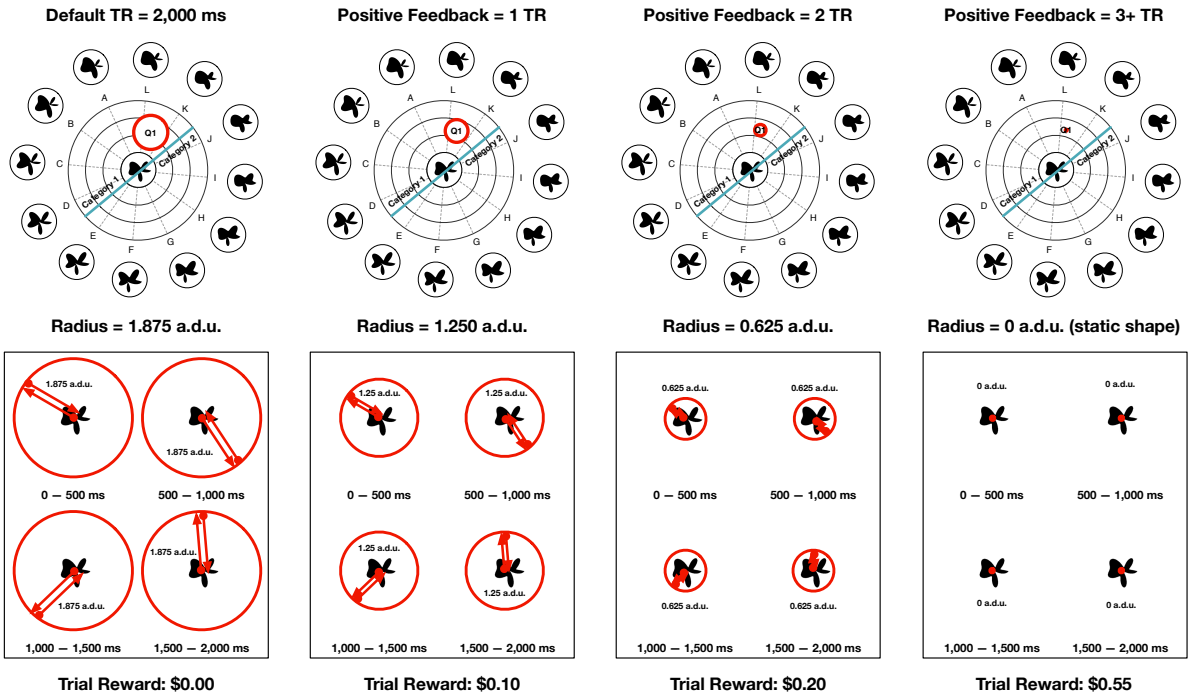


Fig. S14. Neurofeedback trials with varying amounts of feedback.

The four types of TRs shown to participants during the neurofeedback training. At the beginning of each trial, a random shape was selected and made to oscillate with a radius of 1.875 a.d.u. (default TR) (the shape position was static on the screen, but the shape identity morphed gradually according to the trajectory shown in parameter space). If a participant accumulated positive feedback within that trial (cf. Fig. 3D), the radius of the oscillation was reduced in increments of 0.625 a.d.u., otherwise it was left unchanged from the previous TR. Each 2s TR comprised four independent shape oscillations of 500 ms each (see Movie S1 for two example trials). Participants also received a monetary reward for each trial in which they were able to reduce the radius of the oscillation.

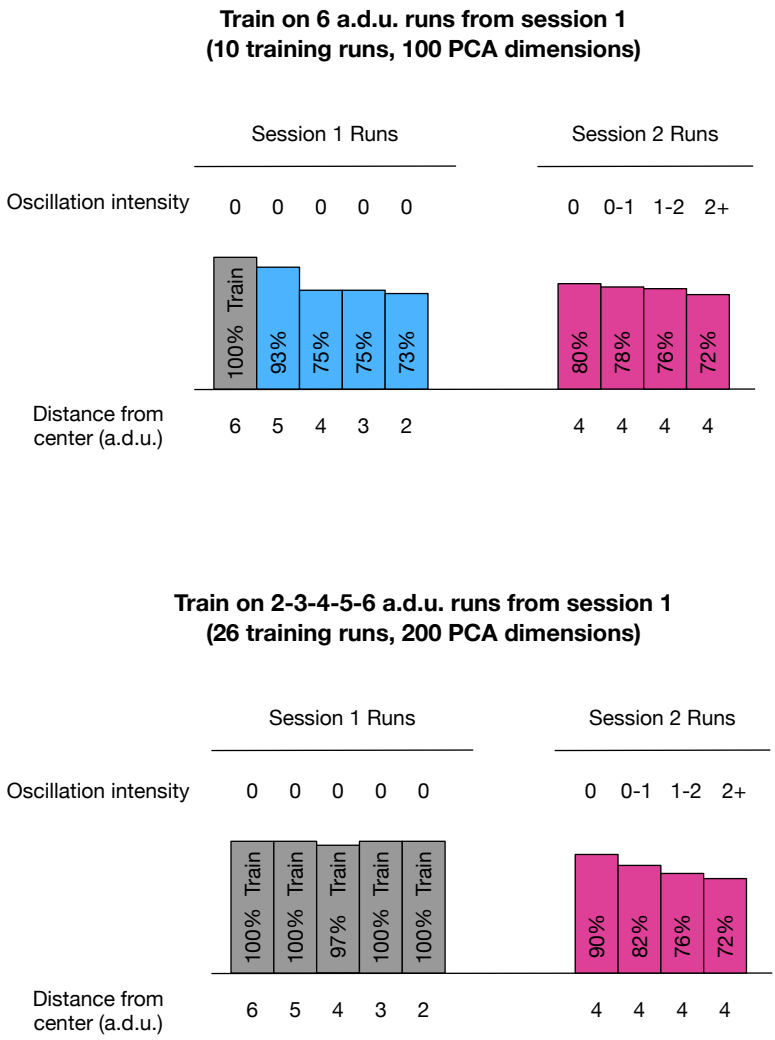


Fig. S15. Gaussian model decoding performance decays closer to the category boundary and with increased oscillation radius.

We collected two fMRI sessions for a pilot participant who performed the localizer task during the first session (four scans on separate days, 26 runs, 20 shapes of varying distance from center per run) and a center dot color change task during the fifth session (8 runs of 20 shapes similar to neurofeedback design, of varying levels of oscillation radius). Gray bars show classifier training accuracy. Blue bars show classifier testing accuracy for runs with varying distance from stimulus space center in session 1. Pink bars show classifier testing accuracy for runs with varying oscillation radius from session 2. Performance of the model decayed for shapes closer to the category boundary (top left) and when the model was trained on static shapes (e.g., localizer runs) and tested on oscillating shapes (e.g., neurofeedback runs). Based on these data, we chose an upper limit of 1.875 a.d.u. for the shape oscillation during neurofeedback trials and did not train shapes that were less than 1 a.d.u. from the category boundary, as the performance of the model was not reliable for combinations of parameters outside of these bounds.

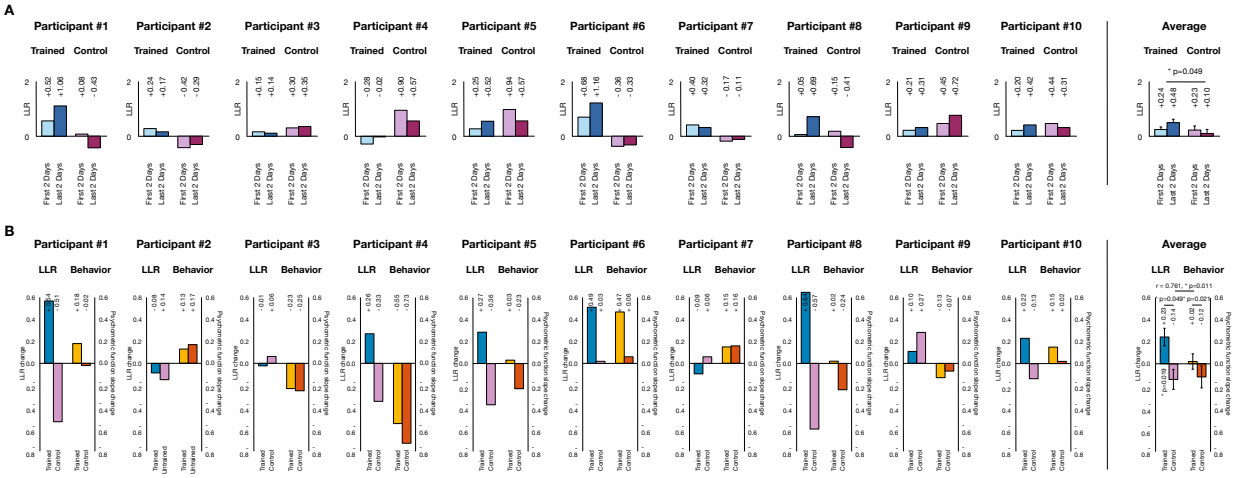


Fig. S16. Neural and perceptual effects of neural sculpting for individual participants.

(A) LLR for each participant's trained (blue) and control (purple) categories during the first two days (light color) and last two days (dark color). We observed a significant interaction effect between training and category (ANOVA; $F(1,76)=4.92$, $p=0.030$); (B) LLR for the trained (blue) and control (purple) categories, plotted together with the psychometric function changes for the trained (light orange) and control (dark orange) categories. Difference between LLR bars yields graph in Fig. 4D and difference between Behavior bars yields graph in Fig. 4E.

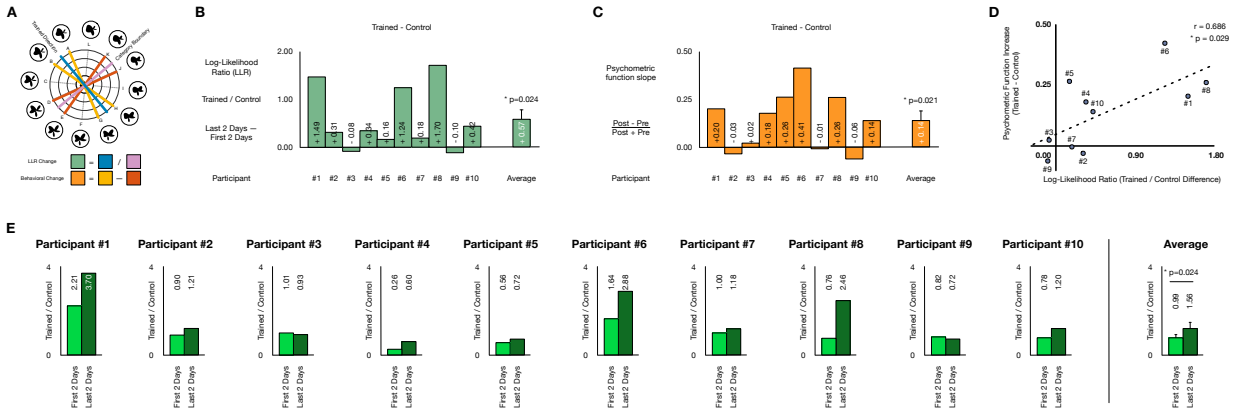


Fig. S17. Neural sculpting effects and correlation with behavior using LLR ratio are similar to those using a difference score.

