nature neuroscience

Article

https://doi.org/10.1038/s41593-026-02311-2

Human learning of noninvasive brain–computer interfaces via manifold geometry

Received: 24 March 2025

Accepted: 17 April 2026

Published online: 09 June 2026

Check for updates

Erica L. Busch[1,2], E. Chandra Fincke[1], Guillaume Lajoie[3,4], Smita Krishnaswamy[2,4,5,6,7] & Nicholas B. Turk-Browne[1,2]

Brain–computer interfaces (BCIs) promise to restore and enhance human capabilities. Yet, their adoption has been limited by slow and inconsistent learning across users. We show that BCI learning is accelerated by leveraging the naturally occurring geometry, or intrinsic manifold, of brain activity, extracted using data diffusion. Participants were trained with real-time functional magnetic resonance imaging to control an avatar in a video game by self-modulating activity in brain regions supporting spatial navigation. We perturbed the mapping between brain activity and avatar movement to test how neural manifolds constrain human BCI learning. When new mappings relied on directions of significant variance on the intrinsic manifold, participants successfully gained control by realigning brain activity along these directions. When new mappings did not follow the intrinsic manifold, participants could not learn to control the avatar. These findings show how manifold geometry in higher-order brain regions guides human learning of complex cognitive tasks, identifying a principle for improving future neurotechnologies.

BCIs enable users to control external devices like computer displays, communication tools or robotic effectors with their brain activity[1–8]. Progress in developing BCIs for neural rehabilitation and cognitive augmentation in humans has been hindered by challenges with neural decoding and user training[8–11]. As such, current human BCIs are prohibitively slow for users to learn and require frequent calibration to maintain performance, and even still many 'nonresponders' are never able to gain control[12–18].

In this project, we design and validate a computational framework that enhances human BCI learning. We hypothesized that some brain states are easier for people to generate, and that tailoring training to these brain states will facilitate BCI learning. To test this hypothesis, we trained healthy human participants to control a video game while their brain activity was measured noninvasively with closed-loop, real-time functional magnetic resonance imaging (rt-fMRI). This technique allowed us to capture whole-brain activity at high-resolution, analyze the data and update the game display based on the results every 2 s (the rate of fMRI data acquisition). rt-fMRI has been used previously as a form of noninvasive BCI for neurofeedback[19–23] and yet suffers the same challenges as other BCIs of slow learning or nonresponse[18,24]. Within this system, we perturbed the relationship between measured brain activity and the video game display to probe how humans learn to generate brain states.

Recent studies on neural prosthetics in nonhuman primates bolster this approach. In these studies, monkeys were trained to control a cursor based on invasive multi-unit recordings from the primary motor cortex. Learning of this task occurred more rapidly when the target brain states conformed to the naturally occurring geometry,

[1]Department of Psychology, Yale University, New Haven, CT, USA. [2]Wu Tsai Institute, Yale University, New Haven, CT, USA. [3]Department of Mathematics and Statistics, Université de Montréal, Montreal, Quebec, Canada. [4]Mila, Quebec Artificial Intelligence Institute, Montreal, Quebec, Canada. [5]Department of Computer Science, Yale University, New Haven, CT, USA. [6]Department of Genetics, Yale University, New Haven, CT, USA. [7]Program in Applied Mathematics, Yale University, New Haven, CT, USA. e-mail: smita.krishnaswamy@yale.edu; nicholas.turk-browne@yale.edu

https://doi.org/10.1038/s41593-026-02311-2

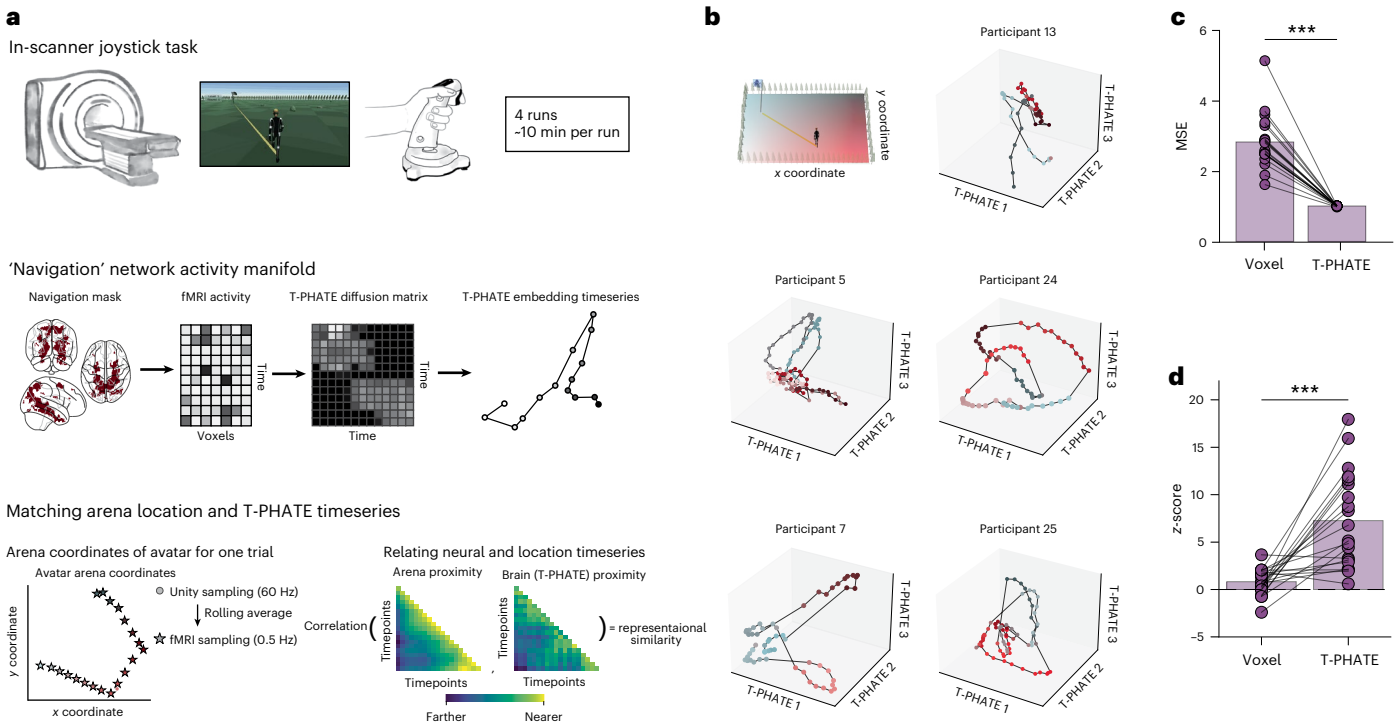

Fig. 1 | Manifold learning and validation. a, Participants practiced self-guided navigation by steering an avatar with a joystick during a video game task (Unity3d) while undergoing fMRI scanning. Multi-voxel fMRI patterns were extracted during this task from a network of brain regions implicated in spatial navigation (see Supplementary Fig. 1 for more views). These patterns were then embedded into a low-dimensional manifold using the T-PHATE algorithm. Each data point is a single fMRI timepoint (one repetition time (TR), that is, the time between consecutive brain volume acquisitions) visualized in three T-PHATE dimensions, with consecutive timepoints connected by a line. For the main analyses, data were embedded into 20 T-PHATE dimensions. The avatar's arena location was downsampled to 0.5 Hz using average pooling to match the temporal resolution of fMRI acquisition and shifted two timepoints (4 s) to correct for the hemodynamic lag. We used representational similarity analysis to compare the similarity structure of brain activity patterns and the avatar's arena location across all pairs of timepoints. This method considers both the continuous movement of the avatar through the arena and that the avatar may return to the same locations at later timepoints. **b**, T-PHATE embeddings of fMRI data show the correspondence between brain activity and the video game arena environment. T-PHATE neural timeseries for five representative participants during four consecutive trials of the joystick task are depicted. The location of the avatar in the game arena is represented by the color of each timepoint, whereas the coordinates in the embedding correspond to the multi-dimensional brain state in T-PHATE space. The clustering of timepoints with similar colors indicates a structured representation of the arena locations in the T-PHATE embedding. **c**, Cross-validated linear models were used to decode the avatar's location from brain data ($N = 20$). These predictions were more accurate when data were embedded in the T-PHATE manifold relative to the original voxel space ($P = 0.0002$) or when embedded with PCA (Supplementary Fig. 2c). **d**, Representational similarity was used to relate distances between brain activity patterns to distances between arena locations ($N = 20$). The representational geometry of T-PHATE embeddings better matched the arena location geometry than the original voxel space ($P = 0.0002$) or PCA embeddings of the same data (Supplementary Fig. 2d). For both **c** and **d**, statistical significance was assessed using a nonparametric randomization test (paired, two-tailed, 10,000 iterations). *** $P < 0.001$.

or 'intrinsic manifold', of the neural population activity[25–28]. This kind of manifold-based BCI learning also promotes more efficient transfer between similar brain states and supports more stable performance over time[29–35]. These foundational findings suggest that it may also be important to extract the intrinsic manifold of brain regions to determine which states to attempt to train in human BCI learning.

Our framework represents an advance over earlier work in several ways. First, given the integral role of higher-order cognitive processes in successful BCI learning[16,36], we targeted a network of brain regions linked to spatial navigation and goal-directed behavior rather than sensorimotor regions (Fig. 1a and Supplementary Fig. 1). Second, to engage these processes during BCI learning, we designed a video game task in which participants navigated an avatar through a virtual reality arena. This focus on modulating higher-order brain regions to control complex behavior lends insight to future applications of BCIs for enriching human cognition beyond the motor domain. Third, given the high dimensionality and noise inherent to fMRI and other noninvasive techniques, we developed state-of-the-art algorithms for learning and extending the intrinsic manifold. Data-diffusion methods have found success in applied mathematics for discovering nonlinear structure in complex biomedical processes[37–39]. We optimized a variant for fMRI—temporal potential of heat diffusion for affinity-based transition embedding (T-PHATE)—that learns a lower dimensional manifold for each participant and highlights brain states related to cognition and behavior better than other dimensionality reduction approaches[40] (Fig. 1), including the linear methods used in most earlier BCI studies.

Participants in our study completed four fMRI sessions. The first was a calibration session to learn the participant's intrinsic manifold with T-PHATE while they practiced navigating the avatar with a joystick. The following three neurofeedback sessions tested how manifold geometry constrained the participant's ability to control the avatar by modulating distributed brain activity patterns (Fig. 2). Each session required the participant to learn a new BCI mapping between fMRI activity and the avatar's heading direction. The second session used an 'intuitive mapping' (IM), which best captured the participant's intrinsic manifold. After learning to control the IM, the third and fourth sessions (order counterbalanced) critically tested perturbations of the mapping that were within or outside the intrinsic manifold (Fig. 2). As hypothesized, participants rapidly learned the IM and the 'within-manifold perturbation' (WMP), defined as a second

