

Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity

Thorsten O. Zander, Laurens R. Krol, Niels P. Birbaumer, Klaus Gramann

Table of Contents

Materials and Methods	2
1. Experimental Set-Up	2
2. Experimental Paradigm	2
3. Participants	4
4. Experimental Procedure	4
7. Performance Measures and Statistics	8
8. Linear Dependency of Peak Amplitude.....	9
Figures	10
Fig. S1. Experimental Paradigm and Procedure.....	10
Fig. S2. Angular Deviance	11
Fig. S3. Class Definition of the Classifier's Training Set	12
Fig. S4. Directional Probabilities	13
Fig. S5. LDA-projected ERPs	14
Fig. S6. Influence of Eye Movements on the ERP at Fz	17
Fig. S7. Mean Peak Amplitudes at Fz	18
Fig. S8. Scalp Activity Series.....	19
Fig. S9. Example Online Cursor Behavior.....	20
Fig. S10. The neuroadaptive control loop	21
Tables	22
Table S1. Online Classification Accuracy.....	22
Table S2. Post Hoc: Final Direction Probabilities.....	23
Table S3. Post Hoc: Peak Amplitude Differences at Fz.....	24
Table S4. Post Hoc: Classifier Output.....	25
Table S5. Pairwise Comparisons: Projected ERP Peak Amplitude Differences	26
References	26

Materials and Methods

1. Experimental Set-Up

Two computers were used for this experiment: One to display the stimuli and another to record EEG. Stimuli were shown on an Iiyama ProLite B2776HDS 27'' display using a resolution of 1920×1080 pixels, located approximately one meter away from the participants. The keyboard used was a standard US layout computer keyboard.

EEG was recorded continuously using 64 active Ag/AgCl electrodes (actiCAP; Brain Products GmbH, Munich) mounted according to the extended 10-20 system on an elastic cap (EASYCAP GmbH, Munich). The signal was sampled at 500 Hz and amplified using BrainAmp amplifiers (Brain Products GmbH, Munich). All electrodes were referenced to FCz.

EEG data was recorded and combined with the paradigm's markers using the Lab Streaming Layer framework (Swartz Center for Computational Neuroscience [SCCN], University of California, San Diego [UCSD]).

2. Experimental Paradigm

The cursor's movements were restricted by the nodes of a visible grid. The cursor started out on one of these nodes and, for each movement, could only move to one of the adjacent nodes. Depending on the cursor's position in the grid, then, there were up to eight possibilities for one movement. The grid consisted of grey open circles connected by grey lines on a black background. The cursor in this paradigm was a red, filled circle. The cursor's size was 4.5% of the display's total height, the grid nodes' size 6%, and the horizontal and vertical grid lines' length 15%. Line thickness between the grid nodes was set to 2 pixels. 'Red' was pure RGB(255,0,0), 'grey' RGB(51,51,51) and 'black' RGB(0,0,0). The dimensions of the grid are discussed below.

An animation allowed the participants to be able to anticipate the moment of each movement. Over the course of one second, a white 'ghost cursor' would grow inside of the actual cursor. As soon as this ghost reached the same size as the actual cursor, it would instantaneously be repositioned to the chosen adjacent node, while also highlighting the grid line connecting the two nodes in white. 'White' here was pure RGB(255,255,255), and the highlighted grid line's thickness was increased to 3 pixels. The movement remained visible for one second, with the red original cursor still on the initial node, the white ghost cursor on the new node, and a white line connecting them. Following that, all whites disappeared and the (red) cursor would instantaneously move to and remain at its new position, on the new node, for another second, before the animation would start over for the next movement. In all, a single trial was three seconds in length.

Grids of three different dimensions were used in this study: One by three nodes 1×3), four by four nodes (4×4), and six by six nodes (6×6).

On each grid, a single red node in one of the corners indicated the current target. Once the cursor had landed on this target node, a new grid was started with the same dimensions, but a different layout. For the 1×3 grids, each new grid was rotated either 45, 90, or 135 degrees relative to its predecessor, and a new target corner was chosen at random ('corner' here is one of the two ending nodes of the grid row). The larger grids did not rotate, but a new target was selected for each new grid such that no two subsequent grids had a target in the same corner. Each newly started grid was displayed for one second before the first movement was initiated, for the participants to orient themselves.

In all grids, the cursor's starting position was one node away from the corner opposite the target's, in a straight line to the target.

Supplementary Figure S1 illustrates the stimuli; Supplementary Movie 3 contains animated stimuli as seen by the participant.

For each individual cursor movement, a movement direction was chosen randomly from the set of possible movements. By default, all possible directions had an equal probability of being selected, and during the offline (calibration) sessions, these initial probabilities remained unchanged---the cursor was unbiased and moved randomly. However, during online application of the pBCI, the directional probabilities were altered based on the classification of each movement as either 'correct' or 'incorrect', biasing the cursor to repeat movements classified as 'correct'.

