Day	Session 1 (~1.5h)	Session 2 (~1.5h)	Session 3 (~1.5h)	Session 4 (~1.5h)	Session 5 (~1.5h)
Task order	categorical perception task	2 recognition runs	2 recognition runs	2 recognition runs	8 recognition runs
		10 feedback runs	10 feedback runs	10 feedback runs	
	8 recognition runs	2 recognition runs	2 recognition runs	2 recognition runs	categorical perception task

Table 1: Multi-session study protocol for each participant.

Methods

Participants

A total of 20 healthy young adults (mean age = 23.55; age range = 19–28; 14 female) with normal or corrected-to-normal visual acuity participated in this study. They completed five fMRI sessions lasting 1.5 hours each and were paid $20/hour. Sessions were scheduled on separate days but as close together as possible, from a minimum of five days in a row to a total span of eight days. Two additional pilots and eight additional participants who did not complete all five sessions were excluded from analysis.

Ethics Statement

All participants provided informed consent to a protocol approved by the Institutional Review Board at Yale University.

Data acquisition

Data were acquired using a 3T Siemens Prisma scanner with a 64-channel head coil at the Brain Imaging Center at Yale University. For recognition and feedback functional runs, an echo-planar imaging (EPI) sequence was used to collect BOLD data (TR=2 s; TE=30 ms; voxel size=3 mm isotropic; FA=90°; IPAT GRAPPA acceleration factor=2; distance factor=25%), yielding 36 axial slices. Each recognition run contained 145 volumes and each feedback run contained 176 volumes. Two field map volumes (TR=5 s; TE=80 ms; otherwise matching the EPI scans) were acquired in opposite phase encoding directions. For anatomical alignment and visualization, we collected a 3D T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) scan (TR=2.5 s; TE=2.9 ms; voxel size=1 mm isotropic; FA=8°; 176 sagittal slices; IPAT GRAPPA acceleration factor=2), and a 3D T2-weighted fast spin echo scan with variable flip angle (TR=3.2 s; TE=565 ms; voxel size=1 mm isotropic; 176 sagittal slices; IPAT GRAPPA acceleration factor=2).

Real-time system

After image reconstruction, the DICOM files were streamed in real-time to Milgram, a high-performance cluster mounted on the Siemens console. The RT-Cloud software package [31] was used for preprocessing and analysis of each image, with the results transmitted to the task computer at the scanner over the network. This output was used to update the task on the next time point.

Data preprocessing

For real-time analyses, motion correction was applied by aligning each new DICOM file with a template volume acquired from the middle of the first recognition run in the current session using 3dvolreg [32]. The BOLD activity of every voxel was normalized by z-scoring over time using the running mean and standard deviation from the prior TRs in the current run; this ensured that baseline differences in the mean or scaling of BOLD activity across voxels did not confound or bias classifier training. The data were masked to include only voxels in the region of interest (ROI) used for feedback prior to classifier training and testing. For offline analyses, the functional data were field map corrected with the topup tool in FSL [33], registered to the middle volume of the current run with MCFLIRT [34], and z-scored based on the full-time series from that run.

Study design

The study consisted of five sessions (Table 1). Session 1 contained eight recognition runs in which the presented, competitor, and control objects were shown repeatedly. These data were used to train classifier models that could distinguish between all pairs of objects. The trained models were then tested on the feedback runs in Sessions 2, 3, and 4 to measure the amount of evidence for the competitor object during the viewing of the presented object. Participants were encouraged to increase the activation level of the competitor through adaptive, closed-loop neurofeedback, inducing coactivation between the presented and competitor objects. Session 5 mirrored the first session with eight recognition runs, in order to assess representational change. We based the use of three neurofeedback sessions (and five total sessions) loosely on other real-time fMRI studies in our lab [26] and in other labs [35].