https://doi.org/10.1038/s41593-026-02311-2

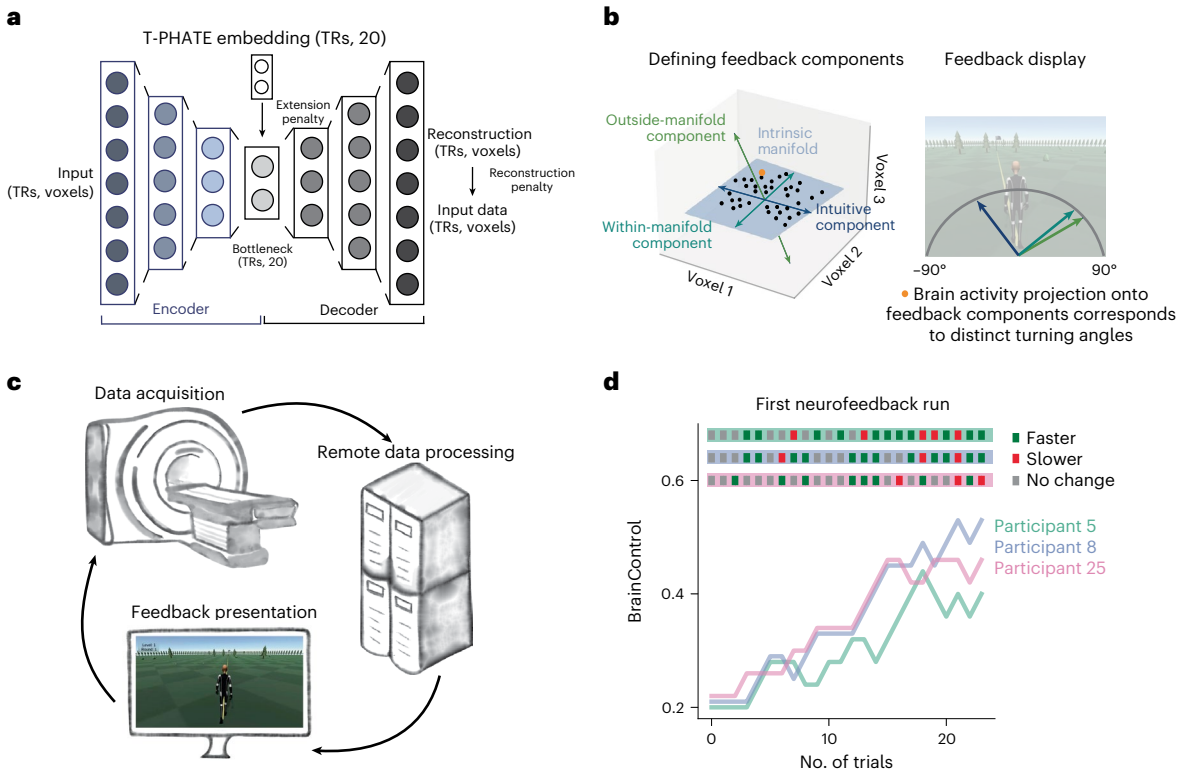

Fig. 2 | Real-time manifold extension and neurofeedback. a, An MRAE uses a T-PHATE layer to penalize the latent space of the autoencoder to reflect the input data's T-PHATE manifold geometry. **b**, Schematic depicting the activity of three voxels over time (black points) with their intrinsic manifold (gray plane) showing an 'intuitive' component (navy vector, capturing a main dimension of variance on the manifold), a 'within-manifold' component (teal, a different, highly explanatory component on the manifold) and an 'outside-manifold' component (green, explaining less variance). These vectors determine the angle of the avatar's movement in the game. A sample data point is highlighted in orange. When projected onto each component, the point results in three different angles of movement in the feedback display. Positive projections (right of the vectors' intersection) result in turns of 0° to 90°; negative projections (left of intersection) result in turns of −90° to 0°. **c**, Closed-loop procedure used rt-cloud[64] software to transfer, analyze and administer continuous feedback based on fMRI scans acquired every 2 s. **d**, The proportion of control that a participant's brain activity exerts over the avatar's movement—the BrainControl parameter—is adjusted at the end of each trial with an adaptive staircasing procedure, increasing with better accuracy (green dashes) and decreasing with worse accuracy (red dashes). The accuracies and staircasing curves of three representative participants are depicted for the first neurofeedback run.

main direction of variance along the intrinsic manifold, but failed to learn the 'outside-manifold perturbation' (OMP), defined as the lowest ranked component in the manifold. Successful BCI learning led to neural realignment, or relative increases in the neural variance accounted for by the trained manifold component, and to enhanced decoding of task information.

Results

Discovering the manifold of human brain regions

Our initial goal was to learn a meaningful manifold of brain activity that could be used in later sessions to guide BCI learning. We first tested whether activity measured from navigation-related brain regions contained dynamic information about the joystick version of the task. A key use of manifold learning methods is for visual exploration of high-dimensional data in lower dimensions[39], so we embedded each participant's fMRI timeseries data for all trials of the joystick task into three dimensions using T-PHATE[40] and labeled each timepoint by the avatar's corresponding coordinate in the game arena (Fig. 1a). Validating that the brain activity embedded with T-PHATE reflected the structure of the arena, points were clustered according to their proximity in the arena and trajectories through the points over time traversed the arena smoothly (Fig. 1b). T-PHATE embeddings better reflected arena locations and trajectories than principal component analysis (PCA) embeddings of the same data (Supplementary Fig. 2).

Earlier work has shown that T-PHATE not only aids in data visualization but also yields cleaner task representations[40,41]. To test this possibility, we trained cross-validated linear regression models for each participant on the original voxel-resolution data and on the T-PHATE embeddings (in 20 dimensions) to predict the avatar's arena coordinates at each timepoint of the joystick task. Regression models were scored as the mean squared error (MSE) between the model-predicted and true coordinates in held-out data, with lower error indicating superior decoding. The models trained on T-PHATE embeddings were more accurate than those trained on voxel-resolution data (mean difference = −1.89, $P = 0.0002$, 95% confidence interval (CI) = [−2.23, −1.60], Cohen's $d = 2.60$; Fig. 1c).

Beyond predicting individual locations, we further tested whether the representational geometry of brain activity in neurofeedback regions reflected the navigational geometry of the game arena. In other words, does T-PHATE improve the alignment of distances between locations in the game and distances between the neural representations of the locations? We used a Mantel test to evaluate representational similarity as z-scores relative to a null distribution[42] and found that distance in the game arena was more correlated with pattern similarity of T-PHATE embeddings than of voxel-resolution activity (mean difference = 5.92, $P = 0.0002$, 95% CI = [3.67, 8.32], $d = 1.23$; Fig. 1d). Together, these results validate that manifold learning enhances access to task-related information.

Extending the manifold to new samples for neurofeedback

Despite the benefits of manifold learning for modeling low-dimensional neural dynamics, a key drawback of most nonlinear dimensionality

reduction methods is that the learned dimensions cannot be easily extended to untrained data[39]. To circumvent this limitation, we developed a manifold-regularized autoencoder (MRAE) that can embed a participant's untrained data in their respective T-PHATE manifold in real time[43,44] (Fig. 2a). MRAE is trained to optimize for two measures of model fit: (1) a reconstruction error penalty, which uses MSE to quantify how well the autoencoder reproduces the original data; and (2) a manifold regularization penalty, which quantifies the error between where the encoder places a sample and its true location on the manifold. These penalties are combined into a single loss function to train MRAE. This training loss decreased rapidly and plateaued by 6,000 training epochs (mean across participants: MSE = 0.21, s.e.m. = 0.0093; Supplementary Fig. 3a). The manifold regularization penalty on its own converged by 5,000 training epochs (MSE = 0.0015, s.e.m. = 0.00004; Supplementary Fig. 3b).

To establish a null baseline, we conducted the full multi-session experiment, including the neurofeedback pipeline calibration and real-time procedure, on fake participants whose simulated fMRI activity consisted only of realistic noise estimated from the brain activity of neurofeedback participants[45]. The data did not contain task-related responses during the calibration session or learning effects during the neurofeedback sessions, and thus we did not expect T-PHATE to learn a useful manifold or for the MRAE training to succeed. Indeed, the overall MRAE training loss (MSE > 1,500, s.e.m. > 500; Supplementary Fig. 3a) and the subscore for the manifold regularization penalty (MSE > 600, s.e.m. > 40; Supplementary Fig. 3b) neither decreased nor converged. Thus, the neural manifolds of real participants engaged in the joystick task reflected learnable, extensible structure that was not found in simulated brains devoid of a task-driven signal (see Supplementary Figs. 4 and 5 for individual participant data).

Manifold-constrained learning of BCI control

We quantified BCI learning by the increase in 'BrainControl', a confidence parameter that tracked the ability of participants to direct the avatar toward the goal efficiently with their brain activity (Fig. 2d). Specifically, BrainControl is the staircased proportion[46] of the avatar's movement direction dictated by the angle decoded from the brain; the remaining proportion reflects the angle straight toward the goal, providing assistive guidance in the correct direction. Change in BrainControl was computed at each trial relative to the first trial of the session (Fig. 3a). In the first neurofeedback session, BCI learning of the IM occurred rapidly, reaching significance after trial 4 and plateauing around trial 40. Adjusting to the WMP in the second or third neurofeedback session (order counterbalanced across participants) led to a slower but steady increase in BrainControl, reaching significance around trial 30. Participants did not adjust to the OMP, as indicated by the lack of change in BrainControl from baseline.

Overall BCI learning was quantified as ΔBrainControl from the first to the last trial of each session type (Fig. 3b). BCI learning varied significantly across session types ($F(2, 51) = 51.87$, $P = 5.11 \times 10^{-13}$, $R^2 = 0.67$). There was a significant positive learning effect in the IM session (mean ΔBrainControl = 49.33, 95% CI = [45.33, 55.33], $P = 0.0001$, $d = 5.55$) and in the WMP session (mean ΔBrainControl = 16.67, 95% CI = [7.78, 25.56], $P = 0.002$, $d = 0.84$), but not in the OMP session (mean ΔBrainControl = −0.67, 95% CI = [−6.67, 5.78], $P = 0.90$, $d = -0.048$). The learning effect was significantly greater in the IM session than the WMP session (mean difference = 32.67, 95% CI = [22.00, 42.89], $P = 0.0002$, $d = 1.42$) or the OMP session (mean difference = 50.00, 95% CI = [43.33, 56.67], $P = 0.0001$, $d = 3.31$), as well as in the WMP session compared with the OMP session (mean difference = 17.77, 95% CI = [5.56, 30.00], $P = 0.009$, $d = 0.64$). Using a linear model, we tested whether the difference between learning in the WMP and OMP sessions was explained by counterbalancing order. There was no significant effect of which perturbation participants received first ($F(1, 32) = 1.03$, $P = 0.32$, partial eta squared (η_p^2) = 0.03), nor an interaction between order and session type ($F(1, 32) = 1.14$, $P = 0.29$, $\eta_p^2 = 0.03$; Supplementary Fig. 6).

BCI learning effects in neurofeedback participants were benchmarked against simulated null fMRI participants that underwent identical procedures without the possibility for learning (Fig. 3c). As expected, these simulated data did not show a significant increase in BrainControl in any session type (P values > 0.1). This analysis verifies that the observed BCI learning effects in the real participants were not an artifact of the staircasing procedure or rt-fMRI processing pipeline.

Realignment of neural activity along manifold

We hypothesized that the mechanism for BCI learning is the alignment of brain activity with the manifold component being trained. That is, by modulating their brain activity in high-dimensional space to increase neural variance along the trained BCI component, participants can more precisely control the movement of the avatar through access to the full range of heading directions mapped to that component. To quantify this neural realignment, we calculated the percentage of total variance in the fMRI signal of each run that could be explained by the feedback component for each session type (Fig. 4a). We then compared the change in this percentage of explained variance (ΔPEV) between the first and last runs of each session.

Neural activity became more aligned with the IM component (mean across participants: ΔPEV = 1.06, 95% CI = [0.35, 1.72], $P = 0.005$, $d = 0.70$) and the WMP component (ΔPEV = 0.71, 95% CI = [0.01, 1.49], $P = 0.042$, $d = 0.42$) during their respective neurofeedback sessions, but not with the OMP component (ΔPEV = −0.45, 95% CI = [−1.36, 0.39], $P = 0.35$, $d = -0.23$; Fig. 4b). Importantly, these increases in alignment for IM and WMP were selective to the trained component: there was no increase in alignment with WMP or OMP components in the IM session, nor with IM or OMP components in the WMP session (Supplementary Fig. 7). The amount of neural realignment was comparable between IM and WMP components in their respective sessions (mean difference = 0.35, 95% CI = [−0.67, 1.32], $P = 0.52$, $d = 0.15$) and greater than for the OMP component in its session (versus IM: mean difference = 1.51, 95% CI = [0.52, 2.49], $P = 0.005$, $d = 0.69$; versus WMP: mean difference = 1.16, 95% CI = [0.034, 2.29], $P = 0.034$, $d = 0.46$).

Three additional analyses provided further evidence of neural realignment. First, total variance in brain activity did not change over the course of learning (Supplementary Fig. 8a; P values > 0.1), which would have complicated interpretation of proportional changes. Second, the ΔPEV effect was specific to the trained component (for example, C_{IM} during IM training; Supplementary Figs. 7 and 8b). Arbitrary components of the manifold did not show significant ΔPEV in any session and the ΔPEV after on-manifold training was greater than arbitrary components (C_{IM}: mean $z = 0.57$, 95% CI= [0.25, 0.88], $P = 0.002$, $d = 0.81$; C_{WMP}: mean $z = 0.41$, 95% CI = [0.03, 0.82], $P = 0.03$, $d = 0.47$); this was not true for outside-manifold training (C_{OMP}: mean $z = -0.17$, 95% CI = [−0.59, 0.24], $P = 0.77$, $d = -0.18$). Third, participants increased the variability of projections onto the trained components with learning (Supplementary Fig. 8c), reflecting an enhanced capacity for generating activity patterns to support task performance. Participants utilized a marginally greater range of patterns after on-manifold learning (C_{IM}: mean variance change = 0.028, 95% CI = [0.00, 0.06], $P = 0.06$, $d = 0.38$; C_{WMP}: mean variance change = 0.013, 95% CI = [0.00, 0.03], $P = 0.079$, $d = 0.35$), but not outside-manifold learning (C_{OMP}: mean variance change = −0.027, 95% CI = [−0.06, 0.00], $P = 0.95$, $d = -0.04$).