Specifically, all movement directions were represented a specific number of times in the set from which each new movement was randomly selected. After classification, the respective directions' numbers were increased or decreased such that the resulting probability of each subsequent direction is given by

$$p = \frac{mx}{n + (m - 1)x}$$

where x is the respective direction's initial number of shares in the set, n is the full set's initial size, and m is a multiplier. For a movement classified as 'correct', m was 2.0 for that movement's direction and 1.5 for the two adjacent directions; for movements classified as 'incorrect', m was 0.5 and 0.75, respectively. All directions started with a share of 100 elements each in the set. All shares below 1 were rounded up at the time of selection.

The adaptation of adjacent directions represents a certain degree of leniency which the participants were assumed to show in their judgements: If the target is a few nodes north of the cursor's current position, a first move to the north-west is imperfect but may still be acceptable. Therefore, the feedback received from a movement to the north-west should not only alter this specific direction's probability, but should also influence subsequent movements to both the north and the west.

All visible on-screen events were marked in the EEG stream and additional information about each cursor movement was logged, including distance to the target and the movement's direction

relative to the target direction (i.e., a straight line unconstrained by the grid represented a 0° movement).

The paradigm was written in Python using the Simulation and Neuroscience Application Platform 1.02-beta (SCCN, UCSD; <https://github.com/sccn/SNAP>).

3. Participants

A total of nineteen participants participated in this study, with an average age of 25.4 years \pm 3.4. Seven were female, and a total of two participants were left-handed. Participants were asked, alternately, to use either their left or their right hand to indicate their judgements. This resulted in ten participants not using their dominant hand. All had normal or corrected to normal vision. Instructions were given in writing, in German. While all knew German, thirteen did not have German as their mother tongue; these were verbally given additional, standardized instructions.

Nineteen participants performed calibration trials, while sixteen of them additionally performed online trials.

4. Experimental Procedure

All participants were informed of the nature of the experiment and the recording and anonymization procedures before signing a consent form. The ethics committee of the Department of Psychology and Ergonomics at the Technische Universität Berlin approved the experiment and the procedures.

After preparation and setting up the EEG cap, which took roughly one hour per participant, participants were seated in a padded chair in a dimly lit room, about one meter away from the stimulus display. In writing, they were instructed to judge each individual cursor movement on the display as either ‘acceptable’ or ‘not acceptable’, with respect to reaching the target, and to indicate their judgement by pressing either ‘v’ or ‘b’, respectively, on a computer keyboard using one and the same finger of one hand. The participants performed this task during all blocks.

Participants were first given four blocks of 50 trials on 1×3 grids. Here, the cursor performed only one movement per grid: Regardless of whether or not that movement reached the target, a new grid was subsequently started. This was because after a movement away from the target, only one movement possibility would remain, i.e. that trial would have no informational value. Breaks between these blocks were self-paced, and participants were given time to practice before the first block was started.

Following these first blocks, participants received additional instructions for the larger grids, emphasizing their task to judge every movement individually and independently of the cursor’s movement history. Participants were again given some time to practice, on a 4×4 grid, and then performed five blocks of 120 trials on grids of that size. Here, if the target had not been reached after 55 trials in one grid, a new grid was started. 55 is twice the median number of random movements required to reach a target on a 4×4 grid.

Taking symmetry and rotation into account, a total of 43 unique cursor movements are possible on the 4×4 grid, i.e., 43 unique pairs of distance and angular deviance. In between the five

calibration blocks, in four sessions, participants were shown, in random order, all 43 unique cursor movements. For each of these, participants were asked to rate it as either ‘acceptable’ or ‘not acceptable’, and on a scale of 1 to 5, with 5 being ‘the best possible movement in that situation’, and 1 being ‘the worst’. Here, unlike during the blocks, there was no time limit imposed on their answers.

EEG recorded during these latter five blocks served to calibrate the classifier, as discussed in Supplementary Method 5. This classifier was then applied to one more block of 120 trials on 4×4 grids, and one last block of 120 trials on 6×6 grids. No maximum number of trials other than the block’s length was set for these online blocks.

5. Feature Extraction and Classification

A BCI based on supervised machine learning needs to be calibrated before it can be applied. This calibration is typically performed on sets of recordings, usually EEG epochs, which are known to contain the signals that need to be detected later. On the basis of these epochs, a classifier is calibrated to optimally distinguish between the different classes of source signals.

The classes the classifier was calibrated on were formed on the basis of the movements’ angular deviance from a straight line towards the target, as illustrated in Supplementary Figure S2. ‘Correct’ movements were those with an absolute angular deviance of 0°; ‘incorrect’ were those with an absolute angular deviance of 135° or more, as in Supplementary Figure S3. These class definitions were determined based on the first three participants’ data.

Note that the participants’ judgements, indicated using button-presses, were ignored: Only a movement’s angle with respect to the target determined its class. Movements between 0 and 135° were not included for calibration.

The open-source toolbox BCILAB (23) version 1.01 was used to define and implement the pBCI. Features were extracted through the windowed means approach. This approach extracted features from the time course of each of the 64 channels by subsampling the ERPs of each epoch. This was done by dividing the time course into a sequence of 8 consecutive 50 ms windows (8 time windows starting at 50 ms after cursor movement) and replacing each window by its average. The resulting features were thus one value for each channel in each time window. Linear discriminant analysis (LDA) was then applied to these features generated from all available calibration trials to distinguish between the features belonging to the two classes. The outcome of the LDA is a linear weighting of all features. In other words, in each time window all channels receive a weight according to their relevance for classification in that specific time window. The linear combination of all features of one single trial with these weights then gives a number between -1 and 1, indicating whether this trial is classified as belonging to class 1 or to class 2.