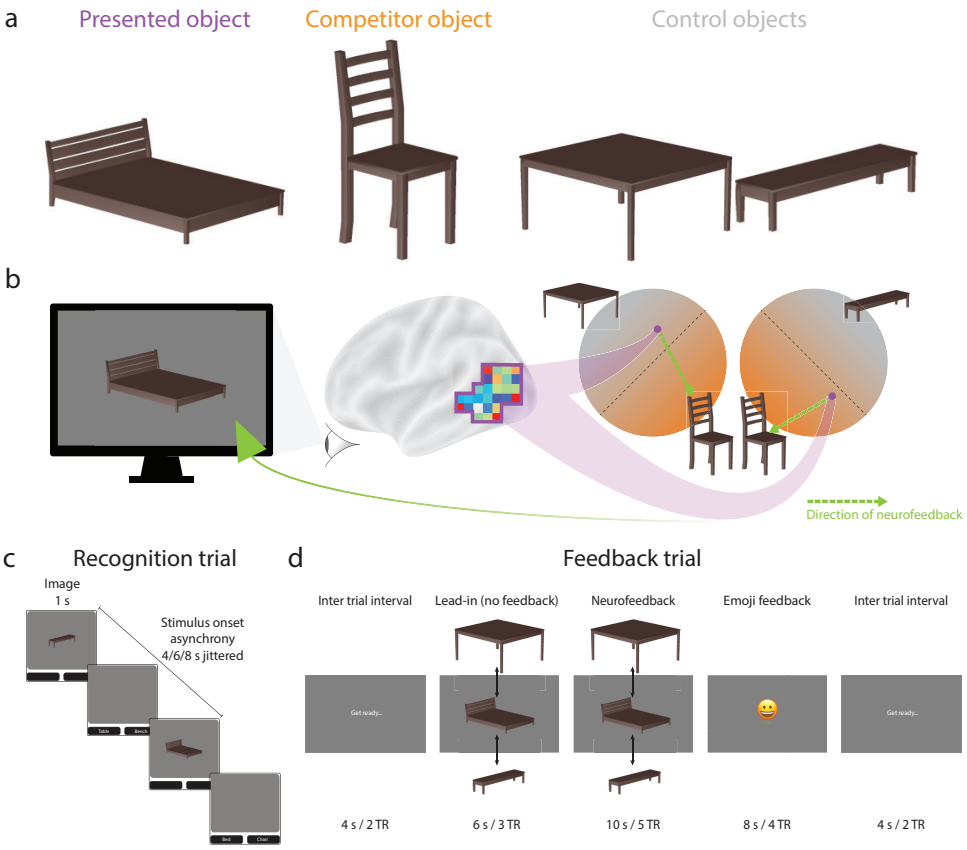

Figure 1: Study design. (a) Each participant received two objects for neurofeedback and the other two objects served as a baseline. (b) The presented object (e.g., bed) was shown on neurofeedback trials and began oscillating in size and shape. The goal of the participant was to make this wobbling stop, which they could achieve by activating the representation of the competitor object (e.g., chair) in their mind. Evidence for the competitor object was quantified based on a classifier trained to decode the competitor object relative to two control objects (e.g., table, bench). The amount of classifier evidence for the competitor needed to reduce the magnitude of wobbling was staircased to maximize coactivation between these objects. (c) Recognition trials showed one of the four objects at a time with no neurofeedback. These trials were used to train the classifier models and to measure neural snapshots of how the object is presented in the hippocampus. (d) Feedback trials were conducted during real-time fMRI to induce coactivation. The main neurofeedback occurred during the object presentation (the amount of wobbling), though participants also received feedback at the end of the trial (monetary reward, valenced emoji). The real-time analysis of each test TR occurred during the acquisition of the next TR by transferring the reconstructed DICOM file instantaneously to a high-performance computing cluster. Because of the hemodynamic lag, we fixed the wobbling at the maximum level for a lead-in period of the first three TRs and began updating it on the fourth TR based on a running average of all TRs collected in the current trial up to that point.