This latter increase in the variance of projection values could be attributed to an alternative subselection neural strategy that is driven by a reduction in the frequency of low-magnitude projections (Supplementary Fig. 9a) rather than realignment (which predicts that the distribution of projection values grows broader and includes low and high values). The realignment strategy would be more effective for

https://doi.org/10.1038/s41593-026-02311-2

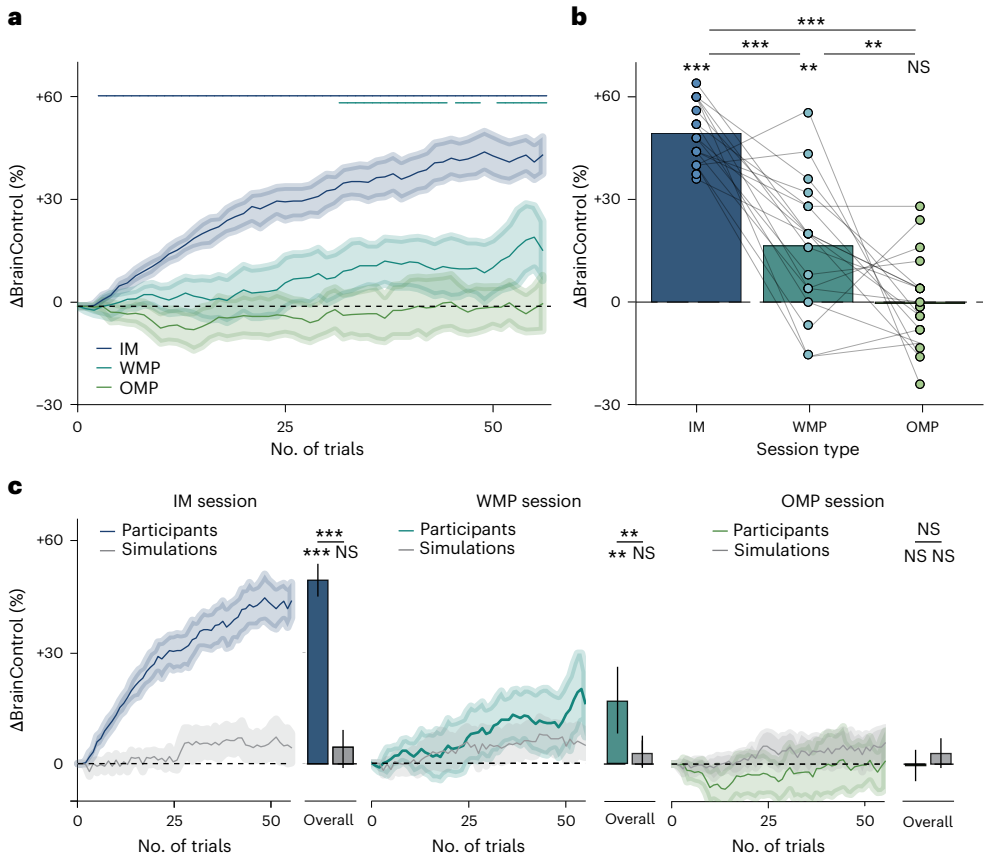

Fig. 3 | BCI learning of different manifold components. a, Learning curves of the change in BrainControl across training trials for each session type. Lines show mean across participants ($N = 18$) and bands show the 95% CI of the mean. The number of trials (55) depicted on the *x* axis corresponds to the minimal trial count successfully completed by all participants across all session types (participants who performed better could complete a few additional trials in the same amount of time, not shown here). Horizontal lines at the top indicate where the mean was significantly greater than chance (ΔBrainControl > 0, $P < 0.05$) for each session type (nonparametric cluster-based correction). **b**, Change in BrainControl from the first trial to the last trial completed in each session type for each participant. The total number of completed trials varied slightly by participant and session. Bars represent the mean across participants; points represent individual participants; lines connect the points from the same participant across the three sessions. IM ($P = 0.0001$) and WMP ($P = 0.002$) session types showed significant learning but OMP sessions did not ($P = 0.90$). Learning during the IM session was significantly greater than that during WMP ($P = 0.0002$) and OMP ($P = 0.0001$) sessions, and WMP was greater than OMP ($P = 0.009$). **c**, Comparing trial-wise and overall BCI learning by session type across real (colors; $N = 18$) and simulated null (gray; $N = 20$) participants. Lines show mean across participants and bands show 95% CIs of the mean. Bars indicate mean overall learning (identical to **b** for real participants) and error bars represent 95% CIs of the mean. For IM ($P = 0.0001$) and WMP ($P = 0.006$) sessions, but not OMP ($P = 0.40$) sessions, real participants showed significantly greater learning than simulated null participants. Statistical significance was assessed using nonparametric randomization tests (10,000 iterations; one-tailed; paired samples in **b** and independent samples in **c**). 95% CIs were estimated via bootstrap resampling (10,000 iterations) $^{***}P < 0.001$, $^{**}P < 0.01$, $^{*}P < 0.05$, not significant (NS) $P \geq 0.1$.

gaining control of the full range of directions the avatar could turn in the game, as the component was mapped continuously onto turning direction such that both high and low projection values were useful for different angles. Nevertheless, to quantify these two strategies we modeled what they would predict for variance change in the IM and WMP sessions and found that the observed data were better fit quantitatively by realignment than by subselection (Supplementary Fig. 9b), particularly during WMP learning. Together, these analyses support our interpretation of neural realignment during successful BCI learning in this task.

Relationship between BCI learning and neural realignment

Neural activity became more aligned with on-manifold components (IM and WMP) during neurofeedback training. We fit a linear model to predict whether the relationship between BCI learning (ΔBrainControl) and neural realignment (ΔPEV) differed across session types, including counterbalancing order as a fixed effect and clustered standard errors by participant. A likelihood ratio test confirmed that the inclusion of random intercepts for participants did not improve model fit ($\chi^2 \approx 0$, $P = 0.50$), justifying the use of ordinary least squares regression. The model accounted for a significant proportion of variance in ΔBrainControl ($F(6, 47) = 35.88$, $P = 9.71\times10^{-9}$, $R^2 = 0.738$) and revealed a significant interaction between ΔPEV and session type ($F(2, 17) = 4.40$, $P = 0.02$): there was no significant effect of ΔPEV in the IM session ($\beta = -2.00$, 95% CI = $[-4.04, 0.03]$, $z = -1.41$, $P = 0.16$) or the OMP session ($\beta = 1.24$, 95% CI = $[-3.01, 5.50]$, $z = 0.57$, $P = 0.57$), but there was a significant effect in the WMP session ($\beta = 7.85$, 95% CI = $[2.38, 13.32]$, $z = 2.81$, $P = 0.005$). Simple slope analyses for each session confirmed that ΔPEV significantly predicted ΔBrainControl during WMP ($\beta = 5.60$, 95% CI = $[0.15, 11.05]$, $t(16) = 2.18$, $P = 0.045$), but not IM ($\beta = -1.67$, 95% CI = $[-4.64, 1.29]$, $t(16) = -1.20$, $P = 0.25$) or OMP ($\beta = -1.09$, 95% CI = $[-4.82, 2.65]$, $t(16) = -0.62$, $P = 0.54$; Fig. 4c).

Consistent with the interpretation that C_{IM} was already the strongest source of variance in the learned manifold, it makes sense that additional neural alignment along C_{IM} did not further benefit BCI control. Such realignment was not possible for the OMP, and there was no significant BCI learning or neural realignment for this session on

https://doi.org/10.1038/s41593-026-02311-2

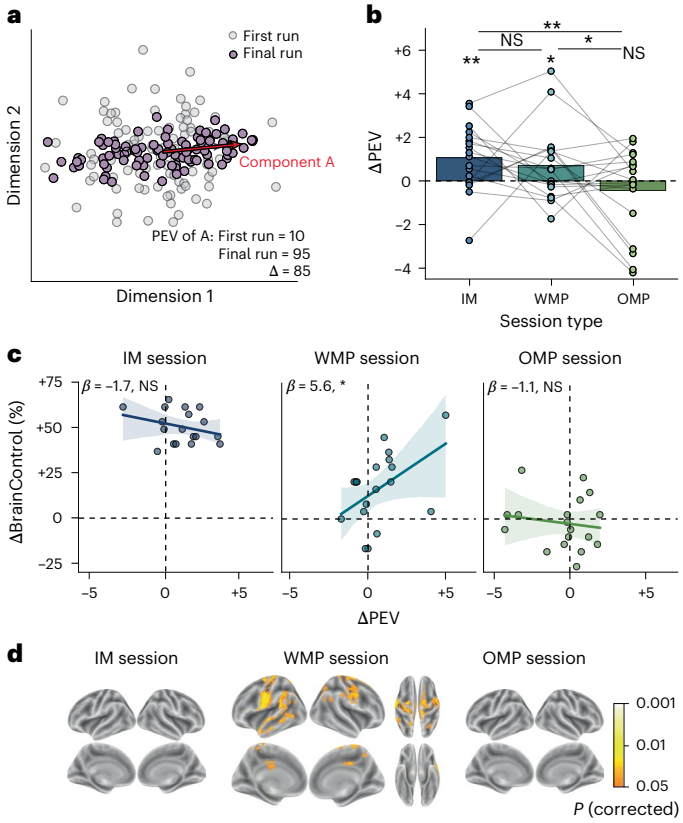

Fig. 4 | Changes in neural activity supporting BCI learning. a, Simulated data showing a component (red vector) that explains 10% of the explainable variance initially (gray points) but 95% after learning (purple points). **b**, Change in the percentage of explained variance (ΔPEV) over neurofeedback training for each session type. Bars represent the mean across participants ($N = 18$); points represent individual participants; lines connect the points from the same participant across the three sessions. Statistical significance was assessed using nonparametric randomization tests (10,000 iterations, one-tailed); 95% CIs were estimated via bootstrap resampling (10,000 iterations). IM ($P = 0.005$) and WMP ($P = 0.042$) sessions, but not OMP ($P = 0.35$) sessions, showed significant neural changes. Neural changes during the IM ($P = 0.005$) and WMP ($P = 0.034$) sessions were greater than those during OMP. **c**, Predicting BCI learning (ΔBrainControl) from changes in neural realignment (ΔPEV) within session type. β and P values were computed using linear regression; lines show the regression estimate and error bands show 95% CIs estimated with 10,000 bootstrap resamples. **d**, Brain maps for each session type depicting searchlights with significantly improved location decoding. Statistical significance was assessed using fMRIB's Software Library's randomise function (one-sided nonparametric randomization test, corrected with threshold-free cluster enhancement). In the WMP session, decoding accuracy of arena location improved across neurofeedback runs in areas spanning the primary motor and somatosensory cortices and superior temporal lobe. See Supplementary Fig. 10 for analysis schematic. $^{**}P < 0.01$, $^{*}P < 0.05$, NS $P \geq 0.1$.

average. The WMP session always occurred after the IM session (either immediately or after OMP) and so participants had to realign activity from the IM to WMP components in order to gain control. This realignment was possible, given that WMP was on the intrinsic manifold, and the extent to which it occurred determined the amount of BCI learning in this session.

Brain-wide changes associated with BCI learning

We also investigated another signature of BCI learning that was unique to the WMP condition: improved task representation. Learning to control activity in one brain region can also alter nontarget brain regions, such as those involved in cognitive strategies and modulation of target regions[47–50]. We thus used an exploratory searchlight analysis to test for improved task representation in regions across the whole brain. Using ridge regression with nested cross-validation, we trained decoders on multi-voxel patterns from searchlights centered on each voxel in the brain to predict the avatar's coordinates in the game arena (Supplementary Fig. 10).