For the feature extraction, the data was first resampled at 100 Hz, and band-pass filtered from 0.1 to 15 Hz. Linear discriminant analysis was regularized by shrinkage. A [5,5]-times nested cross-validation with margins of 5, ensuring the independence and identity of the feature distributions, was used to select the shrinkage regularization parameter, and to generate estimates of the model’s online reliability.

6. Extended Feature Analysis

The method used by the pBCI to extract discriminative information (described in Supplementary Method 5, above) can be analyzed to reveal further insights into the relevant underlying processes. The classification model used here is a multivariate approach, a linear discriminant analysis (LDA), optimized for the discriminability of the extracted features between classes. Each feature represents data at a single sensor for one of the chosen time windows. Hence, the methodology recently introduced by Haufe *et al.* (1) could be applied to interpret the classification model.

For one of the eight 50 ms time windows, Figure 1 in the main manuscript illustrates three aspects of the information used by the pBCI: its scalp topography, its estimated source within the brain, and the activity of that source as filtered from the EEG, i.e. the selected signal of interest.

Signal of interest. Panel (C) of Figure 1 in the main manuscript shows the EEG activity projected by the LDA filter for the third time window, i.e., the combined activity of all 64 channels as filtered by the classifier. Supplementary Figure S5 illustrates this for all eight time windows.

In Supplementary Figure S5, the ERP is shown for eight groups of cursor movements, time-locked to that movement, as well as the difference wave between the two classes ('correct' and 'incorrect'). The difference wave represents the actual signal (and the optimization target) used by the classifier, in the corresponding time window. The time window of the LDA filter used is indicated in grey.

Each figure indicates the response to different cursor movements of those processes whose activity was filtered from the EEG in that time window, over the course of a full second.

Identifying scalp projections. In panel (A) of Figure 1 in the main manuscript, the scalp map shows, on average for all participants, the (interpolated) activation pattern which illustrates how the signal of interest is expressed in the 64 channels (I), in the third time window. Supplementary Movie 1 shows this activation pattern over the full length of the used 400 ms.

For each participant, LDA patterns $A = (a_j)_j$ were generated from the LDA filters $M = (m_j)_j$ originally used for online classification by conjugation with the features' covariance matrix C : $A = CM C^{-1}$. Spatial interpretation of these patterns for each time window reflects a mixture of scalp activations related to discriminative source activity $\hat{A} = (\hat{a}_j)_j$ and class-invariant noise representation N , with $A = \hat{A} + N$. The latter was filtered out by weighting each pattern entry a_j with the correlation of its associated feature activity vector over trials F_j to the binary vector of true class labels L :

$$\hat{a}_j = \text{corr}(F_j, L) \cdot a_j$$

The resulting correlated pattern $\hat{A} = (\hat{a}_j)_j$ can be visualized by topographic plots for each time window, as in Figure 1 of the main manuscript, Supplementary Figure S5, and Supplementary Movie 1.

Identifying relevant sources. To identify the sources underlying these topographic representations, the backward model, i.e. the LDA filter, was combined with an independent component analysis (ICA). The ICA unmixing matrix $W = (I_1, I_2, \dots, I_n)$ was determined on previously manually cleaned data for each participant by using the Adaptive Mixture ICA (AMICA) Toolbox (2), such that $s = Wx$, where s represents the source activation related to a given scalp activation x . For each time window, the relevance for classification R_i of each independent component I_i can then be determined by distributing the LDA filter weights to the independent components via W , weighted by two factors. The first factor is compensating for the amplitude alignment of the LDA filter weights to the feature amplitudes. It is determined by calculating the variance over trials of the feature \hat{F}_i extracted from the time series of the independent component: $V_i = \text{var}(\hat{F}_i)$. A second weight is determined for filtering out noise representations by weighting the independent components with the correlation of their feature activity to the true class labels (as described above for electrode activity).

$$R_i = V_i * \text{corr}(\hat{F}_i, L) * WM$$

Localizing relevant sources. To localize the identified sources, equivalent dipole models that describe the most likely position of the source in a standard head model were identified for selected components by using the EEGLAB plug-in DIPFIT 2.x (3). Components were selected by a threshold criterion for residual variance of the dipole model ($RV < 0.15$) and visual inspection of activation spectra, time courses, and scalp topographies. Only components reflecting cortical, ocular, or muscular activity were included in the analysis.

For each time window, each of the 371 resulting dipoles was weighted by the relevance R_i of its associated independent component, described above. The areas of high relevance were then described by a weighted dipole density plot using the EEGLAB plug-in dipoleDensity (4).

Supplementary Movie 2 shows the result of this analysis over the full length of the used 400 ms, by plotting the dipole density per cubic millimeter weighted by the relevance R_i of each included dipole with a smoothing kernel of 12 mm.