This design employs a within-participant control condition rather than a separate sham or yoked control group. Specifically, one pair of objects served as the target of neurofeedback (trained) and the other pair of objects was a no-neurofeedback baseline (untrained). The assignment of pairs to conditions was counterbalanced across participants. For 10 of the participants (Fig. 1a), the bed (presented object) and chair (competitor object) were the trained pair and the table and bench (control objects) were the untrained pair. For the remaining 10 participants, the assignments were reversed with table (presented object) and bench (competitor object) as the trained pair and bed and chair (control objects) as the untrained pair. By comparing behavioral and neural changes for trained vs. untrained pairs, each participant serves as their own control when assessing the effects of neurofeedback. We based this within-participant control approach on prior neurofeedback studies of learning [28]. Such designs can be efficient because they avoid between-subject variability (e.g., individual differences, cohort effects) that can complicate group comparisons.

Recognition runs

During each trial of the recognition runs (Fig. 1c), participants were presented with one of four rendered furniture objects (bed, bench, chair, table) in one of several potential viewing angles [36,37]. After 1 s, the object was removed from the screen and two furniture labels appeared below, from which participants had to choose which matched the object with an MRI-compatible button box. This response occurred during a jittered interval between trials of 4, 6, or 8 s. Each recognition run contained 48 trials, allowing for 12 repetitions per object and run. Three repetitions of each object appeared in each quarter of a run (no back-to-back repetitions), to ensure an even distribution of objects over time.

The data from the recognition runs in Session 1 were used to train six binary classifiers, corresponding to all combinations of the four objects. We used logistic regression classifiers with L2-norm regularization (penalty=1). Each of the six classifiers contrasted one pair of objects (e.g., chair vs. bench). The training data consisted of patterns of BOLD activity across voxels in a feedback ROI (described below) extracted 4 s after the onset of the object on each trial to account for the hemodynamic lag, labeled by the identity of the object. We validated that the amount of training data was sufficient by quantifying the sensitivity and specificity of the classifiers (Supplementary Fig. 1).

Feedback runs

During each trial of the feedback runs (Fig. 1d), the presented object was shown dynamically on the screen, appearing to wobble in size and shape [29]. BOLD activity patterns were extracted from the feedback ROI and supplied as input to the classifiers. To determine the activation level of the competitor object (Fig. 1b), we averaged the output of the two classifiers trained to discriminate the competitor object from each of the control objects (i.e., competitor vs. control1; competitor vs. control2). Similarly, the activation level of the presented object was determined by averaging the output of the two classifiers trained to discriminate the presented object from each of the control objects (i.e., presented vs. control1; presented vs. control2).

Note that we did not use a classifier trained to discriminate activation of the presented object from activation of the competitor object because we wanted separate estimates of the evidence for these objects. For example, if both objects were active (desirable coactivation), the output of this classifier would be at chance, but this result would also be obtained if neither object is active. Instead, we relied on the control objects as a neutral baseline. An alternative approach for obtaining separate estimates of presented and competitor activation could have been to use a multi-class classifier trained to discriminate activation patterns across all four objects. However, the presented object may still dominate the output of this classifier because it was the only object available perceptually. Indeed, a bias toward the presented object was another reason we did not use a binary presented vs. competitor classifier for neurofeedback, as we suspected that the evidence would always favor the presented object. This is problematic because it may have resulted in a ceiling effect, rendering the classifier insensitive to the subtler fluctuations in competitor activation that were needed for training. We felt that pitting the competitor object against the control objects would place them on a more equal footing, in that none were available perceptually.

Participants received multiple forms of feedback to help motivate them to increase the activation level of the competitor object and foster its coactivation with the presented object, including visual feedback via wobbling, monetary feedback via an increase in their bonus compensation, and valenced feedback via an emoji. If participants successfully raised the activation level of the competitor above an adaptive threshold, the magnitude of wobbling decreased: wobbling began each feedback trial (consisting of 5 TRs) at level 13 (maximal), and reduced to level 9 after 1 TR above the threshold, level 5 after 2 TRs above the threshold, and level 1 (minimal) after 3-5 TRs above the threshold. Participants also received a monetary reward and an emoji after the trial based on the final number of above-threshold TRs: 0 TRs means no money and an unhappy face; 1 TR means no money and a neutral face; 2-3 TRs means 5 cents bonus and a smiling face; and 4-5 TRs means 10 cents bonus and a laughing face. We combined multiple indicators of performance in order to provide participants with rich feedback that was both timely and accurate, and to make the task more fun and engaging.