WMP training resulted in significantly improved task decoding in several brain regions (Fig. 4d), including the somatosensory cortex and areas of the primary motor cortex related to hand and finger movement ($P < 0.05$, threshold-free cluster enhancement corrected). Surprisingly, only 3.2% of the searchlights that showed greater decoding accuracy after BCI learning overlapped with the navigation mask. No regions showed a significant increase in task decoding during the IM or OMP sessions. This was expected for OMP given that there were no indicators of BCI learning for this session type in the first place. In the IM session, neural changes were not found to predict BCI learning, suggesting that this component was already controllable. This interpretation is consistent with the lack of decoding improvement, which indicates that modulatory patterns did not need to change from their original activity to support task performance (Fig. 4d), as well as with the relatively weaker evidence for neural realignment over subselection in models of IM learning compared to WMP learning (Supplementary Fig. 9). Together with the searchlight analysis, these findings indicate that learning a new mapping on an existing manifold alters neural activity within the targeted regions and throughout the cortex.

Discussion

The proliferation of human BCI technologies has been hindered to date by the slowness and variability of learning across individuals. These challenges persist whether the BCI uses invasive or noninvasive neural recordings[8]. Previous rt-fMRI neurofeedback studies have typically required four to ten training sessions to achieve robust changes in perception and cognition[19–23,51,52]. Moreover, one-third of neurofeedback users remain unable to change their brain activity[18,24]. In the current study, all participants successfully learned to self-regulate brain activity to navigate a virtual avatar when neurofeedback was administered along their intrinsic neural manifold. Once participants achieved this control, they were able to relearn—within one session—how to navigate the avatar after a perturbation in the mapping that stayed within their manifold. However, participants were unable to relearn control with the same amount of training when they had to generate outside-manifold brain activity. As in other neurofeedback studies[19,53], successful learning occurred without explicit awareness and using highly idiosyncratic mental strategies across participants (Supplementary Table 1). This highlights the brain's ability to self-modulate via feedback beyond relying upon specific behavioral techniques.

This enhanced on-manifold learning parallels studies in which humans and nonhuman primates learned a novel behavior (cursor control) via an invasive BCI in the motor cortex[25,27–29]. One key difference in the current study is the use of a diffusion-based manifold learning method that was necessitated by the autocorrelation and noise of fMRI data. Given T-PHATE's ability to recover complex signals within noisy samples (Supplementary Figs. 2 and 11), it may benefit a broader range of applications—including invasive studies of language or sensorimotor processes, and noninvasive studies that use neurofeedback to train attention or emotion regulation and to alleviate psychiatric or neurological symptoms—which have so far relied on simpler representations or linear methods[1–7,20,50,54–57].

One reason that manifold learning methods like T-PHATE enhance BCI learning may be that they yield high-quality feedback that promotes durable changes in brain activity and behavior. Previous work found that BCI learning of within-manifold perturbations is supported by neural reassociation—the remapping of existing activity patterns to a new behavior[26]. By contrast, BCI learning in the current study was supported by neural realignment—the generation of novel activity

patterns to drive a behavior. Although neural reassociation can occur more quickly, neural realignment is the behaviorally optimal solution and has been shown to emerge with longer training times[26]. In addition to using T-PHATE to extract a richer, more informative manifold[40], our framework may have enabled rapid neural realignment by encouraging incremental learning with a staircasing procedure, which has been linked to generating new activity patterns[9,27,58–62]. Further, it could be that neural realignment occurs at a macro scale (that is, measurable in population-level blood oxygen level-dependent signals with whole-brain fMRI) before it becomes detectable at the micro scale of direct neuronal recordings. Future studies could examine different temporal and spatial scales of neural realignment to fine-tune BCI procedures and promote behaviorally optimal solutions.

Why does rapid relearning occur after the WMP but not the OMP? The capacity to self-modulate brain activity depends upon how the target brain activity relates to the geometry of the extant brain activity. Our finding that participants failed to learn an OMP in a single session converges with computational models and studies of nonhuman primates, in which such learning can take approximately ten times as much training as a WMP[27,59,61,63]. The consistency of this finding is striking in that it generalizes across species, tasks, neural recording modalities, manifold learning techniques and ways of defining what is outside the manifold. In the present study, we defined the within-manifold component as accounting for the second greatest variance in brain activity and the outside-manifold component as the least variance in the 20-dimensional space (Supplementary Fig. 12). This definition ensured that participants could generate decodable movement patterns along both perturbation components (Supplementary Fig. 13), but that the WMP would require a less dramatic change to the geometry of the existing brain activity (that is, from the intuitive component training, which all participants learned first) than the outside-manifold component.

Because these components were defined relative to one another, the distinction between 'within' and 'outside' the manifold is more continuous than binary, offering differing degrees of learnability. For example, if we had set the third component as our OMP, it may have been learnable because, for many participants, the third component still explained over 10% of variance in the data; in contrast, if we had used a 100-dimensional space, the 100th ranked component may have been less learnable than the 20th used here for OMP because it would have explained minimal variance (Supplementary Fig. 12). Future studies could identify ways of targeting outside-manifold activity patterns of progressive difficulty and offering training that traverses this landscape more efficiently. Further, by defining multiple BCI mappings that are able to be modulated (that is, training multiple on-manifold components simultaneously), future work could explore optimal weighted combinations of components to enhance decoding.

The difficulty of outside-manifold learning offers a potential explanation for why earlier attempts at BCI learning with rt-fMRI neurofeedback have often required multiple sessions to achieve consistent success across participants[19–23,51,52]. The need for multiple sessions is especially true when seeking to train fine-grained control of multivariate activity patterns rather than up- or down-modulation of univariate activity. By designating neurofeedback targets agnostic to the intrinsic manifold, these studies may have inadvertently asked participants to generate outside-manifold activity patterns. This may be desirable in some cases, for example, if the goal is to change a disordered manifold to alleviate psychiatric or neurological symptoms[55,56]. In such cases, better characterizing the initial manifold would support incremental training of outside-manifold activity. Indeed, the intrinsic manifold learned with T-PHATE remained relatively stable over time (Supplementary Fig. 14), suggesting its promise for multiday learning and incremental outside-manifold training.

If the goal of a BCI is to interface the human brain with computer technologies for communication or occupational applications, rather than altering the manifold itself, leveraging the intrinsic manifold will be most efficient. In this case, training multiple on-manifold components simultaneously could support more complex BCIs with two or more dimensions of control (for example, direction plus speed in the current game). Our findings offer a manifold-guided route for improving basic and translational BCIs. We anticipate that these improvements will increase the efficacy of BCIs on wearable devices that are cheaper and easier to scale for public benefit.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41593-026-02311-2.

References

1. Hochberg, L. R. et al. Neuronal ensemble control of prosthetic devices by a human with tetraplegia. *Nature* **442**, 164–171 (2006).
2. Hochberg, L. R. et al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* **485**, 372–375 (2012).
3. Collinger, J. L. et al. High-performance neuroprosthetic control by an individual with tetraplegia. *Lancet* **381**, 557–564 (2013).
4. Willett, F. R., Avansino, D. T., Hochberg, L. R., Henderson, J. M. & Shenoy, K. V. High-performance brain-to-text communication via handwriting. *Nature* **593**, 249–254 (2021).
5. Willett, F. R. et al. A high-performance speech neuroprosthesis. *Nature* **620**, 1031–1036 (2023).
6. Metzger, S. L. et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature* **620**, 1037–1046 (2023).
7. Silva, A. B., Littlejohn, K. T., Liu, J. R., Moses, D. A. & Chang, E. F. The speech neuroprosthesis. *Nat. Rev. Neurosci.* **25**, 473–492 (2024).
8. Patrick-Krueger, K. M., Burkhart, I. & Contreras-Vidal, J. L. The state of clinical trials of implantable brain–computer interfaces. *Nat. Rev. Bioeng.* **3**, 1–18 (2024).
9. Orsborn, A. L. et al. Closed-loop decoder adaptation shapes neural plasticity for skillful neuroprosthetic control. *Neuron* **82**, 1380–1393 (2014).
10. Sussillo, D., Stavisky, S. D., Kao, J. C., Ryu, S. I. & Shenoy, K. V. Making brain–machine interfaces robust to future neural variability. *Nat. Commun.* **7**, 13749 (2016).
11. Luo, S. et al. Stable decoding from a speech BCI enables control for an individual with ALS without recalibration for 3 months. *Adv. Sci.* **10**, 2304853 (2023).
12. Allison, B. Z. & Neuper, C. in *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction* (eds Tan, D. S. & Nijholt, A.) 35–54 (Springer, 2010).
13. Jarosiewicz, B. et al. Advantages of closed-loop calibration in intracortical brain–computer interfaces for people with tetraplegia. *J. Neural Eng.* **10**, 046012 (2013).
14. Perge, J. A. et al. Intra-day signal instabilities affect decoding performance in an intracortical neural interface system. *J. Neural Eng.* **10**, 036004 (2013).
15. Leeb, R. et al. Transferring brain–computer interfaces beyond the laboratory: Successful application control for motor-disabled users. *Artif. Intell. Med.* **59**, 121–132 (2013).
16. Min, B.-K., Chavarriaga, R. & Millán, J. del R. Harnessing prefrontal cognitive signals for brain–machine interfaces. *Trends Biotechnol.* **35**, 585–597 (2017).
17. Thompson, M. C. Critiquing the concept of BCI illiteracy. *Sci. Eng. Ethics* **25**, 1217–1233 (2019).
18. Haugg, A. et al. Predictors of real-time fMRI neurofeedback performance and improvement – a machine learning mega-analysis. *NeuroImage* **237**, 118207 (2021).