The above-mentioned process of selection did not markedly influence this analysis. Compared to all 1191 dipoles and averaged over the eight time windows, the 820 rejected dipoles (68.85%) carried 7.5% of the weights distributed by the classifier. Relative to all 1191 dipoles, a total of 87 dipoles received a relevance weight larger than 1 standard deviation in at least one of the time

windows. These 7.3% of dipoles carried 77.82% of all total weight distributed by the classifier. Four of these highly weighted dipoles (4.6%) were rejected in the process explained above and not included in the analysis. These four represent 1.67% of the weight included in the analysis. Three belonged to the same subject.

Behavioral analysis. Participants indicated their evaluation of each movement as ‘acceptable’ or ‘not acceptable’ through button presses. A paired-samples t-test was conducted to compare the subjects’ mean reaction times. There was a significant difference between ‘acceptable’ button presses (mean 369.7 ms \pm 66.9) and ‘not acceptable’ button presses (mean 310.7 ms \pm 60.7); $t(18) = 8.347$, $p < 0.001$. This calculation included only reaction times between 100 and 1000 ms.

No systematic effect reflecting this delay can be found in the ERPs used for classification however (see Figure 1, and Supplementary Figure S5). This, in combination with an absence of relevance weights in the motor cortex (Supplementary Figure S5) and the ERPs' independence of eye movements (Supplementary Figure S6), leads us to conclude that the behavioral delay has had no effect on the passive BCI.

7. Performance Measures and Statistics

The performance of the cursor was operationalized as the number of movements required to reach the target. Trials at the end of a block that did not contribute to reaching a target or hitting the maximum number of trials for that attempt, were discarded. This measure does include those attempts that hit the maximum number of trials, which happened a total of seven times to seven participants on the 4 \times 4 grid, and to two participants on the 6 \times 6 grid. This resulted in 88 pBCI-supported data points for the 4 \times 4 grid, and 47 for the 6 \times 6 grid.

As a comparison, a sample of around 7000 data points was taken using the same measure from non-supported (random) blocks. This data is non-parametric and varies greatly. Therefore, a resampling approach was taken where the available sample of pBCI-supported measures was repeatedly compared to a new random sample of the same size of non-supported performance measures. This comparison was done through a Wilcoxon rank-sum test. Out of 50 000 such comparisons, 98.02% of tests were significant at the $\alpha=0.025$ level for the 4 \times 4 grid performance measures; for the 6 \times 6 grid, 100.00% of tests were significant at this level. On a 4 \times 4 grid, non-supported cursor movement required on average (median) 27 trials to reach a target, whereas with BCI support, this decreased to an average of 13. On a 6 \times 6 grid, the median values were 90 and 23, respectively.

The pBCI-supported performance was furthermore compared to ‘perfectly supported’ performance, where the cursor was reinforced as during the online sessions, except automatically, with perfect accuracy. Here, all movements with an angular deviance less than 45° were judged to be ‘correct’ and reinforced accordingly, and all others were ‘incorrect’, and their probabilities decreased as described in Supplementary Method 2. The same procedure as above yielded significant differences to the pBCI-supported measures for both the 4 \times 4 grid (99.98% of tests at $\alpha=0.025$) and the 6 \times 6 grid (100%). The median perfectly supported performance measure on the 4 \times 4 grid is 10 movements, and 14 movements on the 6 \times 6 grid.

Additionally, the mean directional probabilities upon reaching the target in both online grids (combined) were calculated. These are illustrated in Supplementary Figure S4. Supplementary Table S2 lists the Bonferroni-adjusted results of pairwise post hoc tests from a one-way ANOVA

($F(7,105)=57.520, p<0.001$) on this data. On average, the classifier has been able to reinforce the ‘correct’ directions significantly more strongly than ‘incorrect’ directions.

8. Linear Dependency of Peak Amplitude

Figures 1 and 2 of the main manuscript show a linear scaling of peak amplitudes with respect to absolute angular deviance away from the target. The eight groups of angular deviations were selected such that for each single participant, at least 50 trials were present in each group, i.e., for an optimal signal to noise ratio while maintaining maximal angular resolution.

For the derivation of the single electrode ERP, BrainVision Analyzer was used (Version 2.0.2.5859, Brain Products GmbH, Munich, 2012). The raw EEG data was first band-pass filtered from 0.5 to 45 Hz, and decomposed into statistically maximally independent source signals through ICA. Individual components that resembled eye movements and eye blinks were manually selected for removal based on their time course and topography (as e.g. described in 26). Figure 1 shows the grand average ERP over all nineteen participants at Fz with these components removed. For a comparison, Supplementary Figure S6 shows the grand average ERP at Fp1, Fp2, and Fz for only the removed components. No systematic response of the sort seen in Figures 1 or 2 is visible, pointing to cortical causes of the differences. At Fz, where the cortical components peak around +3.9 μV for e.g. the 0° condition, the eye components show an amplitude of -0.37 μV .

Statistical analysis focused on the systematic peak differences seen at Fz around 180 ms. A one-way ANOVA indicated a significant influence of angular deviance on peak amplitude ($F(7,126) = 47.243, p < 0.001$). Post-hoc comparisons corrected for false discovery rate are listed in Supplementary Table S3 and illustrated in Supplementary Figure S7. The peak amplitudes differ significantly ($p < 0.001$) between the classes used by the classifier. In between, the peak amplitudes scale linearly with angular deviance, as fitted by a linear regression model using each group’s mean angular deviance as predictor ($b = -0.0035, F = 45.28, p < 0.001; R^2 = 0.33$). Classifier output followed a similar trend, as in Figure 1 of the manuscript.