The wobbling was produced by showing a series of morphed images on a spectrum from 1 to 100, where 1 is the object at one end of the morphing axis and 100 is an object at the opposite end. The four magnitudes of wobbling reflected different ranges of morph values: level 13 changed the image linearly from morph 1 to 40 and back (or 100 to 60 and back); level 9 morph [1, 28] (or [100, 72]), level 5 morph [1,16] (or [100,84]), and level 1 morph [1,4] (or [100,96]). The same number of steps were used for each level, resulting in the fastest motion for 13 and slowest motion for 1.

Participants were informed about these types of feedback and that the feedback depended on their performance in the task. However, they were not instructed that the feedback was based on competitor activation, nor were they instructed on what mental strategy to use. Instead, they were instructed to explore different strategies that seemed to improve feedback. After the study, participants completed a debriefing questionnaire.

The threshold used for determining feedback was adjusted dynamically using an adaptive staircase procedure (Supplementary Table 1). The goal in using staircasing was to start participants at a difficulty level they could achieve at the beginning during their strategy exploration, but then to increase difficulty across runs and sessions such that they would be incentivized to activate the competitor object more and more strongly as they gained control of the feedback. When participants exhibited poor performance, the threshold was decreased, giving them an opportunity to improve and catch up. Conversely, the threshold was increased to create room for further improvement when participants demonstrated higher levels of control.

Feedback ROI

The BOLD activity patterns used to train and test the object classifiers were extracted from a data-driven region of interest (ROI). To define this feedback ROI, we used the neuroSketch dataset [37], in which the same four objects were shown multiple times to other participants. Each of the 300 parcels in the Schaefer atlas [38] classified as gray matter by Freesurfer [39] were retained for further analysis. Individual classifiers were trained for each parcel, and their test performance was quantified using a leave-one-run-out methodology. The performance from each parcel was averaged across all 25 participants in the neuroSketch dataset, resulting in a ranking of parcel performance. To identify the set of parcels that yielded the best performance, we built a mega ROI adding in the voxels from the top-N parcels and re-calculating test performance for each value of N using a leave-one-run-out

approach. The mega ROI composed of the 78 highest-performing parcels yielded the best object classification performance in the neuroSketch dataset.

This mega ROI with 78 parcels served as the starting point for each participant in the current study, but was further refined per individual through a greedy approach. We first removed the voxels from one parcel (77 parcels remaining) and trained and tested a 4-way classifier on the recognition runs from Session 1 with leave-one-run-out cross-validation, and then iterated until all 78 parcels had been the one parcel left out; the iteration in which the remaining 77 parcels yielded the highest decoding performance was retained. Then we repeated the whole procedure, dropping one parcel (76 parcels remaining), iterating until all 77 remaining parcels had been left out, and then retaining the best set of 76 parcels. This process repeated until the voxels from only one parcel remained, yielding performance values for mega ROIs containing 1–78 parcels; the mega ROI with the best performance of all of these combinations was used as the feedback ROI for this participant. As a result, different participants had a different number of parcels in their mega ROI. However, there was good consistency in which parcels were selected, especially in the visual cortex (Fig. 2a).

Our design does not require a control ROI for sham or other feedback. We are not making a claim about which brain region(s) contain the most useful activity for neurofeedback. Instead, our conclusions pertain to the impact of neurofeedback training on representational change in the hippocampus and categorical perception in behavior based on the within-participant comparison between trained and untrained pairs. Thus, the hippocampus and other ROIs (see below) were used to measure the neural outcomes of learning.