(A) Diagram of stimulus space for example participant with emphasized trained direction, category boundary (for LLR computation), and directions used to compute psychometric function slope changes in the behavioral experiments. LLR Ratio (green) represents the LLR in the Trained direction divided by the LLR in the control direction. (B) Main effect of LLR Ratio change as a consequence of neurofeedback training (LLR Ratio difference between first two days and last two days of training). (C) Main behavioral effect of neurofeedback training (replicated from Fig. 4E for convenience). (D) Correlation plot of LLR ratio main effect vs. behavioral main effect. Similar to the LLR difference score measure, we observed a strong correlation between the neural change and perceptual change (r=0.686, p=0.029). (E) LLR Ratio for each participant during their first two days and last two days of neurofeedback training. We observed a significant increase of LLR ratio at the end of the experiment compared to the beginning (p=0.024).

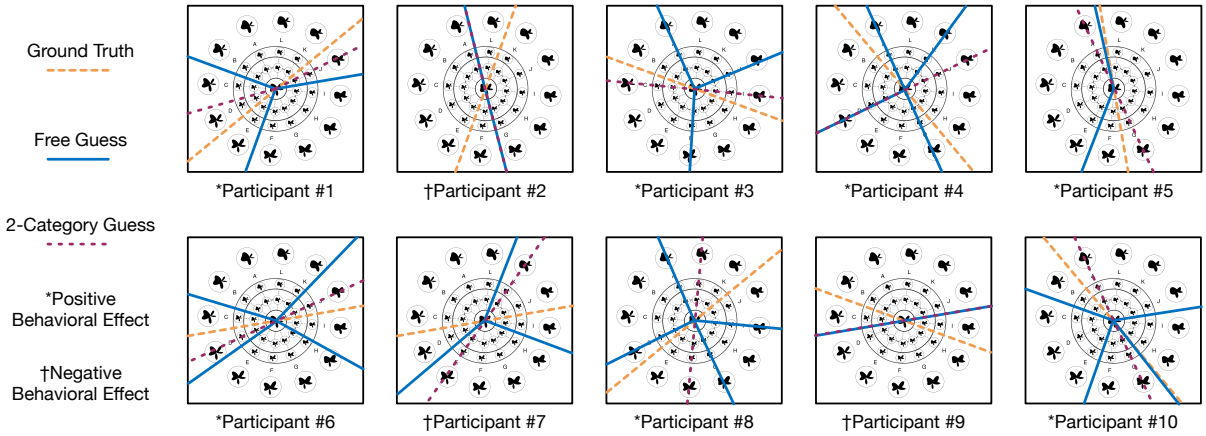


Fig. S18. Category boundary guesses before final debrief.

At the conclusion of the experiment (but before being fully debriefed), participants were told that the shapes they had seen during training belonged to an unspecified number of categories. They were instructed to draw their best guess for the category boundary/boundaries between them. Participants reported an average of 3.3 categories, whose boundaries (blue lines) did not coincide with the category boundaries randomly selected for each of them in the experiment (orange lines). When subsequently forced to split the stimulus space into 2 categories using a straight line, their performance suggested a trend towards being able to access some information about the category boundary given an explicit external constraint (purple line, trained=0°, average absolute displacement=31.5° ± 5.89° s.e.m,; random choice=45°, $t(9)=2.18$, $p=0.058$).

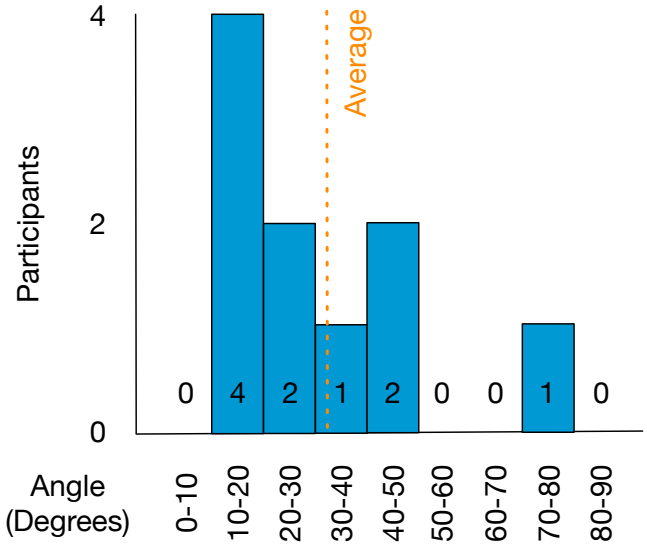


Fig. S19. Histogram of forced 2-category boundary guesses.

Histogram of angle displacement from correct category boundary for forced 2-category guesses in Fig. S18 (trained=0°, average absolute displacement=31.5° ± 5.89°; control=90°, average absolute displacement=58.5° ± 5.89°). Participants' performance suggests a trend towards having some crude information about the category boundary given an explicit external constraint (random choice=45°, t(9)=2.18, p=0.058).

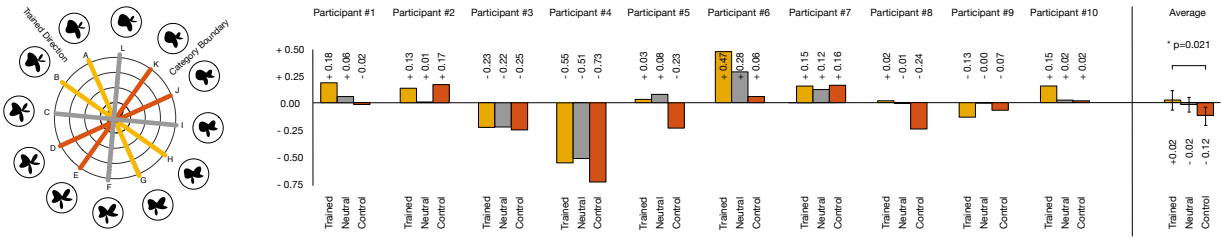


Fig. S20. Neutral direction in shape space was not strongly affected by neurofeedback training.

As a control analysis, we computed the psychometric function slope change for the average of the two directions that were midway between parallel and perpendicular to the trained category boundary (in gray, neutral, e.g., for categories ABCDEF and GHIJKL, the average slope of lines BH and EK). The behavioral change in the neutral direction was almost absent (-0.017±0.064 s.e.m., $t(9)=0.257$, $p=0.803$) and this change was not significantly different from that observed in the trained or control directions, albeit potentially more similar to the former than the latter (neutral<trained $t(9)=1.26$, $p=0.239$; neutral>control $t(9)=2.08$, $p=0.067$). The trained and control bar plots are replicated from Fig.S16B for convenience.

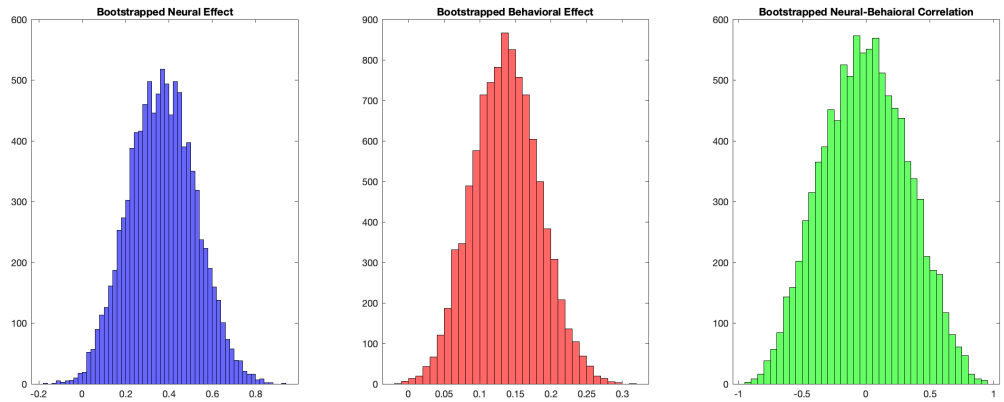


Fig. S21. Bootstrap resampling analysis showed that our results were robust at the individual level.

A sample-with-replacement procedure was used to generate 10,000 random draws of 10 samples (i.e., 10 participants) from the distributions of observed neural and behavioral effects (i.e., LLR difference values and psychometric function normalized difference scores reported in Fig. 4). The histograms of mean observed effects across all samples are shown in blue (neural LLR change), red (behavioral effect), and green (correlation values between each of the neural and behavioral samples). The probability of observing a positive neural effect by chance given our cohort is p(neural>0) = 0.005; the probability of observing a positive behavioral effect by chance given our cohort is p(behavioral>0) < 0.001; the probability of observing a correlation of equivalent magnitude to our reported effect (r=0.761) by chance given the distributions of neural and behavioral effects in our cohort is p(correlation observed>sampled correlation) = 0.006.

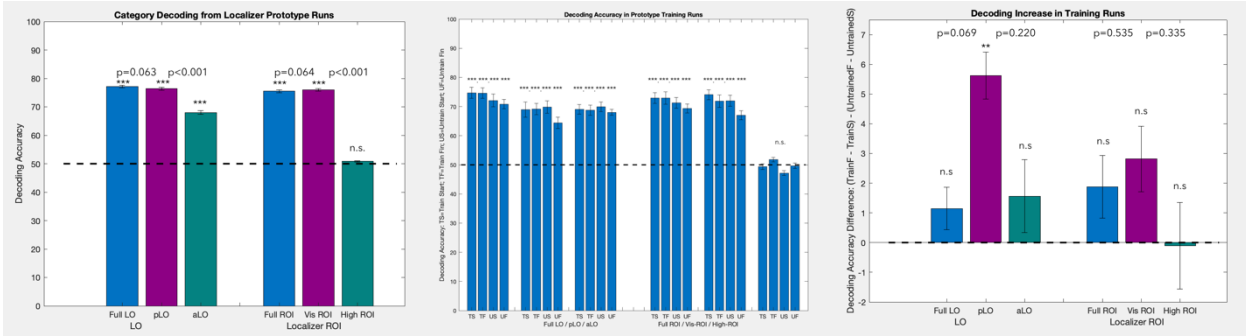


Fig. S22. Split ROI analyses: anterior vs. posterior LO and visual vs. "high-level" ROI.

To measure the strength of category representations within sub-regions of our ROIs, we (1) split LO into equal sized anterior (aLO) and posterior (pLO) halves using the Z coordinate of the Freesurfer map from which the ROI was defined; and (2) split the neurofeedback ROI into "visual" (vis-ROI, all voxels within occipito-temporal cortex) and "high-level" (high-ROI, all remaining ROI voxels outside occipito-temporal cortex) components. We trained new category decoding models within these pairs of sub-ROIs for the runs collected during the two localizer scans (Days 2—3, 12-15 runs per participant). (Left) Decoding accuracy in aLO was markedly lower compared to pLO or to the full LO ROI ($p<0.001$). Furthermore, in the neurofeedback ROI, we could decode categories only in vis-ROI, but not in high-ROI. (Center) To quantify changes in decoding accuracy due to training, we computed the decoding accuracy of these models for the trained (T) and control (U) directions, across the calibration runs collected during first two days (TS, US) and the last two days (TF, UF) of training. In the trained direction, numerically, the decoding accuracy usually remained the same in all regions between the beginning and the end of training. In the control direction, however, accuracy numerically dropped after training, which is consistent with the main results, where we saw feature suppression in the control direction, rather than feature enhancement of the trained direction. While all accuracies were significantly higher than chance individually (except for high-ROI,) they were not significantly different from each other between TS vs. TF or US vs. UF. The only two exceptions were US vs. UF in pLO and in vis-ROI, which is consistent with the main results. (Right) We computed the difference of differences for the data shown in (Center). Numerically, we observed a trend for more posterior sub-regions to undergo more change due to training (i.e., pLO and vis-ROI). However, the only significant effect of decoding accuracy increase after training was observed in pLO ($p=0.004$) and no significant changes were present between any of the regions (although a trend, $p=0.069$ was observed between pLO and the full LO region).