Figures

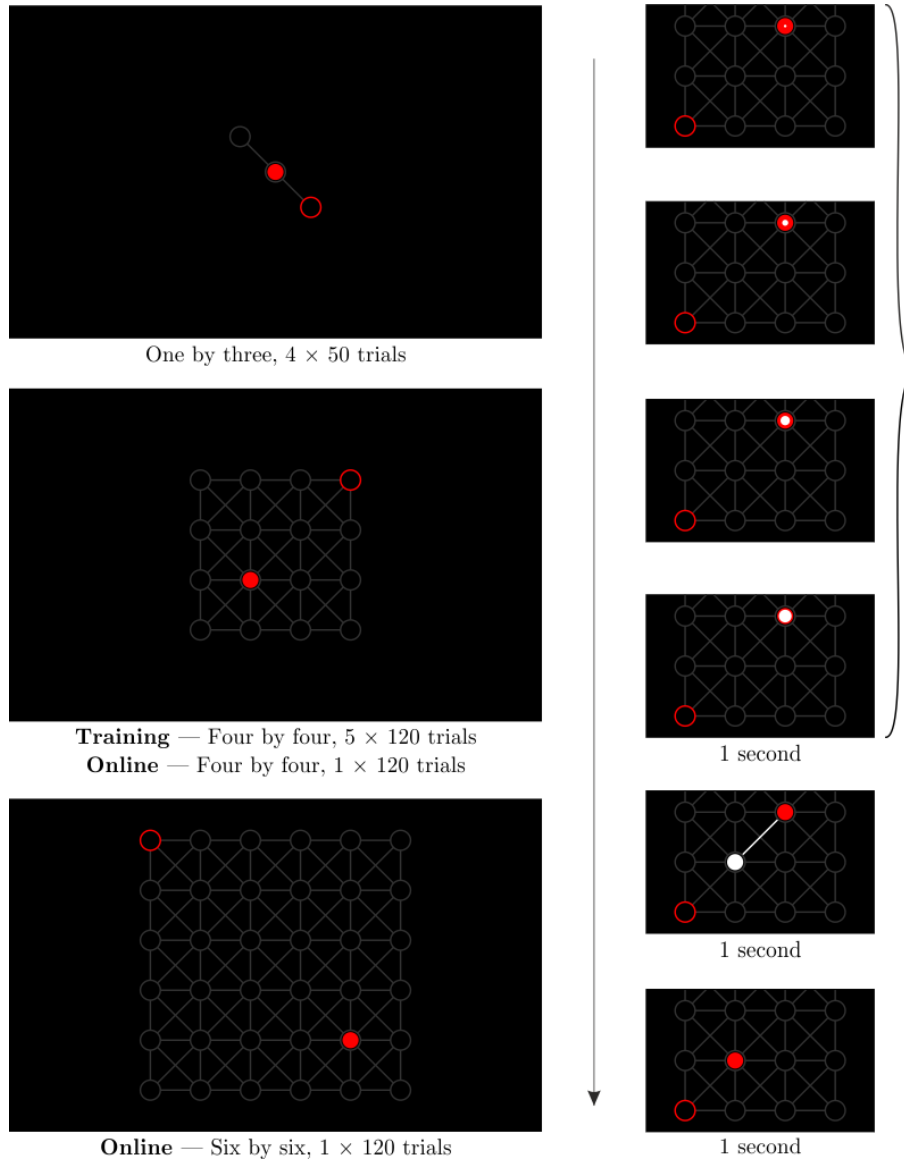


Fig. S1. Experimental Paradigm and Procedure

The experimental paradigm and procedure. *Left*: Three different grid sizes were used in the experiment. Data from five blocks in 4×4 grids was used to calibrate the classifier, which was then applied to another 4×4 block, and one block on a 6×6 grid. Data from the 1×3 grids has not been used in this paper. Each cursor movement consisted of the cursor moving from one node to one of the directly adjacent ones. The cursor could move horizontally, vertically, and diagonally over the grid. Since the target was known and indicated, for each movement, it was possible to determine a measure of correctness for each movement by means of calculating the angular deviance, as in Supplementary Figure S2. *Right*: Detail of a 4×4 grid showing the cursor's movement animation.