Representational change ROIs

To examine how cortical coactivation related to representational change in the hippocampus, we segmented the hippocampus and its subfields anatomically for each participant using the automatic segmentation of hippocampal subfields (ASHS) software package [40] and a reference library of 51 manual segmentations [41,42]. These segmentations defined participant-specific ROIs for the bilateral hippocampus and subfields CA1, CA2/3, dentate gyrus, and subiculum. We also explored broader effects in the cortex by using Freesurfer [43] to create masks for V1, V2, lateral occipital (LO) cortex, inferior temporal (IT) cortex, fusiform gyrus (FG) and parahippocampal cortex (PHC).

Representational change analysis

We assessed representational change in the primary hippocampal ROIs and the exploratory cortical ROIs by comparing the overlap of neural patterns for the presented and competitor objects in Session 1 vs. Session 5 (using the same metric for control objects 1 and 2 as a baseline). The hippocampal ROIs were outside of the (cortical) mega ROIs used for neurofeedback. The cortical ROIs (V1, V2, LO, IT, FG, PHC) partially overlapped with the mega ROI in some cases. The use of some of the same cortical data from Session 1 to assess representational change was not circular because the key dependent measure is the difference between trained and untrained objects, and the selection of the mega ROI was blind to which objects would be trained or untrained. Moreover, the cortical ROIs were defined a priori based on anatomy, and thus the selection of the mega ROI did not impact which cortical ROIs we analyzed or bias the results of these analyses.

For Session 1, we built eight regularized logistic regression classifiers for each participant and ROI, each using seven of the recognition runs for training and the remaining recognition run for testing in a leave-one-run-out manner. The performance of each classifier was quantified using the area under the receiver operating characteristic (ROC) curve.

The ROC curve plots the true positive rate (TPR, or sensitivity) against the false positive rate (FPR, or 1-specificity) to illustrate the diagnostic ability of a binary classifier at various thresholds. TPR is defined as $\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP+FN}}$, where TP denotes the true positives and FN denotes false negatives. FPR is defined as $\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP+TN}}$, where FP denotes false positives and TN denotes true negatives. The area under the ROC curve (AUC) quantifies the classifier's overall ability to distinguish between positive and negative outcomes, ranging from 0.5 (no discrimination) to 1.0 (perfect discrimination). The AUC for each classifier was computed using the 'roc_auc_score' function from Python's sklearn package [44], which requires the classifier's probability outputs and the actual labels.

For Session 5, we tested the eight trained classifiers from Session 1 on the eight recognition runs and averaged the AUCs to obtain an overall performance score for each participant. A decrease from Session 1 to Session 5 indicates reduced discriminability from greater neural overlap. We thus defined a neural integration index for each ROI as the average Session 1 AUC minus Session 5 AUC.

Geometric analysis of shared vs. unique features

We decomposed this measure of overall representational change into the neural features shared between the presented and competitor objects and the neural features unique to each object (and repeated this analysis for control objects 1 and 2 as a baseline). Neural integration (i.e., a reduction in AUC) could reflect an increase in the expression of shared features and/or a decrease in the expression of unique features. We tested both possibilities by comparing how hippocampal patterns of activity evoked by the presented and competitor objects in the recognition runs (and control objects 1 and 2) changed from Session 1 to Session 5. For each participant and recognition run, we mean-centered the timecourse of activity in each voxel from the primary hippocampal

ROI then calculated the average voxel pattern across all trials and recognition runs for each of the four objects, resulting in four voxel-length vectors.

To identify shared and unique dimensions for each participant, we normalized the vectors for the presented and competitor objects to unit length. The shared direction was identified as the sum of the presented and competitor objects (intuitively, a new vector halfway between the two unit vectors). The unique direction was identified as the difference of presented and competitor objects (intuitively, a new vector orthogonal to the sum, connecting the two unit vectors). We then projected the un-normalized vectors from Sessions 1 and 5 onto these two orthogonal directions and calculated the difference in projections for Session 5 minus 1. We repeated this procedure for control objects 1 and 2, including defining their own shared and unique directions, and then calculating the projections on these directions across sessions.