Fig, S23. Psychometric functions for all participants and directions.
To show that no categorization biases exist not just at the cohort level (Fig. S1), but also at the individual participant and direction level, we computed the psychometric functions for each individual direction tested in the shape space (AG, BH, CI, DJ, EK, FL) for each individual participant. No biases were observed for the trained of the control directions in the stimulus space at the individual level.

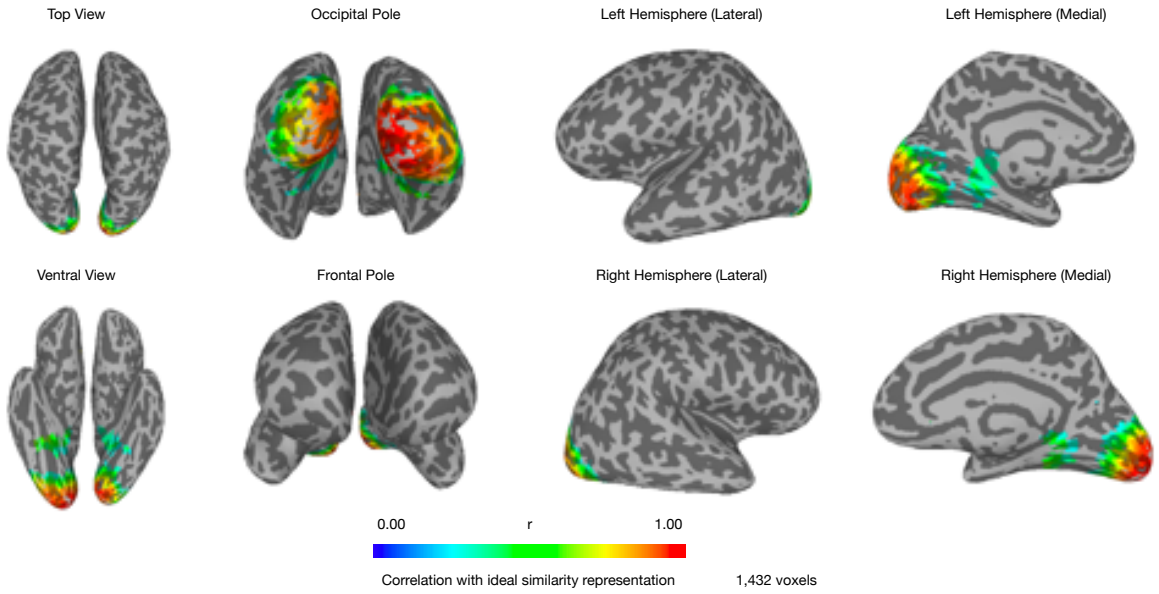


Fig. S24. Neurofeedback ROI for participant #1 (including early visual cortex).
Searchlight analysis indicating all of participant #1's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S3 shows the same brain regions, except without early visual cortex).

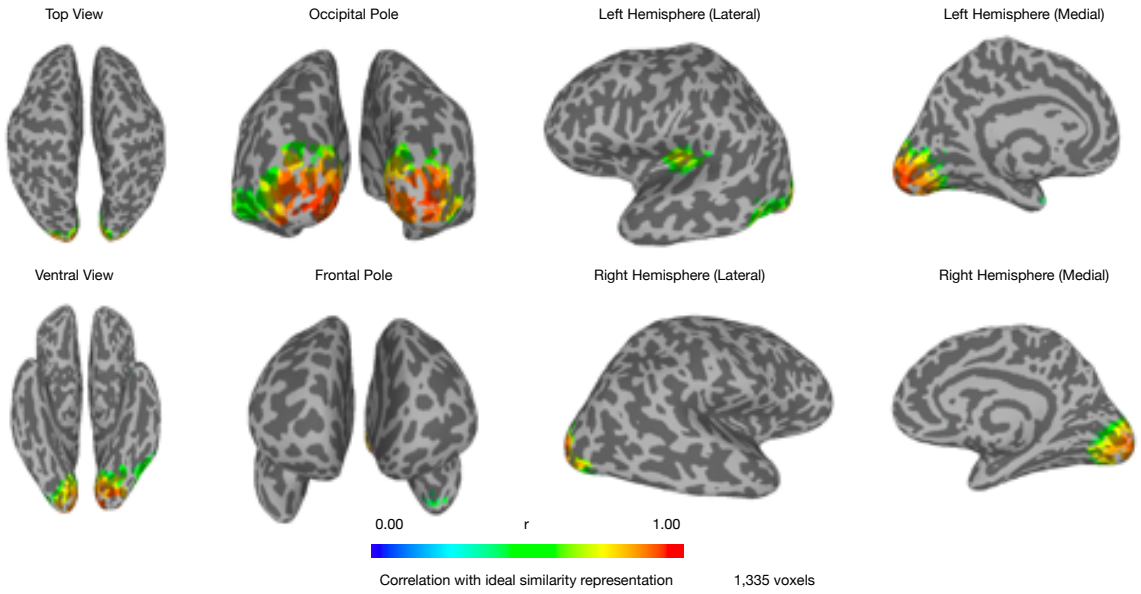


Fig. S25. Neurofeedback ROI for participant #2 (including early visual cortex).
Searchlight analysis indicating all of participant #2's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S4 shows the same brain regions, except without early visual cortex).

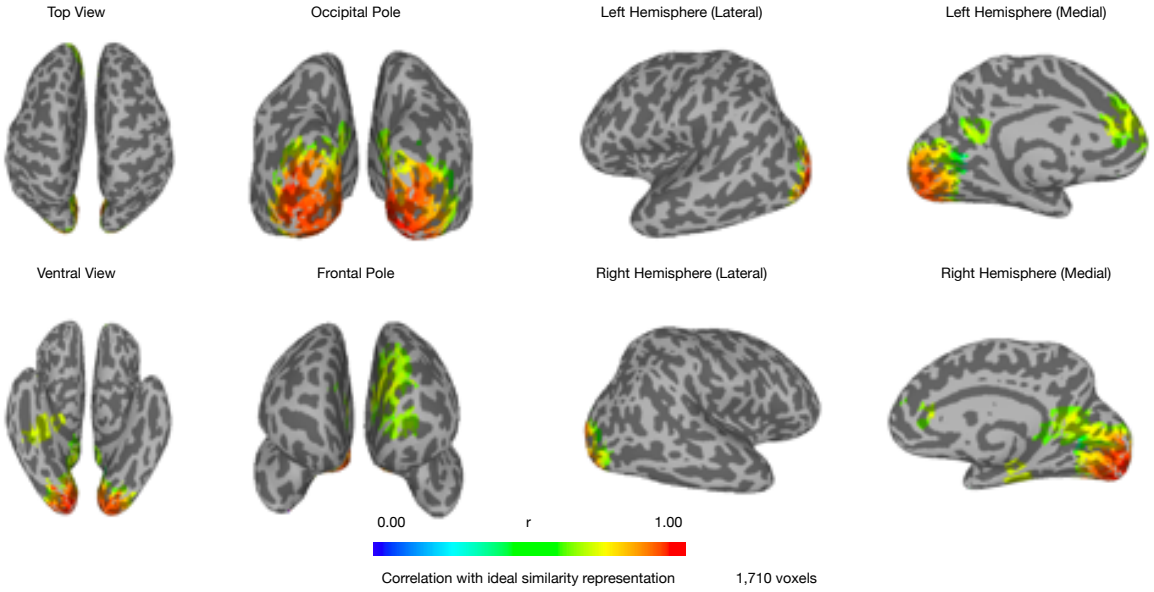


Fig. S26. Neurofeedback ROI for participant #3 (including early visual cortex).
Searchlight analysis indicating all of participant #3's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S5 shows the same brain regions, except without early visual cortex).

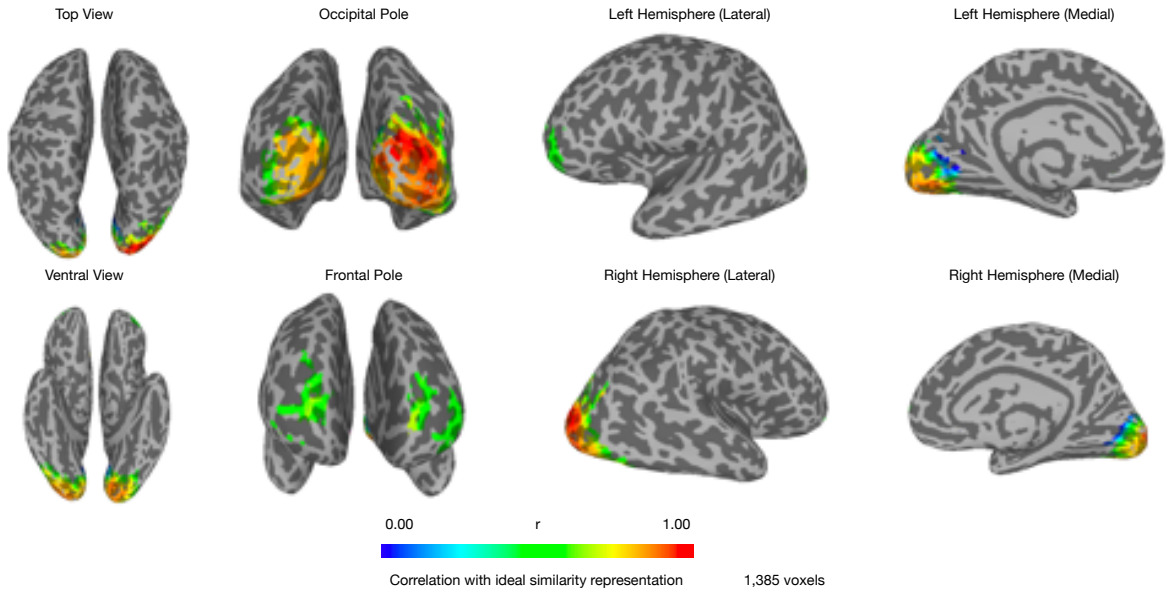


Fig. S27. Neurofeedback ROI for participant #4 (including early visual cortex).
Searchlight analysis indicating all of participant #4's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S6 shows the same brain regions, except without early visual cortex).

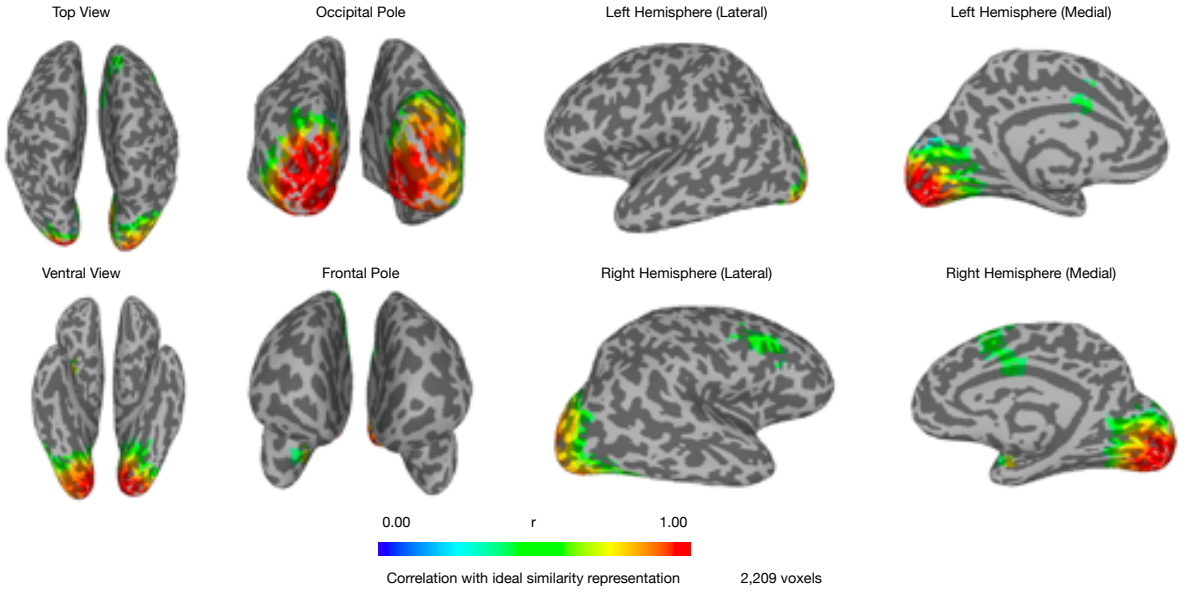


Fig. S28. Neurofeedback ROI for participant #5 (including early visual cortex).
Searchlight analysis indicating all of participant #5's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S7 shows the same brain regions, except without early visual cortex).