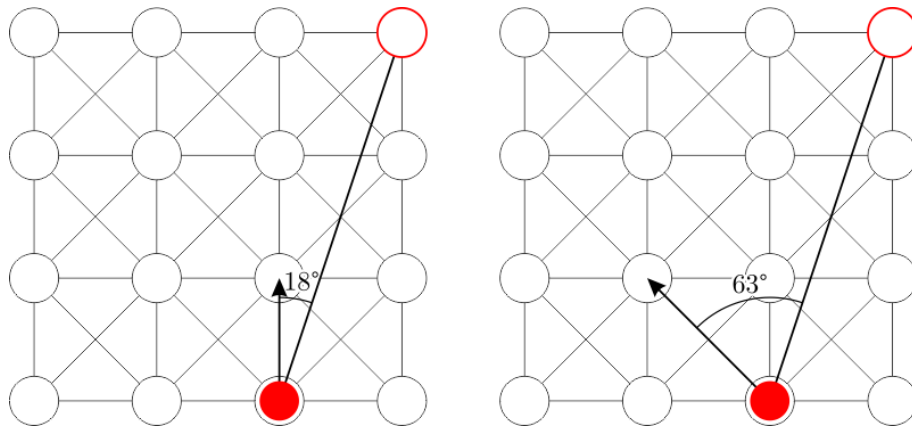


Fig. S2. Angular Deviance

Illustration of the angular deviance of two possible cursor movements (north, and north-west) from the optimal path straight towards the target. Angular deviance was measured as the absolute deviation, in degrees, of the movement direction from a straight line to the target.

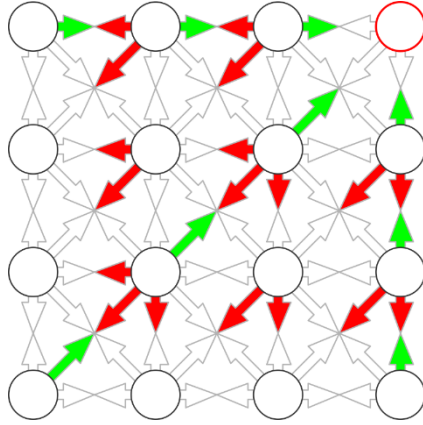


Fig. S3. Class Definition of the Classifier's Training Set

The classifier was calibrated on a subset of trials from a 600-trial calibration session. This set was selected by angular deviance: Movements with an absolute angular deviance to the target of 0° were included as 'correct', while movements with a deviance of 135° or more were included as 'incorrect'. This figure illustrates which of all possible movements on the 4×4 grid were included, and in which category: Green represents 'correct', red 'incorrect'. From the total of 600 trials per participant on the 4×4 grids, this selection left 62.7 ± 7.8 'correct' and 124.4 ± 8.9 'not acceptable' trials per participant for the classifier to be calibrated on.

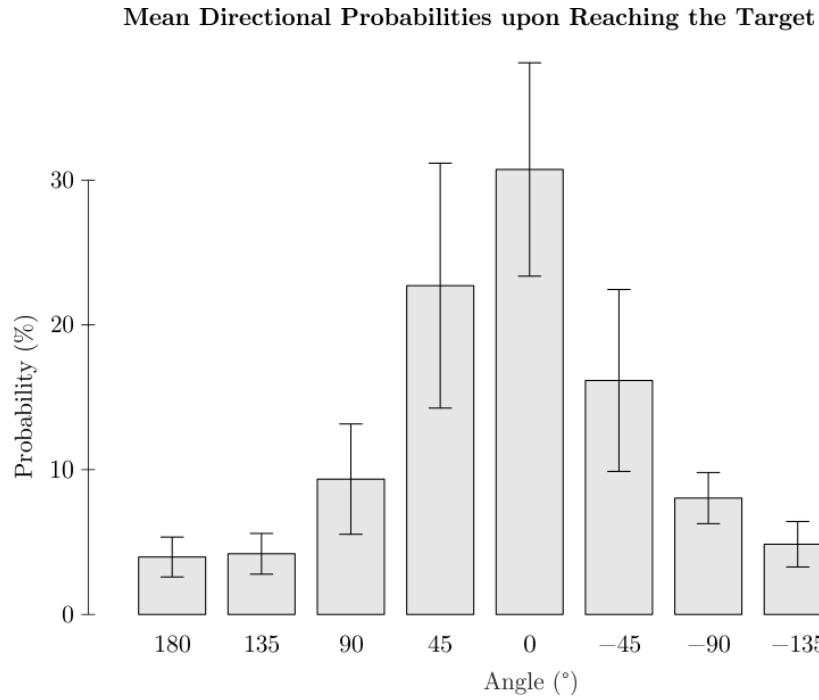


Fig. S4. Directional Probabilities

Mean directional probabilities, relative to the target from the cursor's starting position, upon hitting the target, in the BCI-supported condition, averaged over participants. Error bars represent the standard deviation. See Supplementary Table S2 for significance tests.