Note that we used the Session 1 data to identify the shared and unique directions and then to calculate the pre-training projections. This was necessary because, by definition, the shared features need to be estimated from the initial representations themselves. If measured from other trials or participants, there would be no guarantee that these were the actual shared and unique features. Critically, any bias induced by this procedure applies equally to both the trained axis (presented and competitor objects) and the untrained axis (control objects 1 and 2), as these axes were treated identically in Session 1. Thus, in the analyses below, the change in shared and unique projections for the trained axis is compared to the change for the untrained axis as an important baseline.

We quantified the change in shared features for each participant as the Session 5 minus 1 difference in projections on the shared direction for the trained axis minus the equivalent change for the untrained axis; a positive value indicates a relative increase in shared features. We likewise quantified the change in unique features as the Session 5 minus 1 difference in projections on the unique direction for the trained axis minus the equivalent change for the untrained axis; a negative value indicates a relative decrease in unique features. Using this vector geometry, we recomputed the overall neural integration effect in the hippocampus for each participant as the sum of the magnitude of increase in the shared direction and decrease in the unique direction.

Categorical perception task

To assess the impact of representational change on behavior, we conducted a categorical perception task [29,36] before fMRI data collection in Session 1 and after fMRI data collection in Session 5. During this task, participants categorized images sampled from along a morph continuum between two object endpoints (e.g., bed to chair). Because the objects were rendered from 3-D models with a matching number of vertices, it was possible to smoothly morph between them. The morph percentage (of the second object) was sampled at 13 steps: 18 (i.e., 18% chair, 82% bed), 26, 34, 38, 42, 46, 50, 54, 58, 62, 66, 74, and 82. These morphs were shown 12 times each during both the pre-test and post-test, always from a trial-unique viewpoint. On each trial, participants were briefly presented (1 s) with the morph and asked to make a forced-choice judgment about which object they saw by clicking one of two buttons that appeared below the image. The assignment of labels to left vs. right buttons was randomized across trials. A logistic regression model was used to analyze the relationship between the morphing parameter and categorization responses:

$$p(x) = \frac{1}{1 + e^{-(x-\mu)/s}}$$

For the Session 1 categorical perception task, the slope and μ parameters were estimated. In Session 5, the μ value from Session 1 was fixed (to reduce noise and enhance model capability) and we estimated the slope. The change in slope indicates the type and degree of representational change between the two objects being discriminated. In particular, a decrease in slope (reduced discriminability) when comparing the presented and competitor objects would be consistent with integration of their representations, whereas an increase in slope (improved discriminability) would be consistent with differentiation. We thus defined a behavioral integration index as the Session 1 slope minus Session 5 slope (positive for integration, negative for differentiation). Importantly, these changes can be evaluated with respect to the analogous change observed for the untrained control objects.

Brain-behavior relationship

To the extent that categorical perception is a behavioral readout of neural overlap in one or more ROIs, the behavioral and neural integration scores should be positively correlated across participants. We quantified this association in each ROI by calculating the Pearson correlation coefficient.

Statistics

We used non-parametric statistics where possible to avoid assumptions of parametric tests. To estimate the sampling distribution of an effect, we performed bootstrap resampling at the group level. Namely, from the original sample of 20 participants, for each of 1,000 iterations we sampled 20 participants with replacement and averaged their values. The mean and 95% bounds of the resulting sampling distribution were used to generate the bar plot. For hypothesis testing, we determined the p-value as the proportion of iterations on which the average had the opposite sign from the original effect.

Data availability

Data and analysis code will be shared publicly upon acceptance and are available for the review process at: https://drive.google.com/drive/folders/17ZUIoqzzBoURJyPX0Wj0JlQ5ipl0WmmX (data) and https://github.com/KailongPeng/real_time_paper (code).