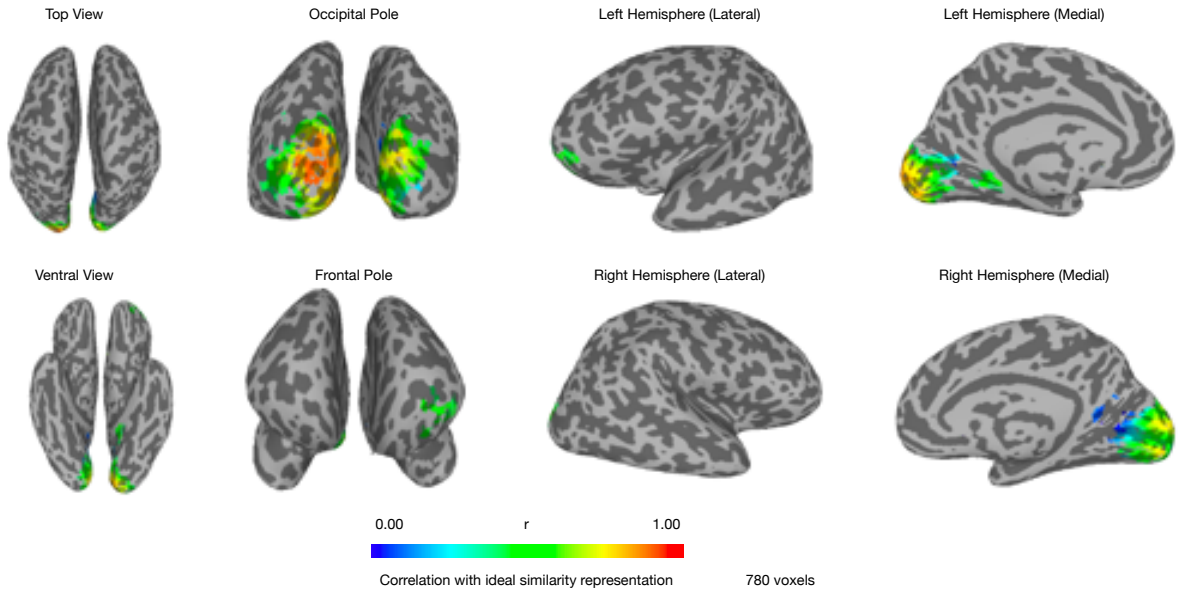


Fig. S29. Neurofeedback ROI for participant #6 (including early visual cortex).
Searchlight analysis indicating all of participant #6's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S8 shows the same brain regions, except without early visual cortex).

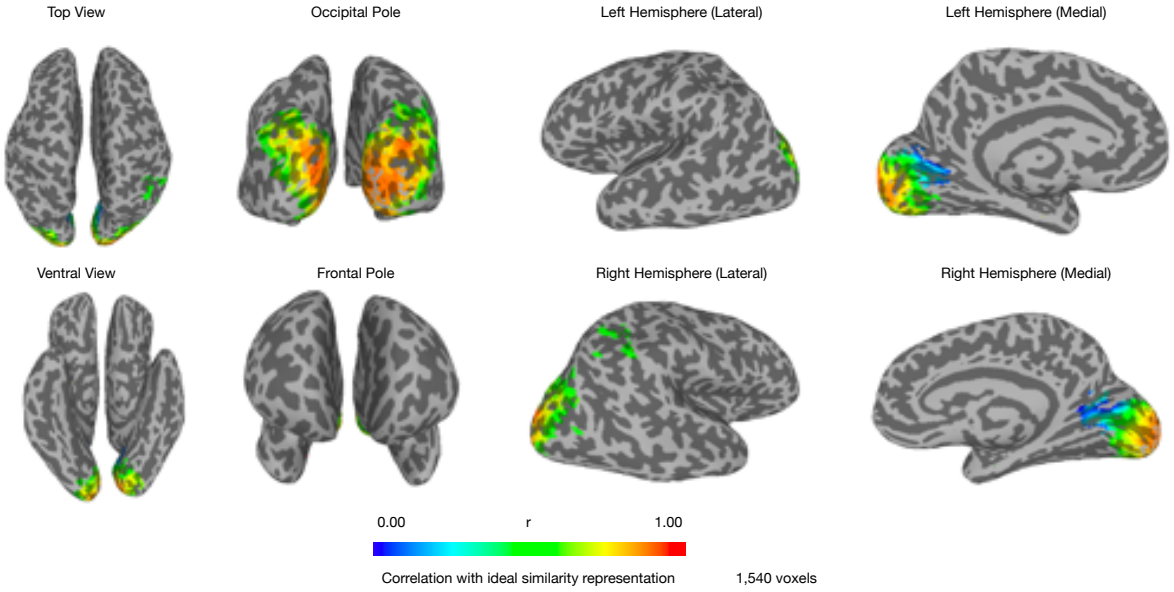


Fig. S30. Neurofeedback ROI for participant #7 (including early visual cortex).
Searchlight analysis indicating all of participant #7's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S9 shows the same brain regions, except without early visual cortex).

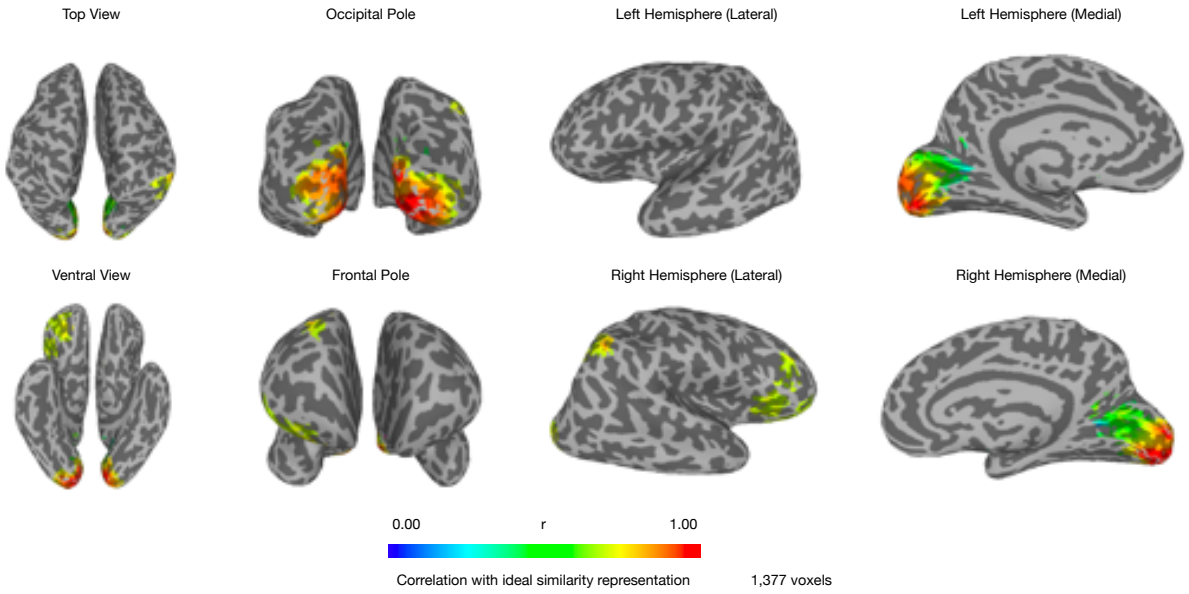


Fig. S31. Neurofeedback ROI for participant #8 (including early visual cortex).
Searchlight analysis indicating all of participant #8's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S10 shows the same brain regions, except without early visual cortex).

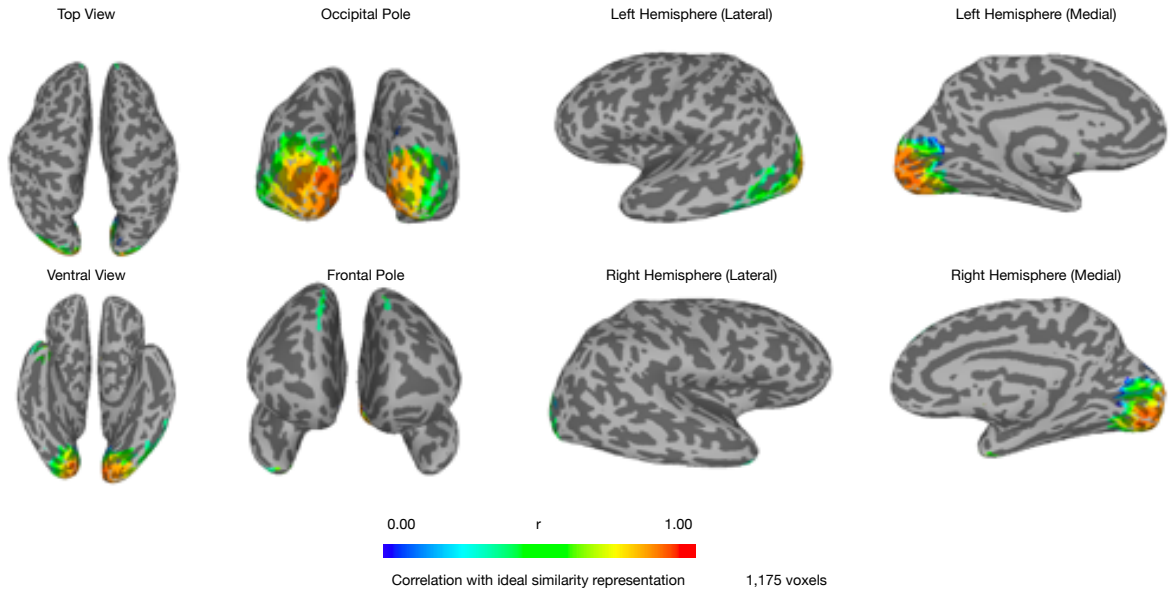


Fig. S32. Neurofeedback ROI for participant #9 (including early visual cortex).
Searchlight analysis indicating all of participant #9's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S11 shows the same brain regions, except without early visual cortex).