19. Shibata, K., Watanabe, T., Sasaki, Y. & Kawato, M. Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science* **334**, 1413–1415 (2011).
20. deBettencourt, M. T., Cohen, J. D., Lee, R. F., Norman, K. A. & Turk-Browne, N. B. Closed-loop training of attention with real-time brain imaging. *Nat. Neurosci.* **18**, 470–475 (2015).
21. deBettencourt, M. T., Turk-Browne, N. B. & Norman, K. A. Neurofeedback helps to reveal a relationship between context reinstatement and memory retrieval. *NeuroImage* **200**, 292–301 (2019).
22. Peng, K. et al. Inducing representational change in the hippocampus through real-time neurofeedback. *Philos. Trans. R. Soc. B* **379**, 20230091 (2024).
23. Iordan, C. R., Ritvo, V. J. H., Norman, K. A., Turk-Browne, N. B. & Cohen, J. D. Sculpting new visual categories into the human brain. *Proc. Natl Acad. Sci. USA* **121**, e2410445121 (2024).
24. Stoeckel, L. E. et al. Optimizing real time fMRI neurofeedback for therapeutic discovery and development. *NeuroImage Clin.* **5**, 245–255 (2014).
25. Sadtler, P. T. et al. Neural constraints on learning. *Nature* **512**, 423–426 (2014).
26. Golub, M. D. et al. Learning by neural reassociation. *Nat. Neurosci.* **21**, 607–616 (2018).
27. Oby, E. R. et al. New neural activity patterns emerge with long-term learning. *Proc. Natl Acad. Sci. USA* **116**, 15210–15215 (2019).
28. Sakellaridi, S. et al. Intrinsic variable learning for brain-machine interface control by human anterior intraparietal cortex. *Neuron* **102**, 694–705 (2019).
29. Hwang, E. J., Bailey, P. M. & Andersen, R. A. Volitional control of neural activity relies on the natural motor repertoire. *Curr. Biol.* **23**, 353–361 (2013).
30. Hennig, J. A. et al. Constraints on neural redundancy. *eLife* **7**, e36774 (2018).
31. Gallego, J. A. et al. Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nat. Commun.* **9**, 4233 (2018).
32. Degenhart, A. D. et al. Stabilization of a brain–computer interface via the alignment of low-dimensional spaces of neural activity. *Nat. Biomed. Eng.* **4**, 672–685 (2020).
33. Dabagia, M., Kording, K. P. & Dyer, E. L. Aligning latent representations of neural activity. *Nat. Biomed. Eng.* **7**, 337–343 (2023).
34. Karpowicz, B. M. et al. Stabilizing brain-computer interfaces through alignment of latent dynamics. *Nat. Commun.* **16**, 4662 (2025).
35. Menéndez, J. A. et al. A theory of brain-computer interface learning via low-dimensional control. *eLife* **14**, 2024.04.18.589952 (2025).
36. Gallego, J. A., Makin, T. R. & McDougle, S. D. Going beyond primary motor cortex to improve brain–computer interfaces. *Trends Neurosci.* **45**, 176–183 (2022).
37. Coifman, R. R. et al. Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. *Proc. Natl Acad. Sci. USA* **102**, 7426–7431 (2005).
38. Moon, K. R. et al. Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Curr. Opin. Syst. Biol.* **7**, 36–46 (2018).
39. Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
40. Busch, E. L. et al. Multi-view manifold learning of human brain-state trajectories. *Nat. Comput. Sci.* **3**, 240–253 (2023).
41. Lamsam, L. et al. The human claustrum tracks slow waves during sleep. *Nat. Commun.* **15**, 8964 (2024).
42. Mantel, N. Ranking procedures for arbitrarily restricted observation. *Biometrics* **23**, 65–78 (1967).
43. Duque, A. F., Morin, S., Wolf, G. & Moon, K. R. Extendable and invertible manifold learning with geometry regularized autoencoders. In *2020 IEEE International Conference on Big Data* 5027–5036 (IEEE, 2020).
44. Huang, J. et al. Learning shared neural manifolds from multi-subject fMRI data. In *2022 IEEE 32nd Interntaional Workshop on Machine Learning for Signal Processing MLSP* 01–06 (IEEE, 2022).
45. Ellis, C. T., Baldassano, C., Schapiro, A. C., Cai, M. B. & Cohen, J. D. Facilitating open-science with realistic fMRI simulation: validation and application. *PeerJ* **8**, e8564 (2020).
46. Hill, N. J., Häuser, A.-K. & Schalk, G. A general method for assessing brain–computer interface performance and its limitations. *J. Neural Eng.* **11**, 026018 (2014).
47. Scheinost, D. et al. Orbitofrontal cortex neurofeedback produces lasting changes in contamination anxiety and resting-state connectivity. *Transl. Psychiatry* **3**, e250–e250 (2013).
48. Ruiz, S., Buyukturkoglu, K., Rana, M., Birbaumer, N. & Sitaram, R. Real-time fMRI brain computer interfaces: Self-regulation of single brain regions to networks. *Biol. Psychol.* **95**, 4–20 (2014).
49. Emmert, K. et al. Meta-analysis of real-time fMRI neurofeedback studies using individual participant data: How is brain regulation mediated? *NeuroImage* **124**, 806–812 (2016).
50. Cohen Kadosh, K. et al. Using real-time fMRI to influence effective connectivity in the developing emotion regulation network. *NeuroImage* **125**, 616–626 (2016).
51. Watanabe, T., Sasaki, Y., Shibata, K. & Kawato, M. Advances in fMRI real-time neurofeedback. *Trends Cogn. Sci.* **21**, 997–1010 (2017).
52. Taschereau-Dumouchel, V. et al. Towards an unconscious neural reinforcement intervention for common fears. *Proc. Natl Acad. Sci. USA* **115**, 3470–3475 (2018).
53. Lubianiker, N. et al. Upregulation of reward mesolimbic activity and immune response to vaccination: a randomized controlled trial. *Nat. Med.* **32**, 572–581 (2026).
54. Hennig, J. A. et al. Learning is shaped by abrupt changes in neural engagement. *Nat. Neurosci.* **24**, 727–736 (2021).
55. Shanechi, M. M. Brain–machine interfaces from motor to mood. *Nat. Neurosci.* **22**, 1554–1564 (2019).
56. Oganesian, L. L. & Shanechi, M. M. Brain–computer interfaces for neuropsychiatric disorders. *Nat. Rev. Bioeng.* **2**, 653–670 (2024).
57. Koizumi, A. et al. Fear reduction without fear through reinforcement of neural activity that bypasses conscious exposure. *Nat. Hum. Behav.* **1**, 1–7 (2016).
58. Sadtler, P. T., Ryu, S. I., Tyler-Kabara, E. C., Yu, B. M. & Batista, A. P. Brain–computer interface control along instructed paths. *J. Neural Eng.* **12**, 016015 (2015).
59. Feulner, B. & Clopath, C. Neural manifold under plasticity in a goal driven learning behaviour. *PLoS Comput. Biol.* **17**, e1008621 (2021).
60. Golub, M. D., Chase, S. M., Batista, A. P. & Yu, B. M. Brain–computer interfaces for dissecting cognitive processes underlying sensorimotor control. *Curr. Opin. Neurobiol.* **37**, 53–58 (2016).
61. Chang, J. C., Perich, M. G., Miller, L. E., Gallego, J. A. & Clopath, C. De novo motor learning creates structure in neural activity that shapes adaptation. *Nat. Commun.* **15**, 4084 (2024).
62. Rajeswaran, P., Payeur, A., Lajoie, G. & Orsborn, A. L. Assistive sensory-motor perturbations influence learned neural representations. Preprint at *bioxRiv* https://doi.org/10.1101/2024.03.20.585972 (2024).
63. Payeur, A., Orsborn, A. L. & Lajoie, G. Neural manifolds and learning regimes in neural-interface tasks. Preprint at *bioRxiv* https://doi.org/10.1101/2023.03.11.532146 (2023).
64. Wallace, G. et al. RT-Cloud: a cloud-based software framework to simplify and standardize real-time fMRI. *NeuroImage* **257**, 119295 (2022).

Methods

Participants

Twenty young adults were recruited from the New Haven community. All participants provided informed consent to an experimental protocol approved by the Yale University Institutional Review Board. Data from two participants were excluded from neurofeedback analyses (one for falling asleep in multiple fMRI runs and one because of a scanner malfunction). One participant dropped out after the first session due to discomfort in the scanner and was replaced with an additional participant. This resulted in a final cohort of 18 participants (9 female; aged 19–35 years, mean age of 25.8 years). This sample size is greater than, or comparable with, other multiday rt-fMRI studies[19,20,22,23,48,52,57,65]. There was no separate control group because all comparisons and baselines are within-participant over time and across repeated-measures conditions. Participants completed four or five fMRI sessions lasting 1–1.5 hours each and were compensated for their participation (US$20 per hour plus retention incentives: US$5 per session cumulative bonus for each return session and US$40 for completion).

Task design

The stimulus was a custom video game programmed with C# for Unity (https://unity3d.com/) with a human avatar in a virtual outdoor arena environment. Participants steered the avatar from a randomly generated starting location to a goal flag across the arena; we refer to one repetition of this task as a 'trial'. A yellow line was always visible, highlighting the shortest path from the avatar's current location to the goal location. For each trial within a run, the flag was placed at the same distance from the avatar's spawning location but at a randomly selected angle. Each run consisted of approximately 10 min of continuous fMRI data collection and resulted in a variable number of trials depending upon how quickly trials were completed. The distance between the spawning location and the flag increased by 5% per run. Scanner triggers were synchronized with the game engine via a custom PsychoPy/TCP script. Task instructions were displayed for 20 s at the start of each neurofeedback run. Full technical specifications of the display and software pipeline are in Supplementary Methods. Task software and sample videos are available at: github.com/ericabusch/avatarRT_task.

fMRI experiment design

The full study comprised a total of four or five sessions per participant. Four functional runs were collected during each session and participants were offered a break between runs. Sessions were scheduled on separate days with a target of 24 h between sessions and a maximum of 48 hours between neurofeedback sessions (mean 27.24 hours, range 13–48 hours). All sessions had to be completed within seven days.

fMRI data were acquired using a 3 T Siemens Prisma scanner with a 64-channel head coil at the Brain Imaging Center at Yale University. Blood oxygen level-dependent signal data were collected with an echo-planar imaging (EPI) sequence (repetition time (TR) = 2 s; echo time = 30 ms; voxel size = 3 mm isotropic; 37 axial slices; flip angle = 71°; IPAT GRAPPA acceleration factor = 2). Functional runs were of variable length due to the self-guided nature of the task; the experimenter aimed to end the task manually at the completion of a trial close to 300 TRs (10 min), resulting in an average of 301.4 TRs per run (range 150–336). Spin-echo field maps and high-resolution T1/T2 anatomical images were also collected. The full acquisition parameters are in Supplementary Methods.

Neurosynth navigation network. Neurofeedback training targeted a large network of voxels involved in self-directed navigation downloaded from neurosynth.org (ref. 66). Searching for 'navigation' yielded 77 published studies and an association test *z*-statistic map covering regions linked to goal-directed navigation, first-person perspective, landmark sequencing and route knowledge—regions previously shown to support decoding of navigation goals[67–70] and paralleled in our findings from the joystick session. The *z*-map in 2-mm Montreal Neurological Institute (MNI) standard space was aligned to each participant's native space functional image, thresholded (bottom 10% of *z*-statistics removed) and cluster-filtered (minimum 10 voxels per cluster), resulting in a mean of 1,354 voxels per participant (range 1,058–1,599; Supplementary Fig. 1). The number of voxels included did not significantly influence a participant's learning in any of the three session types (Supplementary Fig. 16).

The choice to provide neurofeedback from the navigation network was related to our video game task and its ability to drive variable activity in the regions of this network for participants to latch onto, rather than to any special functional properties of specific regions. We speculate this may be a more general principle: given a task that evokes meaningful and dynamic activity in a brain region, manifold-based BCI learning may enable rapid control of this region. The extent to which these findings generalize in this way to other tasks and brain systems is an area for future investigation.

Joystick session. In session 1, participants practiced the video game task by navigating the avatar with an MR-compatible joystick (Current Designs Tethyx), steering the avatar from start to goal within a time limit (60 s in the first run, +10 s per subsequent run). Participants were instructed to explore the environment freely—not necessarily heading directly to the goal—which broadly sampled brain activity across movement patterns and arena locations. As each trial varied in duration, the number of trials per run also varied across participants and runs (mean 19.1, range 8–26). A trial ended upon reaching the flag, after which the game paused for 6 s before the next trial. The technical specifications and practice procedure for the joystick are in Supplementary Methods.

Neurofeedback calibration. Data collected during the joystick session were used to initialize the manifold-based neurofeedback procedure used in subsequent sessions. fMRI data processing used fMRI Expert Analysis Tool version 6.00[71], part of fMRIB's Software Library (FSL) version 6.0.5. EPI and anatomical images were skull-stripped using the Brain Extraction Tool[72]. Susceptibility-induced distortions were measured via the opposing-phase spin-echo volumes and corrected using FSL's topup function[73]. Each functional run was high-pass filtered with a 100-s period cutoff, corrected for head motion with Motion Correction using FLIRT (MCFLIRT)[74], corrected for slice timing and smoothed spatially with a Gaussian kernel (5-mm full-width at half-maximum; FWHM). Then, functional images were registered to the participant's T1-weighted anatomical scan using boundary-based registration[75] and to a 2-mm MNI standard brain using 12 degrees of freedom.

Game outputs from Unity were preprocessed, temporally downsampled, and time-lagged to match the fMRI sampling rate and hemodynamic lag. The fMRI data from the joystick-based navigation trials were divided into training (80%) and testing (20%) sets. The training set was used to fit a 20-dimensional T-PHATE embedding[40]. We selected this dimensionality to (1) obtain low decoding error and high representational similarity (Fig. 1 and Supplementary Fig. 2) and (2) maximize the proportion of explained variance by highly ranked components while maintaining low-ranked components with some explained variance (Supplementary Fig. 12). This choice was made a priori based on pilot participants to avoid overfitting and was fixed across all participants.

We defined three latent components of the T-PHATE manifold: the eigenvectors associated with the greatest (1st component), second greatest (2nd component) and smallest eigenvalues (20th component). These components C corresponded with the IM (C_{IM}), the WMP (C_{WMP}) and the OMP (C_{OMP}) components used for neurofeedback. Participants' eigenspectra varied considerably in the amount of variance explained across components; thus, selecting component 20 for OMP, as opposed

to component 3, for example, helped ensure that this perturbation was comparably challenging for all participants (Supplementary Fig. 12).

Finally, we trained an MRAE with two training objectives: to reconstruct the input fMRI data in voxel space and to learn a latent representation that closely resembles the data's ground-truth T-PHATE manifold geometry. This training allows the MRAE to extend the T-PHATE manifold to incoming data points collected in real time. All T-PHATE, MRAE training, and component determination was performed on the training set of the joystick session data (80%).

After the autoencoder was trained on the training set, its weights were frozen and the testing set was passed through the encoder f, extracting its embedding in the T-PHATE manifold. We then projected the embedded data onto C_{IM}, C_{WMP} and C_{OMP}, as previously defined, to measure the distribution of untrained data along the components of the manifold. The 1st and 99th percentiles of the distributions of untrained data along each component were fixed to represent turning −90° and 90° in the game, respectively. Sample projection mappings from this procedure are shown in Supplementary Fig. 13.