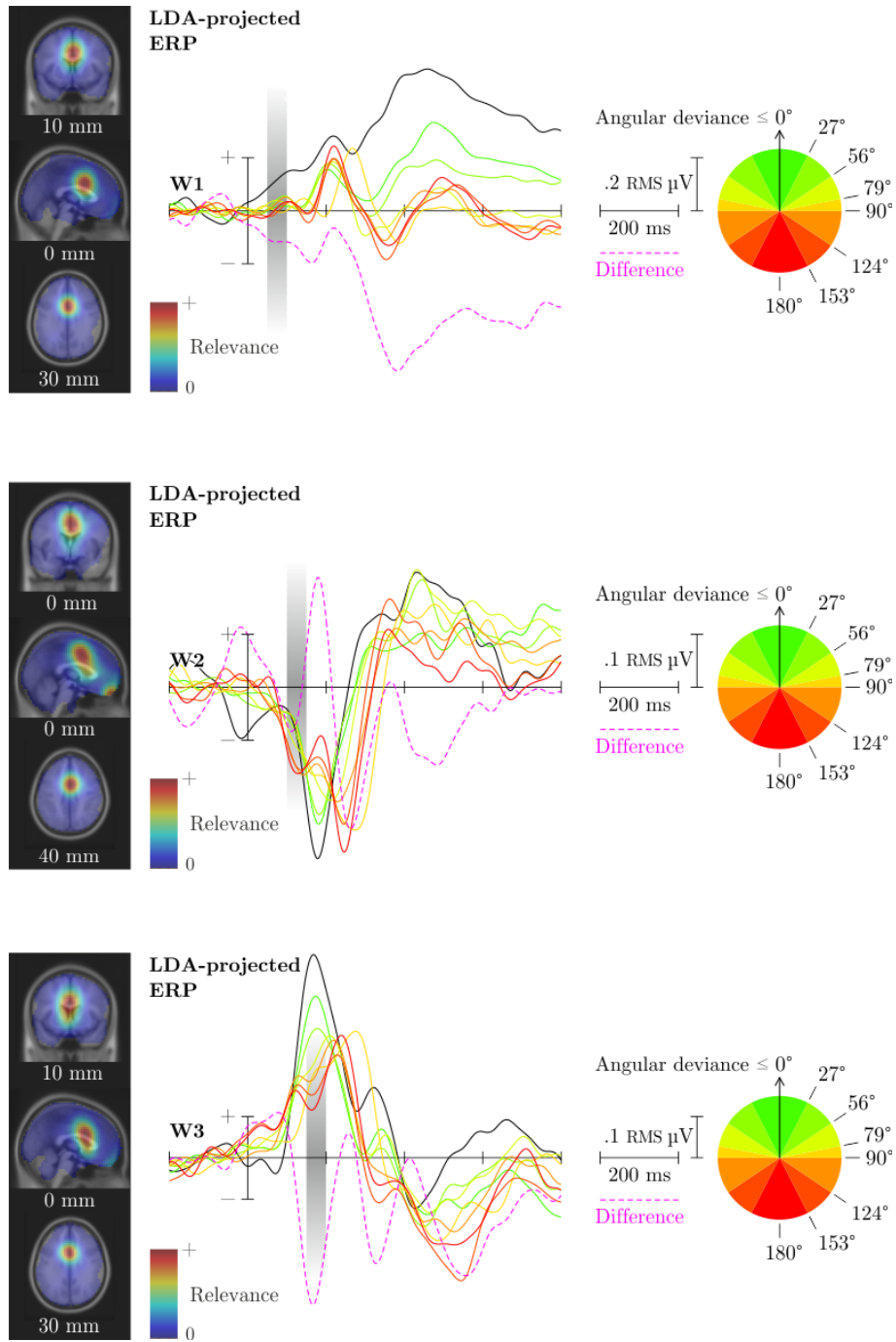


Fig. S5. LDA-projected ERPs

(Figure spans three pages; this is page 1 of 3.)