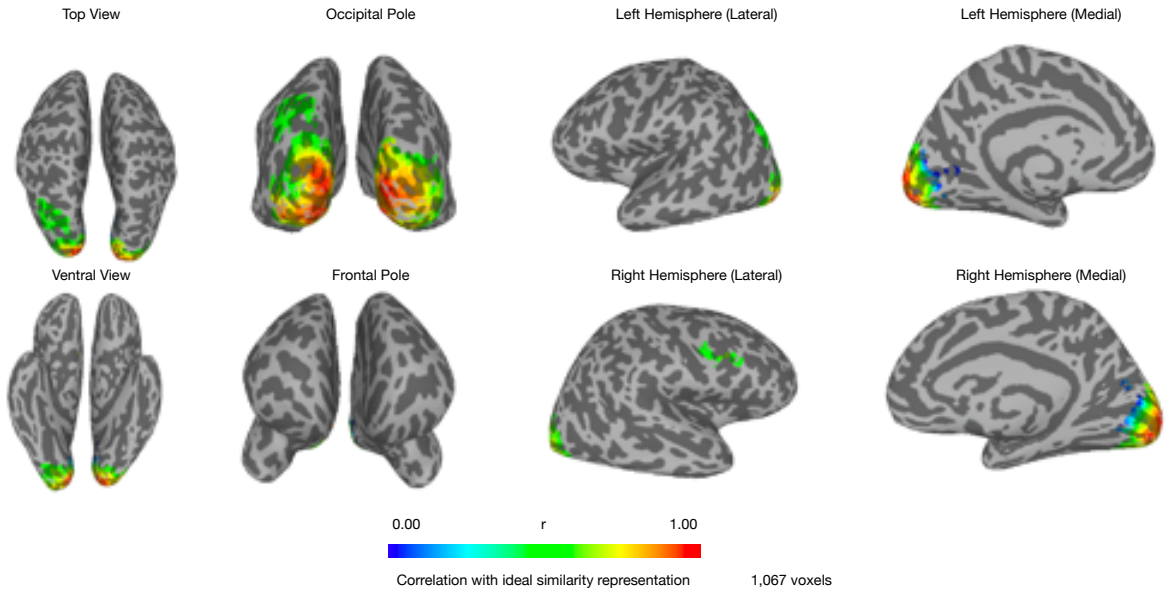


Fig. S33. Neurofeedback ROI for participant #10 (including early visual cortex).
Searchlight analysis indicating all of participant #10's brain regions that represented the shape space parametrically. See Methods for details of the searchlight procedure and threshold selection. (Fig. S12 shows the same brain regions, except without early visual cortex).

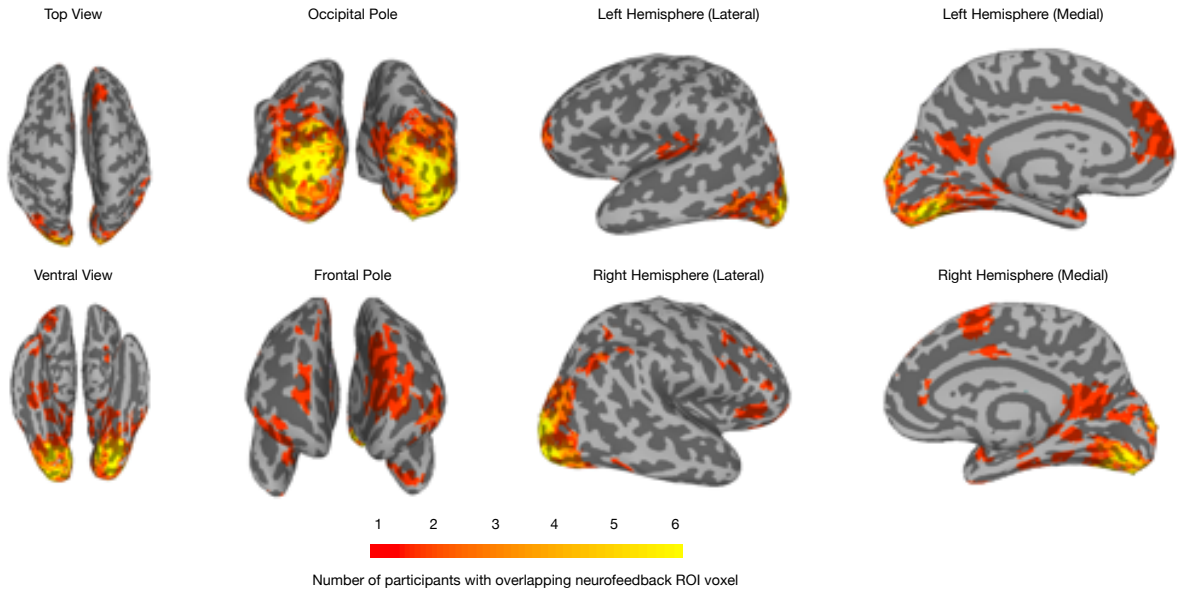


Fig. S34. Group map for neurofeedback ROI.
Group map of neurofeedback ROI overlap for all 10 participants (MNI registration of maps shown in Figs. S3—S12). A high degree of overlap was observed in extrastriate visual cortex (common ROI for up to 6 participants) and, to a smaller degree, in lateral temporal cortex, parahippocampal cortex, and prefrontal cortex (common ROI for up to 3 participants). We note that alignment to a common space is known to introduce biases in the spatial extent of maps, especially for highly specialized ROIs, such as the ones we defined based on an independent decoding accuracy criterion.

			LOCALIZER	LATERAL OCCIPITAL ROI			SEARCHLIGHT COGNITIVE MAP ROI					TRAINED		
	STATUS	PARTICIPANT ID	RUNS	AVG ACCURACY	BEST PARAMS	ACCURACY % PER PARTITION	R THRESHOLD	# VOXELS	AVG ACCURACY	BEST PARAMS	ACCURACY % PER PARTITION	PARAMS	PARTITION	NOTES
1	REJECTED (PERF)	**106**	13	67%	LAG 6, DIM 100	63-63-70-74-68-62								Rejected due to poor performance
2	FINISHED 30/06/2018	**107**	12	78%	LAG 6, DIM 100	78-79-78-77-77-81	0.25	1,432	76%	LAG 6, DIM 150	75-77-73-75-**76**-77	LAG 6, DIM 150	EFGHIJ vs. KLABCD	Finished!
3	REJECTED (PERF)	**108**	12	56%	LAG 6, DIM 150	56-56-55-54-57-55								Rejected due to poor performance
4	WITHDRAWN (3) / REJECTED (PERF)	**109**	14	73%	LAG 6, DIM 100	66-73-77-65-81-72	0.33	783	68%	LAG 6, DIM 150	65-69-72-63-77-64			Quit right before could be rejected due to poor performance
5	FINISHED 17/10/2018	**110**	14	79%	LAG 4, DIM 200	75-81-81-71-88-79	0.40	1,335	74%	LAG 4, DIM 200	68-77-77-67-85-**70**	LAG 4, DIM 200	FGHIJK vs. LABCDE	Finished!
6	WITHDRAWN (3)	**111**	12	74%	LAG 6, DIM 150	74-73-76-76-74-73	0.50	1,755	70%	LAG 6, DIM 150	72-69-68-69-71-72			Scheduled for training, then cancelled x2
7	WITHDRAWN (3)	**112**	13	75%	LAG 6, DIM 150	71-74-78-79-76-72	0.50	954	72%	LAG 6, DIM 200	69-67-73-78-75-67			No email response when invited to return for training
8	WITHDRAWN (1)	**113**												Quit after behavior pre-test
9	REJECTED (MRI)	**114**												Eye surgery revealed when filling out screening form 3rd time
10	WITHDRAWN (2)	**115**												Quit during 3rd run of first localizer scan
11	REJECTED (MRI)	**116**												MRI-unsafe device revealed when filling out screening form 2nd time
12	FINISHED 26/03/2019	**117**	14	73%	LAG 6, DIM 100	66-73-79-63-78-77	0.60	1,710	71%	LAG 6, DIM 100	66-70-**78**-61-76-74	LAG 6, DIM 100	CDEFGH vs. IJKLAB	Finished!
13	FINISHED 30/04/2019	**118**	15	80%	LAG 4, DIM 150	76-85-74-83-76-85	0.50	1,385	79%	LAG 4, DIM 150	75-**85**-73-83-73-85	LAG 4, DIM 150	BCDEFG vs. HIJKLA	Finished!
14	FINISHED 27/06/2019	**119**	15	76%	LAG 6, DIM 150	72-80-68-83-67-83	0.40	2,209	76%	LAG 4, DIM 100	**73**-81-70-81-70-83	LAG 4, DIM 100	ABCDEF vs. GHIJKL	Finished!
15	REJECTED (PERF)	**120**	14	66%	LAG 6, DIM 200	62-66-70-62-69-67								Rejected due to poor performance (fell asleep during localizer scans)
16	FINISHED 20/07/2019	**121**	15	80%	LAG 6, DIM 150	75-88-75-84-75-83	0.40	780	79%	LAG 6, DIM 100	74-88-71-**85**-70-85	LAG 6, DIM 100	DEFGHI vs. JKLABC	Finished!
17	REJECTED (PERF)	**122**	13	70%	LAG 6, DIM 100	68-63-71-78-74-65	0.33	1,444	67%	LAG 4, DIM 150	63-61-71-75-70-63			Rejected due to poor performance
18	WITHDRAWN (1)	**123**												Quit after behavior pre-test
19	FINISHED 29/08/2019	**124**	15	75%	LAG 4, DIM 150	71-79-70-79-70-80	0.50	1,540	73%	LAG 4, DIM 200	69-76-70-**76**-68-76	LAG 4, DIM 200	DEFGHI vs. JKLABC	Finished!
20	REJECTED (PERF)	**125**	14	57%	LAG 4, DIM 150	52-59-60-54-61-54								Rejected due to poor performance
21	FINISHED 04/09/2019	**126**	14	81%	LAG 6, DIM 150	80-79-87-73-85-85	0.66	1,377	79%	LAG 6, DIM 100	78-76-83-75-**83**-82	LAG 6, DIM 100	EFGHIJ vs. KLABCD	Finished!
22	REJECTED (PERF)	**127**	13	70%	LAG 4, DIM 200	66-67-76-79-70-64	0.25	886	65%	LAG 4, DIM 200	62-59-67-75-66-59			Rejected due to poor performance
23	REJECTED (PERF)	**128**	15	75%	LAG 6, DIM 150	66-84-67-83-70-83	0.25	790	67%	LAG 6, DIM 200	59-77-58-74-59-73			Rejected due to poor performance
24	FINISHED 17/10/2019	**129**	14	77%	LAG 6, DIM 100	71-78-84-67-85-77	0.33	1,175	77%	LAG 6, DIM 150	74-78-80-69-84-**75**	LAG 6, DIM 150	CDEFGH vs. IJKLAB	Finished!
25	FINISHED 08/10/2019	**130**	15	75%	LAG 6, DIM 200	70-81-68-80-70-79	0.50	1,067	74%	LAG 6, DIM 150	72-**79**-70-78-70-78	LAG 6, DIM 150	BCDEFG vs. HIJKLA	Finished!