Neurofeedback sessions. After the joystick session, participants returned to the scanner for three or four fMRI neurofeedback sessions. During each neurofeedback session, participants completed four functional runs (~10 min each) during which their task was to use their brain to make the avatar walk to the flag as directly as possible. As in the joystick session, they were given 60 s to complete each trial in the first run and +10 s for each subsequent run.

The participants were instructed to "generate a mental state that made the avatar walk to the flag". We emphasized to participants that there was no single 'right' strategy to accomplish the task; they would need to try a variety of strategies and possibly switch between strategies. Indeed, participants reported a wide range of strategies in a post-study questionnaire (Supplementary Table 1). They were reminded at the start of each run that the avatar's movement reflected their brain states over the past 4–6 s, and that the avatar's movement was a delayed reaction to their thoughts.

We outlined the staircasing training procedure to participants as followed. Participants were told that there were 'bumpers' (akin to those on a bowling lane) to aid in keeping the avatar walking relatively straight to the goal at first (that is, at lower game levels). As brain activity resulted in more accurate movement, participants would 'level up' and the bumpers would go down. Thus, the avatar could veer farther off-course and require brain activity to be more accurate in order to go straight. Participants were instructed that their goal was to continue 'leveling up'. The first neurofeedback run started at level 0, and at the end of each trial, participants were presented with a feedback display explaining performance as the percentage of distance traveled over and above the shortest possible path to the goal. They were then informed if this error was higher, lower, or equal to the average error of their previous three trials and encouraged to continue leveling up. The staircasing procedure began after the third trial. Participants 'leveled up' if their error decreased, 'leveled down' if their error increased, or stayed the same if their error matched earlier trials.

At the start of each neurofeedback run, there was a 20-s (10 TR) delay before the first trial. Task instructions were presented during this time, reminding participants that their goal was to get to the flag as quickly as possible and to remember that there would be a delay between their brain state and the avatar's movement. Then the first trial began and participants completed as many trials as they could in approximately 10 min (300 TRs), with a 6-s break between trials. After that, the scan was ended manually by the experimenter and participants rested until the next run. The first trial of the next run began at the final 'level' of the previous run.

Real-time data processing. Every 2 s, a whole-brain volume was acquired as a DICOM image and sent via fiber network from the local Siemens scanner console to a remote, HIPAA-aligned high-performance compute cluster. The first volume of each run V_1 was smoothed with a 5-mm FWHM Gaussian kernel to match the preprocessed data from the joystick session and saved as that run's reference volume, to which all subsequent volumes were aligned. FLIRT[74] was used to align V_1 to V_{ref}, a reference mean functional image from the joystick session. The resulting transformation matrix M served to transform each subsequent volume to V_{ref}, the space in which the region of interest mask of navigation-related brain voxels was defined. Each volume V_t was smoothed, motion-corrected with MCFLIRT to V_1, aligned to V_{ref} via M and masked to extract voxels in the feedback region of interest.

The first 10 brain volumes in each run were used to estimate mean and standard deviation parameters for each voxel. These parameters were then used to normalize (that is, z-score) the activity of each voxel in the 11th volume and updated with each time step to reflect previous volumes. That is, for a given voxel's activity vector $\mathbf{x}$ consisting of t timepoints, its normalized activity at time t, v_t, was determined as follows:

$$v_t = \frac{x_t - \mu}{\sigma}, \quad \text{where } \mu = \frac{1}{t-1}\sum_{k=1}^{t-1} x_k \quad \text{and} \quad \sigma = \sqrt{\frac{1}{t-1}\sum_{k=1}^{t-1}(x_k - \mu)^2} \quad (1)$$

After selecting voxels from the navigation mask and normalizing them, the data V_t were passed through the trained MRAE encoder f, such that $f(V_t)$ yielded the embedding of V_t onto the T-PHATE manifold. Finally, $f(V_t) \cdot C_{\mathrm{ses}}$ mapped the T-PHATE-embedded data onto the feedback component for that session C_{ses}, which determined the angle of movement $\alpha \in (-90°, 90°)$. The value of α was transmitted back to the scanner presentation computer via the fiber network link to a script running PsychoPy, which was time-locked in communication with the Unity video game via a TCP connection and used to determine the next step, γ, such that:

$$\gamma = \alpha_{\mathrm{ideal}} - \alpha_{\mathrm{ideal}} \times \mathrm{BrainControl} + \alpha \times \mathrm{BrainControl} \quad (2)$$

where α_{ideal} was the angle that would keep the avatar along the straightest path to the goal and BrainControl was the proportion of control α had over the avatar's movement.

At the end of each trial T, the error ϵ_T was computed as the ratio of the total distance traveled by the avatar relative to the shortest possible path between the start and goal locations. BrainControl was then staircased depending upon the ϵ_T and $\epsilon_{\mathrm{prior}}$ (defined as the average error over the prior two trials)[47], such that the BrainControl at trial $T+1$ was determined by:

$$\mathrm{BrainControl}_{T+t} = \begin{cases} \mathrm{BrainControl}_T + \mathrm{step}, & \text{if } \epsilon_T < \epsilon_{\mathrm{prior}} \\ \mathrm{BrainControl}_T - \mathrm{step}, & \text{if } \epsilon_T > \epsilon_{\mathrm{prior}} \\ \mathrm{BrainControl}_T, & \text{otherwise} \end{cases} \quad (3)$$

Participants were given feedback along the intuitive component for the four runs of the second session (first neurofeedback session). BrainControl did not plateau during this session for one participant and so they repeated the IM session in their second visit and completed five sessions total instead of four. For subsequent neurofeedback sessions, participants began the first trial of the first run with BrainControl set to its final value from the end of their IM session plus 20%. They began with one run of IM training to recalibrate the feedback procedure before receiving WMPs or OMPs in the second run.

We always began neurofeedback with IM in the second session because this component was most consistent with the participant's intrinsic manifold learned in the first session. We reasoned that starting with one of the perturbations (WMP or OMP) instead may have altered the intrinsic manifold, rendering a subsequent IM session no longer the intuitive mapping. By first receiving neurofeedback in what

we hypothesized would be the easiest condition, participants also had the opportunity to get familiar with the task and with modulating their brain activity. Perturbation order was counterbalanced across participants to help control for potential temporal confounds, such that half of participants received WMP during the third session and OMP during the fourth session, and the other half received OMP then WMP. The comparison of these conditions was the key test of the hypothesis that WMP would be more learnable than OMP. We verified that there was no effect of session order (Supplementary Fig. 6). To validate our BCI learning effects, we simulated 20 realistic, null participants.

Neural manifold learning

A key innovation of this study is the use of diffusion geometry methods[37–39] to perform neural manifold learning and define an intrinsic manifold of neural activity to target with neurofeedback. Earlier studies using low-dimensional data representations as input for BCIs relied on linear dimensionality reduction methods (for example, PCA, factor analysis) to approximate the intrinsic manifold of brain activity as a linear subspace. More recent studies have shown that high-dimensional neural population activity can be summarized within a lower-dimensional nonlinear manifold[40,76,77]. The linear approximation of a nonlinear manifold in previous work may not fully capture the intrinsic manifold of brain activity and what activity patterns are on versus outside the manifold (Supplementary Fig. 11).

fMRI signals have high noise in both space and time, so extracting behaviorally meaningful brain activity from fMRI often requires aggregating activity across many trials or participants, or reducing dimensionality via averaging, to improve the signal-to-noise ratio. Recently, we developed and applied a new manifold learning algorithm (T-PHATE) to represent fMRI activity from single participants during cognitively complex tasks[40]. Qualitatively, T-PHATE's strength for use with fMRI data is reflected in the organized clusters and trajectories shown in the T-PHATE embeddings of fMRI activity from the joystick task (Fig. 1b) relative to the disorganized and shattered PCA embeddings of the same data (Supplementary Fig. 2a,b). Quantitatively, this is clear in the higher decoding error and lower representational similarity of PCA embeddings (Supplementary Fig. 2c,d); by retaining a greater number of components, linear dimensionality reduction representations begin to approximate the quality of nonlinear embeddings, suggesting that they inadequately characterize the nonlinearities of the brain activity manifold.

We designed a three-step procedure to learn and apply nonlinear manifolds in rt-fMRI: (1) learning the intrinsic manifold of single-participant fMRI activity via T-PHATE, which builds upon the classic diffusion maps algorithm[37]; (2) training an MRAE to quickly embed new brain volumes onto the learned T-PHATE manifold and projecting them to the BCI mappings; and (3) identifying within- and outside-manifold components based on the components of the MRAE latent space as BCI mappings. These three steps are elaborated below.

T-PHATE. The T-PHATE algorithm is a dual diffusion-based manifold learning method for discovering data geometry and latent dynamics from complex, biological timeseries data. We used T-PHATE to learn a low-dimensional manifold of brain activity in our study because T-PHATE accurately captures cognitively meaningful signals and task information in single-participant fMRI data[40,78]. T-PHATE takes as input multi-voxel activity patterns (that is, a matrix with timepoints/samples as rows and voxels/features as columns) and learns two 'views' among pairs of samples: a PHATE-based[39] affinity matrix and a temporal autocorrelation-based affinity matrix. Extending the classic diffusion maps algorithm[37], PHATE provides an accurate, de-noised representation of local similarities (via an α-decay kernel) and global relationships (via a diffusion potential distance), without many assumptions about a hypothesized manifold structure[39]. The autocorrelation matrix models the temporal dynamics across data samples by computing the correlation of each voxel's timeseries with lagged versions of itself. This kernel captures both the temporal dynamics related to the measured blood oxygen level-dependent signal and those related to the temporally diffuse cognitive processes occurring within a given multi-voxel pattern. The PHATE and autocorrelation views are converted into transition probability matrices and then combined with alternating diffusion, before embedding into an m-dimensional representation using metric multi-dimensional scaling. T-PHATE embeddings were performed for individual participants. fMRI timeseries data input to the T-PHATE algorithm were masked to include only voxels in the navigation mask, z-scored within voxel and fMRI run, and concatenated across runs.

MRAE. T-PHATE embeddings improved access to task-relevant information from the brain data, as shown by the improvements in location decoding and arena representation (Fig. 1c,d and Supplementary Fig. 2c,d). However, a key limitation of nonlinear methods such as T-PHATE is that the learned manifolds are not readily extensible to new data samples. To avoid needing to re-fit the T-PHATE algorithm with each incoming fMRI volume in real time (which would be prohibitively slow), we trained an MRAE to (1) reconstruct the voxel-resolution fMRI data and (2) learn the correspondence between input fMRI data and the corresponding points on the T-PHATE manifold[43,44]. This latter operation was trained via a manifold regularization penalty, which minimizes a geometric loss function of the distance between fMRI samples in the autoencoder's bottleneck and the initial T-PHATE embedding. The encoder thus learned a nonlinear mapping between the fMRI data in voxel space and the fMRI data in T-PHATE space; after MRAE training was complete, we could use the encoder to embed new fMRI samples onto the T-PHATE manifold in real time and it faithfully interpolated along that manifold.

The MRAE encoder (f) and decoder (g) each had three fully connected layers, with a bottleneck latent space layer in between. As input, we gave the model both the training data Y (in this experiment, 80% of the data from the joystick session) and the T-PHATE embedding of the training data ($E(Y)$). The input layer of f and output layer of g had n_{voxels} units, which was participant-dependent (that is, the number of voxels in an individual's navigation network mask). Hidden layers had 256, 128, 64, 20, 64, 128 and 256 units, respectively, and leaky rectified linear unit activations were applied on all layers. We used the Adam optimizer, a batch size of 64, and a learning rate of 0.001. The model was trained with a reconstruction error penalty for each sample seen at training time, to minimize the MSE between Y and $\hat{Y}$, or $f(g(Y))$. Models were trained for 10,000 epochs; each epoch consisted of a complete pass through the entire training dataset to calculate errors and update network parameters. Thus, given participant's data Y, where Y has k timepoints, encoder f, and decoder g, the reconstruction penalty was defined as:

$$L_{f,g}^{\text{reconstruction}}(Y) = \sum_k |\hat{Y_k} - Y_k|^2 \tag{4}$$

The model was also trained with manifold regularization, which pushed the bottleneck layer to have the same geometry as the initial T-PHATE embedding of Y, E. This was computed by:

$$L_{f,g}^{\text{geometric}}(Y) = \sum_k |f(Y_k) - E_k|^2 \tag{5}$$

To combine $L^{\text{geometric}}$ and $L^{\text{reconstruction}}$, we used a coefficient γ that controlled the amount of geometric regularization to the hidden layer of the MRAE. We used $\gamma = 0.01$ for all participants, following previous work on manifold and geometry regularized autoencoders. The combined loss L thus became:

$$L = \sum_k \left(|Y_k - \hat{Y}_k|^2 + \gamma |f(Y_k) - E_k|^2 \right) \quad (6)$$

which was optimized over encoder f and decoder g. After training, these weights were frozen and the testing set of data from the joystick task (the remaining 20% of data) were passed through f to obtain their embeddings in the latent space.