LEGEND

STATUS	Finished	Finished full experiment on listed date
	Rejected (Perf)	Did not meet performance criteria on localizer scans
	Rejected (MRI)	Revealed to be incompatible with MRI scan safety criteria after participating in behavior pre-test, but before first localizer scan
	Dropped (N)	Withdrew participation in the study after N sessions
PARTICIPANT ID	Unique identifier assigned to each participant recruited for the study between May 2018 and October 2019	
LOCALIZER RUNS	12–15	Number of fMRI localizer runs participant underwent during first two scans (11m37s per run)
LO FREESURFER ROI	Avg Accuracy	Average decoding accuracy across all 6 partitions of the category space (LORO cross-validation across LOCALIZER RUNS)
	Best Params	Combination of parameters (PC dimensions x hemodynamic lag) that yielded the highest Avg Accuracy
	Accuracy % Per Partitior	Decoding accuracy for each of the 6 partitions: ABCDEF vs. GHIJKL, BCDEFG vs. HIJKLA, etc.
SEARCHLIGHT COGNITIVE MAP ROI	R Threshold	Cognitive map ROI represents all voxels that correlated with ideal RSA at r>R (see Methods for full description of how R was determined)
	# Voxels	Number of voxels in neurofeedback ROI
	Avg Accuracy	Average decoding accuracy across all 6 partitions of the category space (LORO cross-validation across LOCALIZER RUNS)
	Best Params	Combination of parameters (PC dimensions x hemodynamic lag) that yielded the highest Avg Accuracy
	Accuracy % Per Partitior	Decoding accuracy for each of the 6 partitions, bolded value indicates partition chosen for neurofeedback training (see Methods for description of how the partition was chosen)
TRAINED	Params	Parameters used for building final neurofeedback training neural model in neurofeedback ROI (identical to SEARCHLIGHT COGNITIVE MAP ROI -> Best Params)
	Partition	Shorthand description of which half of the shape space circle corresponds to the partition used for neurofeedback training, e.g., shapes closest to ABCDEF represent category 1, while shapes closest to GHIJKL represent category 2
NOTES	Additional details about each participant who was rejected and/or withdrew their participation from the study	

Table S1. Participant fMRI run details, model decoding accuracy, exclusion criteria, neurofeedback ROI statistics, and training parameters.

We recruited 25 participants for the study, of which 6 withdrew before completing the study, 2 were rejected due to not meeting MRI scanning safety criteria, 7 were rejected due to poor performance during the localizer scans, and 10 finished the full experiment. The embedded table legend provides detailed explanations for each column and field.

PROCEDURE					
IF FIRST NEUROFEEDBACK TRAINING RUN OF SESSION 1				THEN	THRESHOLD = 70%
ELSE IF FIRST NEUROFEEDBACK TRAINING RUN OF DAY N					THRESHOLD = LAST THRESHOLD FROM DAY N-1
ELSE IF PROGRESS WAS	0-1	FOR THE PREVIOUS	1 RUN		DECREASE THRESHOLD BY 5% (MIN. 60%)
	2-3		3 RUNS		
	4-5		5 RUNS		
	6		ANY RUNS		KEEP THRESHOLD UNCHANGED
	7-8		5 RUNS		INCREASE THRESHOLD BY 5% (MAX. 90%)
	9-10		3 RUNS		
	11+		1 RUN		

Table S2. Adaptive procedure for setting the neurofeedback threshold during training.

To account for high variability of fMRI data across multiple days and runs, the LLR feedback threshold for each training run was adjusted given participant performance (how many trials generated positive feedback) on all previous runs since the beginning of the training day using an adaptive procedure designed to converge on giving feedback for approximately 1/3 of the twenty trials in each run.

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
1	107	#1 25/06/2018	1	70%	0
			2	65%	7
			3	65%	4
			4	65%	4
			5	65%	7
			6	65%	6
		#2 26/06/2018	1	65%	2
		#3 27/06/2018	1	65%	3
			2	65%	2
			3	65%	1
			4	60%	5
			5	60%	9
			6	60%	9
			7	60%	8
		#4 28/06/2018	1	60%	5
			2	60%	7
			3	60%	11
			4	65%	5
			5	65%	3
		#5 29/06/2018	1	65%	3
			2	65%	1
			3	60%	5
			4	60%	9
			5	60%	8
			6	60%	8
			7	60%	7
		#6 N/A			
	NUMTRIALS: 520				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
2	110	#1 10/10/2018	1	70%	6
			2	70%	7
			3	70%	2
			4	70%	6
			5	70%	11
		#2 11/10/2018	1	70%	8
			2	70%	7
			3	70%	9
			4	70%	6
			5	70%	9
		#3 12/10/2018	1	70%	10
			2	70%	6
			3	70%	4
			4	70%	1
			5	65%	0
			6	60%	7
		#4 14/10/2018	1	60%	12
			2	65%	11
			3	70%	8
			4	70%	12
			5	75%	3
			6	75%	6
			7	75%	0
		#5 15/10/2018	1	75%	6
			2	75%	5
			3	75%	4
			4	75%	2
			5	75%	4
			6	75%	7
		#6 16/10/2018	1	75%	5
			2	75%	3
			3	75%	2
			4	75%	1
			5	70%	6
			6	70%	10
			7	70%	6
	NUMTRIALS: 720				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
3	117	#1 21/03/2019	1	70%	9
			2	70%	9
			3	70%	6
			4	70%	5
			5	70%	9
			6	70%	6
		#2 22/03/2019	1	70%	3
			2	70%	6
			3	70%	1
			4	65%	10
			5	65%	8
			6	65%	12
			7	70%	2
		#3 23/03/2019	1	70%	4
			2	70%	5
			3	70%	5
			4	70%	5
			5	70%	3
			6	65%	9
		#4 24/03/2019	1	65%	14
			2	70%	6
			3	70%	6
			4	70%	8
			5	70%	4
			6	70%	2
			7	70%	3
		#5 25/03/2019	1	70%	8
			2	70%	4
		#6 N/A			
	NUMTRIALS: 560				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
4	118	#1 23/04/2019	1	70%	4
			2	70%	6
			3	70%	2
			4	70%	8
			5	70%	4
			6	70%	6
		#2 24/04/2019	1	70%	2
			2	70%	3
			3	70%	3
			4	65%	8
			5	65%	7
			6	65%	11
		#3 25/04/2019	1	65%	7
			2	65%	8
			3	65%	6
			4	65%	7
			5	65%	9
			6	65%	6
		#4 26/04/2019	1	65%	8
			2	65%	5
			3	65%	4
			4	65%	8
			5	65%	9
			6	65%	11
			7	70%	2
		#5 27/04/2019	1	70%	5
			2	70%	11
			3	75%	2
			4	75%	3
			5	75%	2
			6	70%	9
			7	70%	6
		#6 29/04/2019	1	70%	7
			2	70%	5
			3	70%	5
			4	70%	8
			5	70%	10
			6	70%	7
	NUMTRIALS: 760				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
5	119	#1 21/06/2019	1	70%	5
			2	70%	5
			3	70%	8
		#2 22/06/2019	1	70%	6
			2	70%	8
			3	70%	6
			4	70%	7
			5	70%	13
			6	75%	1
			7	70%	7
		#3 23/06/2019	1	70%	6
			2	70%	7
			3	70%	7
			4	70%	7
			5	70%	10
			6	70%	9
			7	75%	4
		#4 24/06/2019	1	75%	2
			2	75%	2
			3	75%	4
			4	75%	7
			5	75%	3
			6	75%	5
			7	75%	1
		#5 25/06/2019	1	75%	2
			2	75%	3
			3	75%	8
			4	75%	5
			5	75%	5
			6	75%	11
			7	80%	6
		#6 26/06/2019	1	80%	4
			2	80%	0
			3	75%	3
			4	75%	5
			5	75%	9
			6	75%	6
	NUMTRIALS: 740				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
6	121	#1 14/07/2019	1	70%	4
			2	70%	3
			3	70%	1
			4	65%	4
			5	65%	3
			6	65%	9
		#2 15/07/2019	1	65%	10
			3	65%	2
			4	65%	8
			5	65%	3
			6	65%	6
			7	65%	4
		#3 16/07/2019	1	65%	9
			2	65%	9
			3	65%	9
			4	70%	9
			5	75%	4
			6	75%	1
		#4 17/07/2019	1	75%	0
			2	70%	4
			3	70%	2
			5	70%	1
			6	65%	4
			7	65%	0
		#5 18/07/2019	1	65%	2
			2	65%	6
			3	65%	6
			4	65%	7
			5	65%	4
			6	65%	7
			7	65%	4
		#6 19/07/2019	1	65%	6
			2	65%	4
			3	65%	7
			4	65%	6
			5	65%	6
			6	65%	4
			7	65%	9
	NUMTRIALS: 760				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
7	124	#1 23/08/2019	1	70%	6
			2	70%	5
			3	70%	4
			4	70%	9
			5	70%	7
			6	70%	8
			7	70%	5
		#2 24/08/2019	1	70%	9
			2	70%	10
			3	70%	5
			4	70%	5
			5	70%	4
			6	70%	6
			7	70%	6
		#3 25/08/2019	1	70%	5
			2	70%	5
			3	70%	5
			4	70%	10
			5	70%	8
		#4 26/08/2019	1	70%	4
			2	70%	2
			3	70%	2
			4	70%	4
			5	70%	6
			6	70%	7
			7	70%	1
		#5 27/08/2019	1	70%	9
			2	70%	7
			3	70%	11
			4	75%	3
			5	75%	1
			6	70%	6
			7	70%	5
		#6 28/08/2019	1	70%	7
			2	70%	4
			3	70%	3
			4	70%	5
			5	70%	10
			6	70%	6
			7	70%	2
	NUMTRIALS: 800				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
8	126	#1 28/08/2019	1	70%	0
			2	65%	2
			3	65%	3
			5	65%	2
			6	60%	6
			7	60%	9
		#2 29/08/2019	1	60%	6
			2	60%	5
			3	60%	2
			4	60%	4
			5	60%	3
			6	60%	4
			7	60%	4
		#3 30/08/2019	1	60%	9
			2	60%	12
			3	65%	10
			5	70%	10
			6	75%	9
			7	80%	7
		#4 31/08/2019	1	80%	0
			2	75%	8
			3	75%	1
			4	70%	8
			5	70%	9
			6	70%	0
			7	65%	5
		#5 02/09/2019	1	65%	10
			2	65%	10
			3	65%	11
			4	70%	1
			5	65%	10
		#6 03/09/2019	1	65%	9
			2	65%	7
			3	65%	4
			4	65%	4
			5	65%	10
			6	65%	3
			7	65%	4
	NUMTRIALS: 760				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
9	130	#1 02/10/2019	1	70%	13
			2	75%	7
			3	75%	6
			4	75%	7
			5	75%	8
			6	75%	7
		#2 03/10/2019	2	75%	7
			4	75%	3
			5	75%	5
		#3 04/10/2019	1	75%	9
			2	75%	7
			3	75%	6
			4	75%	10
			5	75%	10
			6	75%	10
			7	80%	5
		#4 05/10/2019	1	80%	1
			2	75%	9
			3	75%	5
			4	75%	6
			5	75%	0
			6	70%	11
			7	75%	8
		#5 06/10/2019	1	75%	7
			2	75%	1
			3	70%	12
			4	75%	8
			5	75%	8
			6	75%	5
			7	75%	1
		#6 07/10/2019	1	75%	9
			2	75%	9
			3	75%	4
			4	75%	4
			5	75%	4
			6	75%	2
			7	75%	2
	NUMTRIALS: 740				

	PARTICIPANT ID	SESSION	RUN	LLR THRESHOLD	PROGRESS
10	129	#1 11/10/2019	1	70%	5
			2	70%	0
			3	65%	6
			4	65%	5
			5	65%	5
		#2 12/10/2019	1	65%	6
			2	65%	7
			3	65%	5
			4	65%	13
			5	70%	5
			6	70%	4
			7	70%	3
		#3 13/10/2019	1	70%	11
			2	75%	7
			3	75%	4
			4	75%	3
			5	75%	7
			6	75%	5
			7	75%	4
		#4 14/10/2019	1	75%	7
			2	75%	6
			3	75%	9
			4	75%	7
			5	75%	9
			6	75%	9
			7	75%	0
		#5 15/10/2019	1	75%	1
			2	70%	9
			3	70%	5
			4	70%	6
			5	70%	5
			6	70%	5
			7	70%	1
		#6 16/10/2019	1	70%	11
			2	75%	3
			3	75%	4
			4	75%	1
			5	70%	3
			6	70%	5
			7	70%	8
	NUMTRIALS: 800				

LEGEND	
PARTICIPANT ID	Unique identifier assigned to each participant recruited for the study between May 2018 and October 2019, includes total number of neurofeedback training trials for this participant (520–800)
SESSION	Date of number of neurofeedback training session
RUN	Number of functional neurofeedback training run for current session
LLR THRESHOLD	Minimum log-likelihood ratio used to give positive feedback on current RUN (see Methods and Table S4 for details about how this was set)
PROGRESS	Number of trials in current RUN where positive feedback was given to participant (min. 0, max. 20)

Table S3. Neurofeedback training run details.