Defining manifold components. We used manifold components to map the data from the T-PHATE manifold to the avatar's movement. We learned these components using PCA over the MRAE latent space (post-training), which yielded the eigenvectors of the T-PHATE covariance matrix. We selected the three eigenvectors that captured the greatest, second greatest and least variance in the 20-dimensional manifold-embedded brain activity. We validated that this approach of using the T-PHATE-based eigenvectors faithfully reflected the nonlinear latent structure of noisy data (Supplementary Fig. 11). Experimentally, we refer to these eigenvectors as the neurofeedback mapping components, and we take the 1st and 2nd components to be C_{IM} and C_{WMP} and the 20th component to be C_{OMP}. These components were fitted individually on each participant's training data from the joystick session and held constant through the experiment (that is, not refitted on subsequent days).

Data analysis

Joystick session validation. We validated that the T-PHATE embeddings captured task-relevant signals during the joystick session. We visualized the first three dimensions of the T-PHATE embedding timeseries and colored each point based on the avatar's coordinate in the game arena at that timepoint. We depict embeddings from four consecutive trials in five representative participants showing clustering by game arena coordinates, with similarly colored points placed nearby despite being derived from brain activity collected in different trials (Fig. 1b). To assess the benefits of embedding the joystick session data with T-PHATE relative to more typical linear dimensionality reduction methods, we performed the same visualization with PCA (Supplementary Fig. 2a,b).

We next attempted to decode the avatar's coordinates from the manifold embeddings. We trained a linear model (ridge regression, implemented with the himalaya package in Python[79]) to predict the avatar's location (X, Y coordinates, normalized within trial to account for different spawning locations) at each timepoint from the T-PHATE embedding timeseries in the navigation network, using leave-one-run-out cross-validation. On the held-out run, models were scored as the MSE between the model-predicted (X, Y) coordinates and the avatar's true (X, Y) coordinates. Hyperparameter optimization was performed with leave-one-run-out cross-validation in the inner fold to select the best regularization penalty in the ridge regression. To assess the benefit of T-PHATE, we performed the same analysis using the original voxel-resolution data (~1,300 voxels) before running T-PHATE (Fig. 1c) and the voxel-resolution data embedded with PCA (Supplementary Fig. 2c).

We also evaluated the representational similarity of brain activity and arena locations. In other words, are more proximal locations in the game arena represented more similarly in the brain than more distant locations? We quantified this similarity by extracting fMRI activity patterns for each timepoint and calculating the Pearson correlation of the voxel or component patterns from all pairs of timepoints to populate a timepoint-by-timepoint similarity matrix. We generated a parallel timepoint-by-timepoint similarity matrix for distances in the game arena by calculating one minus the Euclidean distance of the avatar's location between all pairs of timepoints.

For each participant, we tested for the second-order representational similarity of the brain's correlation matrix, calculated from voxel-resolution, PCA, and T-PHATE embeddings, and the arena's proximity matrix using a Mantel test[42,80] (Fig. 1d for voxel-resolution and T-PHATE; Supplementary Fig. 2d for PCA). This metric captures how the neural embeddings reflect local and global task structure: subsequent timepoints should be represented more similarly, as the avatar moves continuously through the space, and more temporally distant timepoints may also be represented similarly, as the avatar can return to the same arena coordinates later in time. The Mantel test takes the Spearman correlation of the true brain and arena proximity matrices and a null distribution created by randomly permuting one of the matrices and recomputing the Spearman correlation (10,000 iterations), then reports a z-score and P value of the true correlation relative to the null distribution.

BCI learning during neurofeedback. The staircased BrainControl parameter scaled with task accuracy (that is, stepping up with better performance and down with worse performance), so ΔBrainControl serves as a metric of BCI learning in each session. Across trials, we tested whether ΔBrainControl was significantly greater than zero using randomization testing with cluster-based correction. As a baseline, we computed the ΔBrainControl for each session type using the simulated null participants.

Neural realignment with BCI learning. To quantify the neural changes underlying BCI learning, we computed the proportion of neural variance explained by each component of the manifold. Namely, using the 20-dimensional T-PHATE embedding of each fMRI timepoint (excluding rest between trials) $f(V_t)$, we computed the variance explained by each eigenvector. We divided this value by the overall variance of the data and multiplied by 100 to get the PEV along each component. We focused on the PEV along the IM ($f(V_t) \cdot C_{IM}$), WMP ($f(V_t) \cdot C_{WMP}$) and OMP ($f(V_t) \cdot C_{OMP}$) components. We then took the final run's PEV and subtracted it from the first run's PEV to get ΔPEV, our measure of neural realignment.

In the main analyses (Fig. 4), we report ΔPEV for each component in its corresponding session (that is, ΔPEV for the IM component during the IM session). Supplementary Results report ΔPEV for each component in noncorresponding sessions (that is, ΔPEV for the IM component during the WMP session) to test the specificity of the observed neural learning effects (Supplementary Fig. 7). We bolstered our characterization of the observed neural changes as realignment by considering the overall variance in brain activity over learning and the specificity of the ΔPEV findings to the component being trained versus arbitrary manifold components, as well as simulated changes that could result from a realignment or subselection distributional shifts of different sizes. The details of these analyses are in Supplementary Methods.

Whole-brain decoding. The navigation mask of brain regions used for neurofeedback showed neural realignment, but changes could also be reflected elsewhere in the brain. Using a whole-brain searchlight analysis, we computed the accuracy of decoding the avatar's coordinates in the game arena from multi-voxel patterns of fMRI data during the first and last runs of neurofeedback for each session type. Searchlights were centered on every voxel in the brain, each surrounded by a sphere with a radius of 3 voxels (343 total voxels per searchlight).

For each searchlight, we trained a linear model (ridge regression, himalaya package in Python[79]) to predict the avatar's location (X, Y coordinates) at each timepoint from the fMRI activity pattern across voxels in that searchlight using leave-one-trial-out cross-validation. Models were scored based on the MSE between the predicted and true coordinates in the held-out trial, averaged across folds to get one voxelwise map of errors for the first and last neurofeedback runs of each session type (lower distance means better decoding). To quantify the change in decoding, we subtracted the last run's map from the first run's map and normalized the difference by the sum of the first and last run's maps. The resulting value for each voxel is thus bound between −1 and 1, with 1 indicating a larger relative distance in the first run than the final run (that is, improved location decoding).

We tested the statistical significance of the normalized difference in each session type using FSL's randomise function[81] with variance smoothing and a 5-mm sigma, corrected for multiple comparisons with threshold-free cluster enhancement[82]. Note that these analyses were, by necessity, performed in voxel space, as the searchlights were different sets of voxels than those used to define the T-PHATE manifold.

Statistical analyses

All statistical comparisons were conducted using nonparametric randomization tests (10,000 iterations). For paired comparisons and one-sample tests against zero, condition labels were randomly shuffled within participant to construct a null distribution of the mean (difference) across participants. For independent-sample comparisons, group labels were randomly shuffled across all observations, preserving group sizes, to construct the null distribution of the mean difference across participants. The *P* value was calculated as the proportion of iterations meeting or exceeding the observed statistic in absolute value (or exceeding 0 in magnitude for one-sample tests). One-sided tests were used for directional hypotheses and two-sided tests for nondirectional hypotheses. Significance at the trial level (Fig. 3a) was evaluated using randomization tests, with cluster-based multiple-comparisons correction across adjacent trials.

CIs (95%) were estimated by using bootstrap resampling to generate a sampling distribution of the effect[83]. On each of 10,000 iterations, we sampled the same number of participants with replacement. For one-sample tests, we calculated the mean across resampled participants. For paired comparisons, we computed the difference between conditions within each resampled participant and then the mean difference across participants. For independent-sample comparisons, we resampled participants separately within each group before calculating the group means and difference.

To evaluate individual differences in brain–behavioral relationships, we fit an ordinary least squares regression model using the statsmodels package in Python[84]. We clustered standard errors by participant to account for the repeated-measures structure and included counterbalancing order as a fixed effect. CIs and *P* values were computed using a Wald *t*-distribution approximation.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Raw and preprocessed functional and anatomical images, behavioral data, model weights and neurofeedback masks in native space are available via Dryad at https://doi.org/10.5061/dryad.9cnp5hr0w (ref. 85). Source data are provided with this paper.

Code availability

The task code was programmed with C# for Unity (unity3d.com; Version 2019.4.12f1) and is available at github.com/ericabusch/avatarRT_task. The rt-fMRI experiment used the open-source rtCloud framework (rt-cloud.readthedocs.io)[64]. Our experiment, preprocessing and analysis scripts are available at github.com/ericabusch/avatarRT_analysis. MRAE code is available at github.com/ericabusch/MRAE. T-PHATE is available as a Python package at github.com/KrishnaswamyLab/TPHATE and via Zenodo at https://doi.org/10.5281/zenodo.7637522 (ref. 78).

References

65. Cortese, A., Amano, K., Koizumi, A., Kawato, M. & Lau, H. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat. Commun.* **7**, 13669 (2016).
66. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665–670 (2011).
67. Rodriguez, P. F. Neural decoding of goal locations in spatial navigation in humans with fMRI. *Hum. Brain. Mapp.* **31**, 391–397 (2010).
68. Kamps, F. S., Lall, V. & Dilks, D. D. The occipital place area represents first-person perspective motion information through scenes. *Cortex* **83**, 17–26 (2016).
69. Mueller, C. et al. Building virtual reality fMRI paradigms: a framework for presenting immersive virtual environments. *J. Neurosci. Methods* **209**, 290–298 (2012).
70. Sherrill, K. R. et al. Hippocampus and retrosplenial cortex combine path integration signals for successful navigation. *J. Neurosci.* **33**, 19304–19313 (2013).
71. Woolrich, M. W., Ripley, B. D., Brady, M. & Smith, S. M. Temporal autocorrelation in univariate linear modeling of fMRI data. *Neuroimage* **14**, 1370–1386 (2001).
72. Smith, S. M. Fast robust automated brain extraction. *Hum. Brain. Mapp.* **17**, 143–155 (2002).
73. Andersson, J. L. R., Skare, S. & Ashburner, J. How to correct susceptibility distortions in spin-echo echo-planar images: application to diffusion tensor imaging. *Neuroimage* **20**, 870–888 (2003).
74. Jenkinson, M., Bannister, P., Brady, M. & Smith, S. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* **17**, 825–841 (2002).
75. Greve, D. N. & Fischl, B. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage* **48**, 63–72 (2009).
76. Gao, S., Mishne, G. & Scheinost, D. Nonlinear manifold learning in functional magnetic resonance imaging uncovers a low-dimensional space of brain dynamics. *Hum. Brain. Mapp.* **42**, 4510–4524 (2021).
77. De, A. & Chaudhuri, R. Common population codes produce extremely nonlinear neural manifolds. *Proc. Natl Acad. Sci. USA* **120**, e2305853120 (2023).
78. Busch, E. KrishnaswamyLab/TPHATE: Version 1.1. *Zenodo* https://doi.org/10.5281/zenodo.7637522 (2025).
79. Nunez-Elizalde, A. O., Huth, A. G. & Gallant, J. L. Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage* **197**, 482–492 (2019).
80. Carr, J. Jwcarr/mantel. *GitHub* https://github.com/jwcarr/mantel.git (2022).
81. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain. Mapp.* **15**, 1–25 (2002).
82. Smith, S. M. & Nichols, T. E. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* **44**, 83–98 (2009).
83. Stine, R. An introduction to bootstrap methods: examples and ideas. *Sociol. Methods Res.* **18**, 243–291 (1989).
84. Seabold, S. & Perktold, J. statsmodels: econometric and statistical modeling with python. In *9th Proc. Python Science Conference* 92–96 (SciPy, 2010).
85. Busch, E. et al. Data and code from: Human learning of noninvasive brain–computer interfaces via manifold geometry [Dataset]. *Dryad* https://doi.org/10.5061/dryad.9cnp5hr0w (2026).