We report details on every feedback run for every training session of every participant, including number of trials with positive feedback and corresponding feedback LLR threshold for each run computed using the procedure detailed in Table S2.

LLR COMPUTATION	Main Effect: TRAINED - CONTROL				BEHAVIOR CORRELATION	
	mean	s.e.m.	t	p	r	p
TR 1 (main result)	0.369	0.154	2.27	0.049	0.761	0.011
TRs 1-2	0.400	0.154	2.46	0.036	0.654	0.040
TRs 1-3	0.335	0.137	2.32	0.045	0.603	0.065
TRs 1-4	0.268	0.108	2.35	0.044	0.562	0.091
TRs 1-5	0.191	0.079	2.30	0.047	0.524	0.120

Table S4. Effects of neurofeedback training are similar when using LLR measured from any number of TRs from each trial, compared to using the first TR only.

LLR difference score computed using model estimates from a variable number of TRs across each neurofeedback trial (1-5; participants received at most 5 instances of feedback per trial). We observed similar results in all cases to when using only the first TR from each trial, with minor weakening of the reported main effect of neurofeedback training as more TRs were added, likely due to repetition suppression effects (*27*) across the length of the 16s trials.

	PARTICIPANT ID	PRE-TEST								POST-TEST								TRAINED CATEGORIES	PSYCHOMETRIC FUNCTION SLOPE SHARPENING: (POST-PRE/(POST+PRE)									
		DATE	AG	BH	CI	DJ	EK	FL		DATE	AG	BH	CI	DJ	EK	FL			AG	BH	CI	DJ	EK	FL	TRAINED	NEUTRAL	CONTROL	EFFECT (T-U)
1	**107**	18/06/2018	2.479	1.519	2.261	3.027	2.236	2.185		30/06/2018	3.251	2.442	1.956	3.045	2.078	3.174		EFGHIJ vs. KLABCD	0.135	0.233	-0.072	0.003	-0.037	0.185	0.184	0.056	-0.017	0.201
2	**110**	20/07/2018	1.589	1.676	1.482	2.363	1.734	1.317		17/10/2018	1.793	1.908	2.233	2.158	2.164	2.069		FGHIJK vs. LABCDE	0.060	0.065	0.202	-0.045	0.110	0.222	0.133	0.007	0.166	-0.033
3	**117**	26/02/2019	1.771	2.070	1.551	1.650	2.267	1.284		26/03/2019	1.073	1.136	1.018	1.105	1.144	0.994		CDEFGH vs. IJKLAB	-0.245	-0.291	-0.207	-0.198	-0.329	-0.127	-0.228	-0.222	-0.249	0.021
4	**118**	28/03/2019	2.682	2.527	2.037	1.926	1.390	1.299		30/04/2019	0.282	0.531	0.540	0.317	0.608	0.501		BCDEFG vs. HIJKLA	-0.810	-0.653	-0.581	-0.717	-0.391	-0.443	-0.554	-0.512	-0.731	0.177
5	**119**	12/06/2019	2.114	1.173	2.702	1.716	2.823	1.982		27/06/2019	1.047	2.748	2.587	2.011	1.702	1.528		ABCDEF vs. GHIJKL	-0.338	0.402	-0.022	0.079	-0.248	-0.129	0.029	0.077	-0.233	0.262
6	**121**	24/06/2019	2.083	2.404	1.360	1.538	1.361	1.542		20/07/2019	2.945	4.300	1.513	1.735	2.447	11.856		DEFGHI vs. JKLABC	0.171	0.283	0.053	0.060	0.285	0.770	0.471	0.284	0.057	0.414
7	**124**	13/08/2019	1.970	1.566	2.535	1.450	2.685	1.941		29/08/2019	2.347	2.713	2.703	2.630	2.552	3.024		DEFGHI vs. JKLABC	0.087	0.268	0.032	0.289	-0.025	0.218	0.153	0.121	0.161	-0.008
8	**126**	21/08/2019	1.370	2.039	2.124	1.723	2.227	1.783		04/09/2019	1.407	2.128	2.009	1.248	1.132	1.842		EFGHIJ vs. KLABCD	0.013	0.021	-0.028	-0.160	-0.326	0.016	0.017	-0.006	-0.243	0.260
9	**130**	19/09/2019	1.703	1.647	1.708	1.994	0.941	1.363		08/10/2019	1.315	1.608	1.389	0.990	1.086	1.655		BCDEFG vs. HIJKLA	-0.129	-0.012	-0.103	-0.336	0.072	0.097	-0.132	-0.003	-0.070	-0.062
10	**129**	19/09/2019	1.414	1.753	1.491	1.302	1.876	1.511		17/10/2019	1.637	1.479	1.876	1.237	2.500	2.089		CDEFGH vs. IJKLAB	0.073	-0.085	0.114	-0.026	0.143	0.161	0.152	0.024	0.015	0.137
																					AVERAGE EFFECT				0.022	-0.017	-0.115	0.137

LEGEND

Trained Directions Psychometric Function Slope
Neutral Directions Psychometric Function Slope
Untrained Directions Psychometric Function Slope

PARTICIPANT ID		Unique identifier assigned to each participant recruited for the study between May 2018 and October 2019
PRE-TEST	Date	Date of behavioral pre-test
	AG	Psychometric function slope estimate in behavioral pre-test for AG direction
	BH	Psychometric function slope estimate in behavioral pre-test for BH direction
	CI	Psychometric function slope estimate in behavioral pre-test for CI direction
	DJ	Psychometric function slope estimate in behavioral pre-test for DJ direction
	EK	Psychometric function slope estimate in behavioral pre-test for EK direction
	FL	Psychometric function slope estimate in behavioral pre-test for FL direction
POST-TEST	Date	Date of behavioral post-test
	AG	Psychometric function slope estimate in behavioral post-test for AG direction
	BH	Psychometric function slope estimate in behavioral post-test for BH direction
	CI	Psychometric function slope estimate in behavioral post-test for CI direction
	DJ	Psychometric function slope estimate in behavioral post-test for DJ direction
	EK	Psychometric function slope estimate in behavioral post-test for EK direction
	FL	Psychometric function slope estimate in behavioral post-test for FL direction
TRAINED CATEGORIES		Shorthand description of which half of the shape space circle corresponds to the partition used for neurofeedback training, e.g., shapes closest to ABCDEF represent category 1, while shapes closest to GHIJKL represent category 2
SLOPE SHARPENING	AG	Psychometric function slope sharpening between pre- and post-test: AG direction (normalized difference score)
	BH	Psychometric function slope sharpening between pre- and post-test: BH direction (normalized difference score)
	CI	Psychometric function slope sharpening between pre- and post-test: CI direction (normalized difference score)
	DJ	Psychometric function slope sharpening between pre- and post-test: DJ direction (normalized difference score)
	EK	Psychometric function slope sharpening between pre- and post-test: EK direction (normalized difference score)
	FL	Psychometric function slope sharpening between pre- and post-test: FL direction (normalized difference score)
	Trained	Average sharpening for two directions at 75 degrees and 105 degrees to category boundary
	Neutral	Average sharpening for two directions at 45 degrees and 135 degrees to category boundary
	Control	Average sharpening for two directions at 15 degrees and 165 degrees to category boundary
	Effect	Trained - Control

Table S5. Estimated psychometric function slopes and changes between pre- (day 1) and post-tests (day 10), for all 6 directions in shape space and 10 participants.

The embedded table legend provides detailed explanations for each column and field.

PARTICIPANT	QUESTION 1
	Did you suspect that there were multiple categories of shapes during the experiment? If so, how many? On this map of the shape space, please draw the categories (you may draw any category boundaries, regardless of the diameters indicated, e.g., AG, BH, etc.). *(Author's note: boundaries drawn shown in Fig. S18).*
	ANSWERS
107	Initially, I thought there were 3 categories of shapes, but on the last day I realized they might have all been the same shape somehow.
110	Yes, maybe 4-5.
117	Yes, 3.
118	Yes, about 4 categories. *(Author's note: participant wrote 4, crossed it out, then wrote 5, crossed it out, then wrote 4 again while thinking about the question for more than minute).*
119	I sometimes suspected 2 and sometimes 3, depending on the day.
121	I identified 4 categories.
124	3-4.
126	Yes. I thought there were two main categories of shapes, w/ each category having variants of the origin shape.
130	Yes, 2.
129	I thought there were about 4-5 categories; yes, I did suspect there were multiple categories.

PARTICIPANT	QUESTION 2
	At the beginning of the experiment, we selected a random category boundary in this space by choosing a random diameter (straight line). If you had to guess, which straight line do you think separates Category 1 from Category 2? How confident are you about your choice on a scale from 1 to 5 (1 = not at all; 5 = very confident)? On this map of the shape space, please draw what you think the most likely category boundary was (you may draw any straight line, regardless of the diameters indicated, e.g., AG, BH, etc.). *(Author's note: boundaries drawn shown in Fig. S18).*
	ANSWERS
107	4. *(Author's note: the participant drew a line on the form close to DJ, see Fig. S18).*
110	GA, 3.
117	A-J-F. 4. C-I 2. *(Author's note: the first category split refers to Question 1 above, the second category split refers to Question 2).*
118	DJ. 3.
119	A-G (3).
121	DJ. 4.
124	3 (EK).
126	Direction: F-> L. confidence: 4.
130	CI, 3.
129	DJ, confidence: 3. *(Author's note: after clarifying with the participant, the actual boundary they chose was AG = orthogoinal direction to DJ.)*

PARTICIPANT	QUESTION 3
	When performing the main experimental task in the scanner (i.e., 'Generate a mental state that makes the shapes stop wobbling'), did you explicitly think about other shapes that were different from the ones on the screen? If so, to which of the A--L shapes at the exterior of the circle were they most similar to? Please share anything else you remember about the strategy or strategies you employed during the experiment.
	ANSWERS
107	I thought about the particular shape at hand and the shapes like A and D were much easier to stop. To get the shapes to altogether stop I repeatedly said "stop" in my head in a yodel kind of voice. But I never was able to stop the shapes like H.
110	No, I was mostly thinking/visualizing the shape stopping.
117	No. I imagined a static form of the shape very close to the shape and focused my attention on the boundaries. I also tried to focus on the static white background. *(Author's note: the shapes were black on an equiluminant grey background.)*
118	No, I tried to think about something that concentrated me completely. Like one shape reminded me of my mom's side profile. LOL. So I would think of her.
119	Yes, I imagined a star *(Author's note: participant drew a five-pointed star on the form mid-sentence)* , which was most similar to shapes A thru E.
121	I tried so many strategies, some ridiculously complex. I settled on recognizing & categorizing the shape -- deciding which category I thought it belonged to and focusing on that shape at rest.
124	Yes, I tried to look at the shape out of focus so that it would gradually morph ito different shapes. Sometimes this seemed to work.
126	I didn't really think about other shapes. I first quickly determined which of my two categories the shape on the screen belongs to, and from that point I picked a moving spot on the shape to focus on continuously. If the shape was in the "mounds" category, I would focus on the curve of the divet, if the shape was in the "butterfly" category, I would focus on the moving point that points into the "body/wings".
130	Shapes BCD were similar, as well as FGHI. I thought about filling up space for shapes BCD and shrinking space for FGHI.
129	No, I did not think of other shapes in order to make the current shape stop wiggling. My strategy was to repeat a word/phrase, as well as mentally visualize a corresponding scene when necessary.

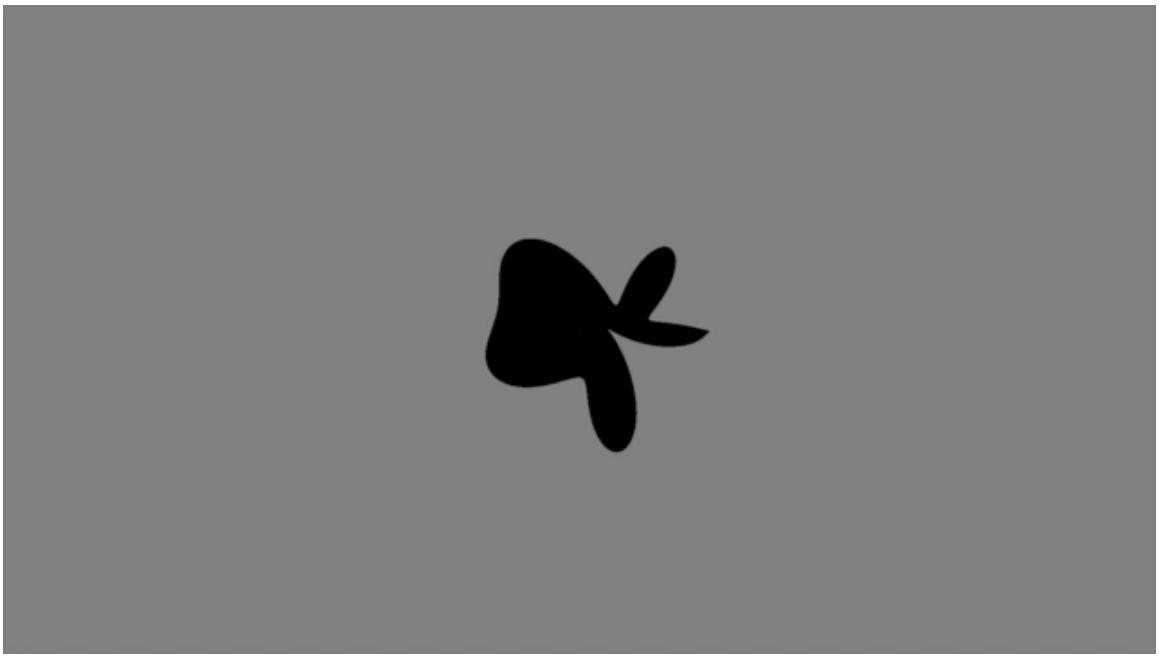
PARTICIPANT	QUESTION 4
	Were any of the shapes or shape types more memorable than others, i.e. did you think about certain trials as 'This is a pointy shape, I need to think about X'? Did you have explicit names for any of the shapes? Did these names evolve or change over the course of the experiment (i.e., days)?
	ANSWERS
107	I called shapes like A "rabbits", shapes like H "butterflies", and shapes like C "starfish". These names were the same over the course of the experiment. The A and C shapes were most memorable because I actually got more to stop.
110	C bug, F tree/cloud on a stick, and I little girl/Powerpuff girl.
117	Yes, shape H was distinct and felt was easiest to reduce jiggling. Shape E/D was memorable as a more difficult one to stop jiggling.
118	I had names for them. 1 girl, 2 elephant, 3 dude, 4 flower. They became easier to categorize as the days passed.
119	Yes. At first, A--D, J--L were walrus, E--H was butterfly, I--K was laughing frog. By the end, B--E was star, G--J was pacman.
121	Yes, I had names for the shape types and the names/categories changed over the runs. The central white pixel made an eye and the shapes were recognizable faces.
124	Yes. Shapes B--L reminded me of a bee, shapes E--H of a flower. No they did not change.
126	The names for my categories are mentioned above as "mounds" and "butterfly". These names came about in probably the third scanning session, when I really figured out my above mentioned techniques.
130	I remembered shape L as a rabbit, J as doll, E as clover, and C as Snoopy. The names remained the same.
129	Yes! For shapes that were more pointy in particular, I had to come up with a scene to accompany my repeated word (as these shapes were harder to stop from wiggling). I didn't have a name for shapes, but I did think that one shape looked like a pokemon.

PARTICIPANT	QUESTION 5
	Did you focus on specific features of the shapes during the experiment in order to perform the task (e.g., the top pointy end, the bottom blobby end)? Did you feel that this worked, i.e. your performance improved by doing so? How confident are you about the effectiveness of this strategy on a scale from 1 to 5 (1 = useless; 5 = very effective)?
	ANSWERS
107	For the non-fmri portions I focused on curved or sharp edges and found this helpful (5). *(Author's note: the participant is referring to the two-alternative forced choice pre- and post-behavioral experiments.)* But for the scanner I viewed a shape more holistically as art of one of my categories.
110	Yes, like the wings on the bug and the curve of the girl's face.
117	I focused on a point in the shapes with the largest surface area. 3. I think this worked well for shape I.
118	Yes, I focused on specific parts when categorize them. It woked in terms of quickly categorizing a shape into i.e. girl and using a strategy that previously worked. 3.
119	Yes *(Author's note: participant drew a five-pointed star on the form mid-sentence with an arrow pointing towards the gap between two vertices),* this part of B--E. It worked for these shapes. I never quite got it right for the other shapes.
121	I continuously made strategies about shape features since focusing on particular elements seemed to work at first. I had more success moving back and considering/focusing on the shape as a whole. I think my strategy was to see the shapes as old friends, to concentrate on recognizing them.
124	I tended to focus on parts of the shape that weren't wiggling as much or on all wiggly parts at the same time, I'd say these strategies worked the best for me. 4.
126	This is also mentioned in my answer to question 3. My "curve of the divet" seems to be your "bottom blobby end. Confidence: 5. When I was relaxed and able to focus it worked very well, but sometimes I simply got distracted or had too many thoughts going through my head for the technique to work.
130	I focused on the front (left) indent and is that didn't seem immediately clear, I look at the middle to compare the balance between the left and right. 4.
129	For shapes with less pointy edges, I focused on the center. For shapes with more pointy ends, I actually focused on the crevices between the pointy parts. I think that this strategy, after practicing, was 4.5 on a scale of effectiveness.

PARTICIPANT	QUESTION 6
	Please share anything else you think might be relevant about your experience with our study.
	ANSWERS
107	It was a very interesting task that I didn't think I'd be able to do at first but was pleasantly surprised when I was able to stop shapes.
110	Some shape types were certainly harder to stop than others, the Bug (C,B,A,D,E) were all hard.
117	Visual fixation & mental distraction; I felt that whenever I was thikning about something else or visualizing something else in my head, while maintaining a solid/steady gaze was efective & less mentally taxing.
118	Having multiple days was useful. I felt it was a bit harder after the day break but not that bad. *(Author's note: this was one of the two participants who requested a 48h gap between two of the sessions for personal reasons).*
119	N/A.
121	This is so fascinating.
124	N/A.
126	Although the goal was to be totally focused, sometimes when I stopped the shape I had a single thought or scenario going through my head, almost as if that helped me stop the shape.
130	N/A.
129	N/A.

Table S6. Final debriefing form and summary of participant answers.

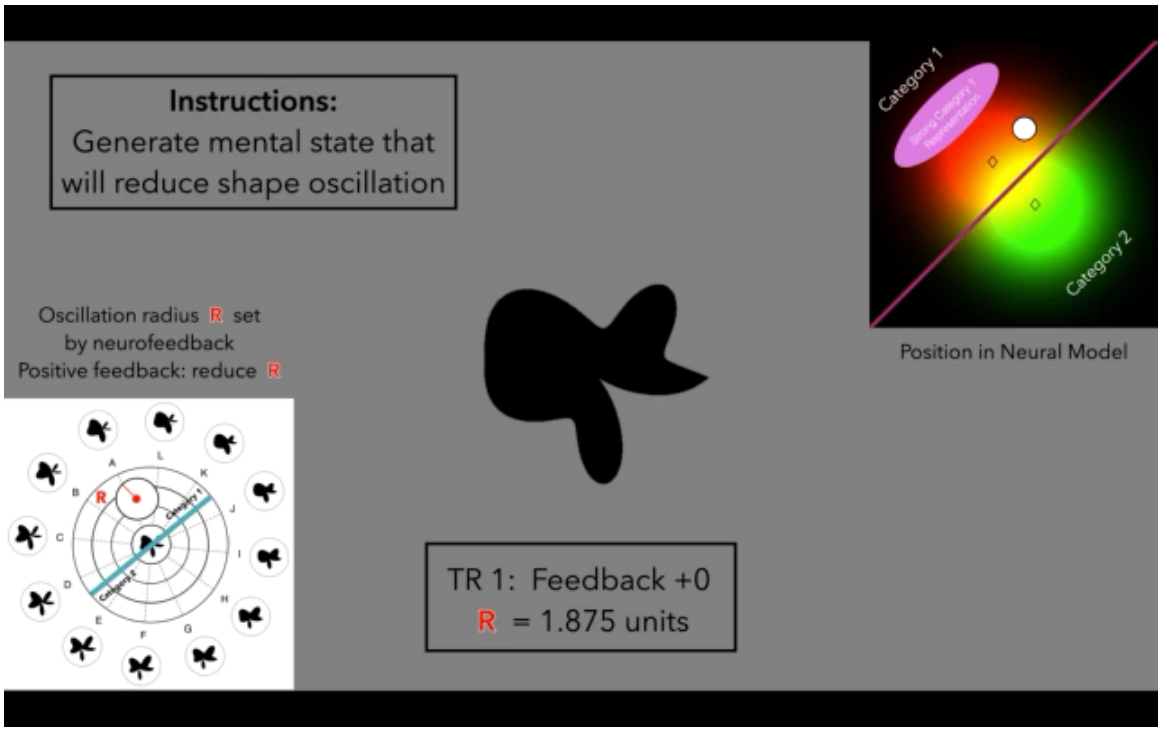
At the conclusion of the study, but before being fully briefed on the goals of the experiment, participants were asked to complete a questionnaire about their experience, strategies for completing the task, as well as guesses (and drawings) of potential category boundaries on a never-before-seen picture of the circular stimulus space.



External Link to Movie (opens in browser): http://www.princeton.edu/~miordan/sculpting.html

Movie S1. Demo of oscillating shape neurofeedback trial.

Video shows two example trials during a neurofeedback training run with no feedback (first trial, shape oscillation radius 1.875 a.d.u.) and simulated positive feedback that eventually stops the shape from oscillating (second trial, shape oscillation radii: 1.875 a.d.u, 1.250 a.d.u., 0.625 a.d.u.. 0 a.d.u.).



External Link to Movie (opens in browser): http://www.princeton.edu/~miordan/sculpting.html

Movie S2. Detailed demo of oscillating shape neurofeedback trial.

Video shows an example trial during a neurofeedback training run with simulated positive feedback that eventually stops the shape from oscillating (shape oscillation radii: 1.875 a.d.u, 1.250 a.d.u., 0.625 a.d.u.. 0 a.d.u.). Top right corner shows simulated trajectory of current stimulus in neural shape model. Bottom left corner shows extent of shape oscillation in circular shape space.