Acknowledgements

E.L.B. was supported by an NSF Graduate Research Fellowship (award no. 2139841). G.L. was supported by Canada CIFAR AI Chair and Canada Research Chair in Neural Computations and Interfacing.

S.K. was supported by the NIH (grant nos. R01GM135929 and R01GM130847), an NSF Career Grant (grant no. 2047856) and a Sloan Fellowship (grant no. FG-2021-15883). N.B.T.-B. was supported by the NIH (grant no. R01MH069456), NSF (grant no. 1839308) and CIFAR. Additional internal funding was provided by the Faculty of Arts and Sciences and the Wu Tsai Institute at Yale University. We thank L. Behm, T. Botch, M. Conley, G. Feng, J. Heffner, C.-H. Kao, A. Letrou, K. Peng, J. Trach, T. Yates, X. Zhang, I. Zhou and Y. Zhu for help with data collection.

Author contributions

E.L.B., S.K. and N.B.T.-B. conceived the study. E.L.B., G.L., S.K. and N.B.T.-B. designed the experiment. E.L.B. and E.C.F. conducted the neurofeedback experiment and the simulation experiments. E.L.B. analyzed the data with support from E.C.F. and feedback from all authors. E.L.B., S.K. and N.B.T.-B. contributed to the development of data processing algorithms. E.L.B. and N.B.T.-B. wrote the manuscript with extensive feedback from all authors.

Competing interests

S.K. is a cofounder and CSO of Latent Alpha. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41593-026-02311-2.

Correspondence and requests for materials should be addressed to Smita Krishnaswamy or Nicholas B. Turk-Browne.

Peer review information *Nature Neuroscience* thanks Aaron Batista, Stephen LaConte and Takeo Watanabe for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Smita Krishnaswamy, Nick Turk-Browne

Last updated by author(s): Mar 31, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data were collected on a 3T Siemens Prisma scanner with custom code written in Psychopy (version 2022.2.4 ; scanner experiment code available here https://github.com/ericabusch/avatarRT_task/tree/main/experimenter_computer). Real-time fMRI used the open-source rtCloud framework (rt-cloud.readthedocs.io). The task performed in the scanner was run with custom code programmed in C# for Unity3D and presented with Unity (Version 2019.4.12f1; task code available: https://github.com/ericabusch/avatarRT_task/tree/main/Assets/Scripts).
Data analysis	Data analyses was performed with custom scripts in python version 3.7. All analysis code is available here: https://github.com/ericabusch/avatarRT_analysis along with conda environments containing versions for python packages including SciPy, Sklearn, T-PHATE, MRAE, pytorch.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Raw and preprocessed functional and anatomical images, behavioral data, model weights, and neurofeedback masks in native space are available at the following link: https://doi.org/10.5061/dryad.9cnp5hr0w. Source data are available for the results presented in all figures.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	We report data from 20 participants, including 10 identifying as female.
Reporting on race, ethnicity, or other socially relevant groupings	We did not collect or report information about race or ethnicity or other socially relevant groupings.
Population characteristics	We recruited 20 healthy young adults from the New Haven community (10 female, age range = 19--35 years, mean age = 25.8 years).
Recruitment	Participants were recruited from the New Haven community via community recruitment and word of mouth (i.e., fliers, email lists). . All participants provided informed consent to an experimental protocol approved by the Yale University Institutional Review Board. Participants completed four or five fMRI sessions lasting 1–1.5 hours each and were compensated for their participation ($20/hour plus retention incentives: $5/session cumulative bonus for each return session and $40 for completion).
Ethics oversight	The study protocol was approved by the Yale University Institutional Review Board

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Data are quantitative experimental, with a repeated within-subject, repeated-measures design.
Research sample	The research sample included healthy young adults (ages 19-35 years) from the New Haven community. Ten participants reported self-identifying as female. This sample size is greater than or comparable with other multi-day real-time fMRI studies.
Sampling strategy	The sample size was chosen based upon prior literature conducting multi-day real-time fMRI studies. Our sample size is comparable with or greater than those included in prior literature.
Data collection	Participants were blind to the hypothesis and experimental condition during the course of the experiment. Researchers were not blind as to condition. fMRI data were collected using a 3T Siemens Prisma MRI scanner. No one was present during the experiment aside from the participant and two researchers who were operating the scanner and experiment. Data about the task were collected using custom Unity3D scripts and provided as above.
Timing	Data for all 20 participants were collected between April 5, 2023, and September 15, 2023.
Data exclusions	Two participants were excluded from neurofeedback analyses (i.e., included for the joystick results during the first session, as presented in Figure 1, but excluded from analyses for second session onward). One participant was excluded due to a scanner malfunction during Session 4 resulting in lost data for that session, preventing the within-subject comparison. The other participant was excluded for falling asleep during multiple fMRI runs.

Non-participation	One participant dropped out during the second session due to discomfort during the MRI scan.
Randomization	Participants were randomly assigned to either receive the within-manifold perturbation or outside-manifold perturbation first, such that 10 participants were included in each condition.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Plants

Seed stocks	n/a
Novel plant genotypes	n/a
Authentication	n/a

Magnetic resonance imaging

Experimental design

Design type	Naturalistic video game task design
Design specifications	Participants completed 4 runs of approximately 10 minutes per run during each MRI session. Runs were comprised of trials that were self paced but on average 30 s per trial, with 6 s break between trials. Participants were given a break of self-determined length between runs. Participants completed 4--5 fMRI sessions total on separate days, but with a target of 24 hours between session (maximum time = 48 hours) (mean = 27.4 hours, range = 13--48 hours). All sessions had to be completed within 7 days.
Behavioral performance measures	Participants were scored based on their accuracy during the video game task (i.e., the distance they traveled in navigating the avatar from the start location to the goal location relative to the shortest possible distance). Participants were presented with this as the percentage of extra distance the avatar traveled during a given run. This was used to adjust the difficulty of the task, where the more difficult the task got, the better participants were performing. We compared the mean performance and standard deviation of performance within condition to determine if participants were performing the task as expected, and we also compared this to simulated fMRI data performing the same task

Acquisition

Imaging type(s)	functional
Field strength	3T
Sequence & imaging parameters	Functional data were acquired using a 3T Siemens Prisma scanner with a 64-channel head coil at the Brain Imaging Center at Yale University. BOLD data were collected with an echo-planar imaging (EPI) sequence (TR = 2 s; TE = 30 ms; voxel size = 3 mm isotropic; FA = 8 degrees; IPAT GRAPPA acceleration factor = 2; distance factor = 25%). We acquired two spin-echo field map volumes in opposite phase encoding directions for field map correction (TR = 5 s; TE = 80 ms; otherwise matching the EPI sequences). During the first session we collected two high-resolution anatomical images for

spatial registration: a 3D T1-weighted magnetization-prepared rapid acquisition gradient echo (MPRAGE) scan with an IPAT GRAPPA acceleration factor of 2 (TR = 2.5 s; TE = 2.9 ms; voxel size = 1 mm isotropic; FA = 8 degrees; 176 sagittal slices), and a 3D T2-weighted fast spin echo scan with variable flip angle and IPAT GRAPPA acceleration factor of 2 (TR = 3.2 s; TE = 565 ms; voxel size = 1 mm isotropic; 176 sagittal slices).

Area of acquisition: whole-brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software: Data collected during the first fMRI session were preprocessed using fMRI Expert Analysis Tool (FEAT) version 6.00, part of fMRIB's Software Library (FSL) version 6.0.5. EPI and anatomical images were skull-stripped using the Brain Extraction Tool (BET). Susceptibility-induced distortions were measured via opposing phase spin echo volumes and corrected using FSL's topup command. EPI images were high-pass filtered with a 100 s period cutoff, corrected for head motion using MCFLIRT, corrected for slicetiming, and smoothed spatially using a Gaussian kernel (5 mm full-width half-maximum).
During real-time analysis, these steps were abbreviated to be conducted rapidly: the first volume of each functional run was smoothed with a 5 mm FWHM Gaussian kernel and FMRIB's Linear Image Registration Tool (FLIRT) with nearest neighbor interpolation was used to derive a transformation T between a template volume from the participant's first session data and this first volume of the functional run. All subsequent volumes were smoothed with the 5 mm FWHM Gaussian kernel, motion corrected using MCFLIRT, and aligned to the template volume via T.

Normalization: Functional images were registered to a participant's T1 weighted anatomical scan using boundary-based registration (BBR) with 3 degrees of freedom and to standard space using a nonlinear registration 12 degrees of freedom.

Normalization template: Data were registered to the 2mm MNI standard brain for deriving a transformation matrix for the functional ROI selected and for performing the whole-brain searchlight analysis. Native space was used for computing neurofeedback scores in real-time, and data were normalized in real-time to a reference volume collected from the participant during their first session.

Noise and artifact removal: Head motion was corrected for using Motion Correction using FLIRT (MCFLIRT) in FSL version 6.0.5. Due to the temporal constraints on real-time fMRI we did not perform artifact removal for the first session analyses and for real-time analyses.

Volume censoring: Due to the temporal constraints on real-time fMRI we did not perform volume censoring.

Statistical modeling & inference

Model type and settings: In Figure 1C, we report a multivariate decoding analysis using a multiple linear regression to decode task features (i.e., the avatar's location in the arena) from 1) the high-dimensional voxel-resolution pattern and 2) the T-PHATE embedding of this pattern. We report the mean-squared error of these models. In Figure 1D, we use the same input data and use a representational similarity analysis to relate the similarity of task features (the avatar's location in the arena) and the similarity of neural activity. In figure 4B, we report the change in the percentage of variance explained by the neurofeedback component over the course of neurofeedback learning. In Figure 4C, we fit an ordinary least squares regression model using the statsmodels package in Python. We clustered standard errors by participant to account for the repeated-measures structure and included counterbalancing order as a fixed effect. We used a whole-brain searchlight analysis in figure 4D to decode task features from multi-voxel patterns.

Effect(s) tested: We used ANOVA to test whether participants differed in their ability to learn the BCI as an effect of an interaction between session type and session order.

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☒ Both

Anatomical location(s): ROI-based analyses were used for the joystick and neurofeedback sessions. The region of interest was determined using the search term "navigation" in the meta-analysis site Neurosynth.org . This yielded an association test z-statistic map, which displays voxels frequently reported in articles including the search term and FDR corrected at 0.01 . We also perform a whole brain searchlight analysis.

Statistic type for inference

(See Eklund et al. 2016)

All statistical comparisons were conducted using nonparametric randomization tests (10,000 iterations). For paired comparisons and one-sample tests against zero, condition labels were randomly shuffled within participant to construct a null distribution of the mean (difference) across participants. For independent-sample comparisons, group labels were randomly shuffled across all observations, preserving group sizes, to construct the null distribution of the mean difference across participants. The P-value was calculated as the proportion of iterations meeting or exceeding the observed statistic in absolute value (or exceeding 0 in magnitude for one-sample tests). One-sided tests were used for directional hypotheses and two-sided tests for nondirectional hypotheses. Significance at the trial level (Figure 3A) was evaluated using randomization tests, with cluster-based multiple-comparisons correction across adjacent trials.
Confidence intervals (95%) were estimated by using bootstrap resampling to generate a sampling distribution of the effect. On each of 10,000 iterations, we sampled the same number of participants with replacement. For one-sample tests, we calculated the mean across resampled participants. For paired comparisons, we computed the difference between conditions within each resampled participant and then the mean difference across participants. For independent-sample comparisons, we resampled participants separately within each group prior to calculating the group means and difference83.
To evaluate individual differences in brain-behavioral relationships, we fit an ordinary least squares regression model using the statsmodels package in Python. We clustered standard errors by participant to account for the repeated-measures structure and included counterbalancing order as a fixed effect. Confidence intervals and P-values were computed using a Wald t-distribution approximation.

Correction

Cluster-based multiple comparison correction across adjacent trials (Figure 3A).
Threshold-free cluster enhancement multiple comparisons correction in the whole-brain searchlight analysis

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

Dimensionality reduction used the T-PHATE algorithm as outlined in the methods section.
Manifold regularized autoencoder (MRAE) was used to extend the T-PHATE manifold. This is an autoencoder which has three fully connected layers in the encoder and three in the decoder with a bottleneck latent space layer between them. This model is trained using an Adam optimizer, batch size of 64, learning rate of 0.001, with a reconstruction error penalty to minimize the mean squared error between input and output samples and a manifold regularization penalty to minimize the mean squared error between the sample's embedding in the bottleneck layer and its ground-truth location in the TPHATE embedding space.