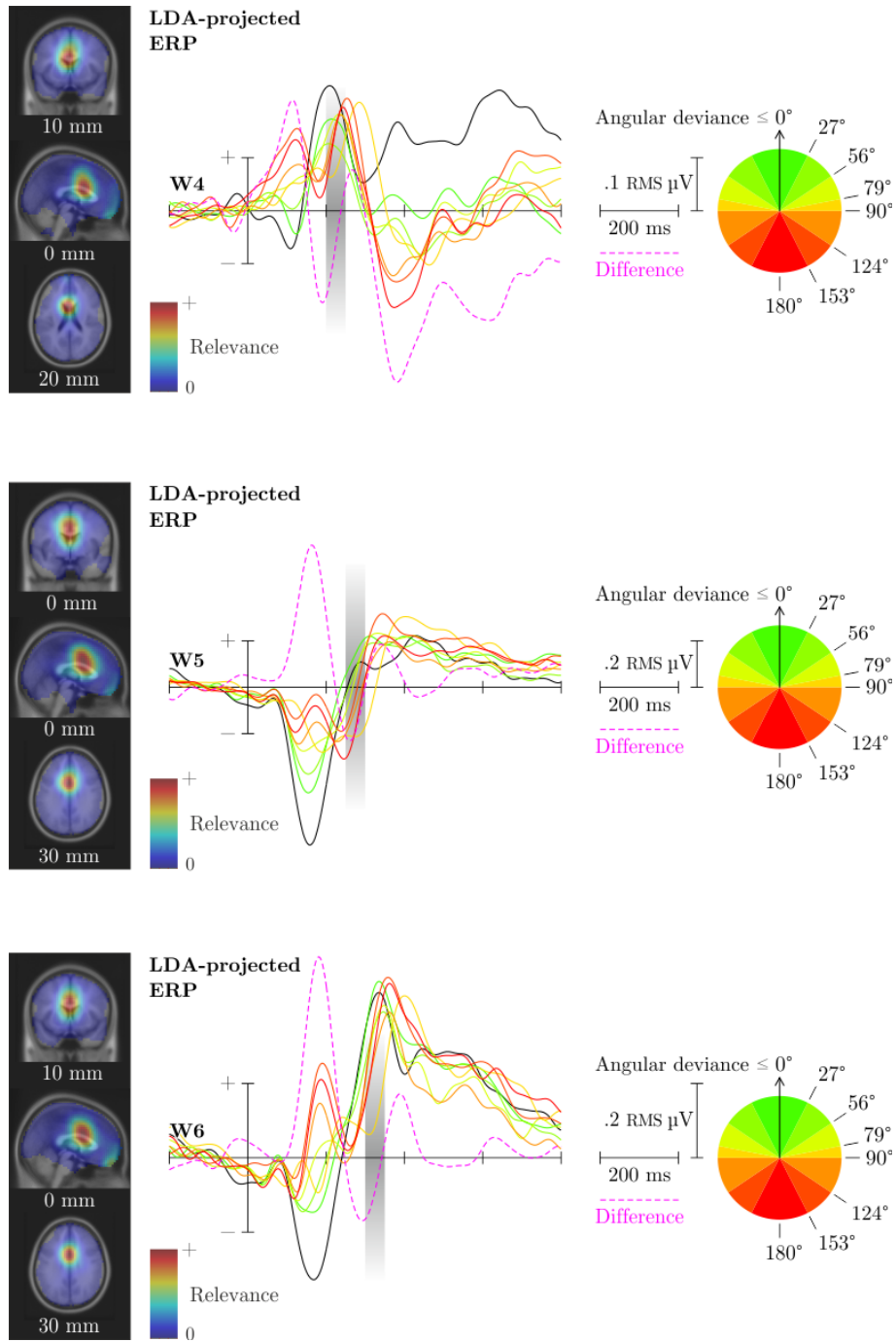


Fig. S5. LDA-projected ERPs

(Figure spans three pages; this is page 2 of 3.)

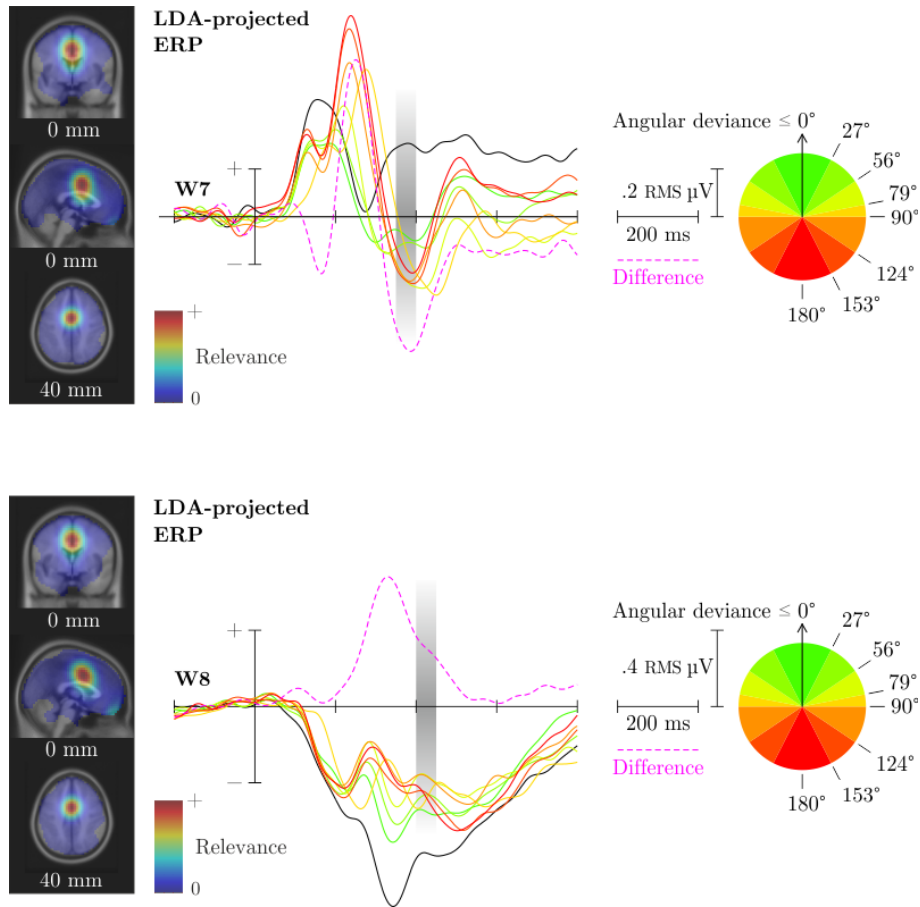


Fig. S5. LDA-projected ERPs

(Figure spans three pages; this is page 3 of 3.)

The classifier was calibrated on features spanning eight time windows of 50 ms each. For each time window, the above figures show on the *left*: Source localisation via weighted dipole density of the process(es) focused on in that time window; *right*: Full-length ERPs, combined as per that time window's LDA filter, of eight groups of cursor movement, as well as the projected difference wave between 'correct' and 'incorrect' classes (i.e. the filter's optimisation target). Each figure's actual time window, which actually contributed to classification, is highlighted in grey.

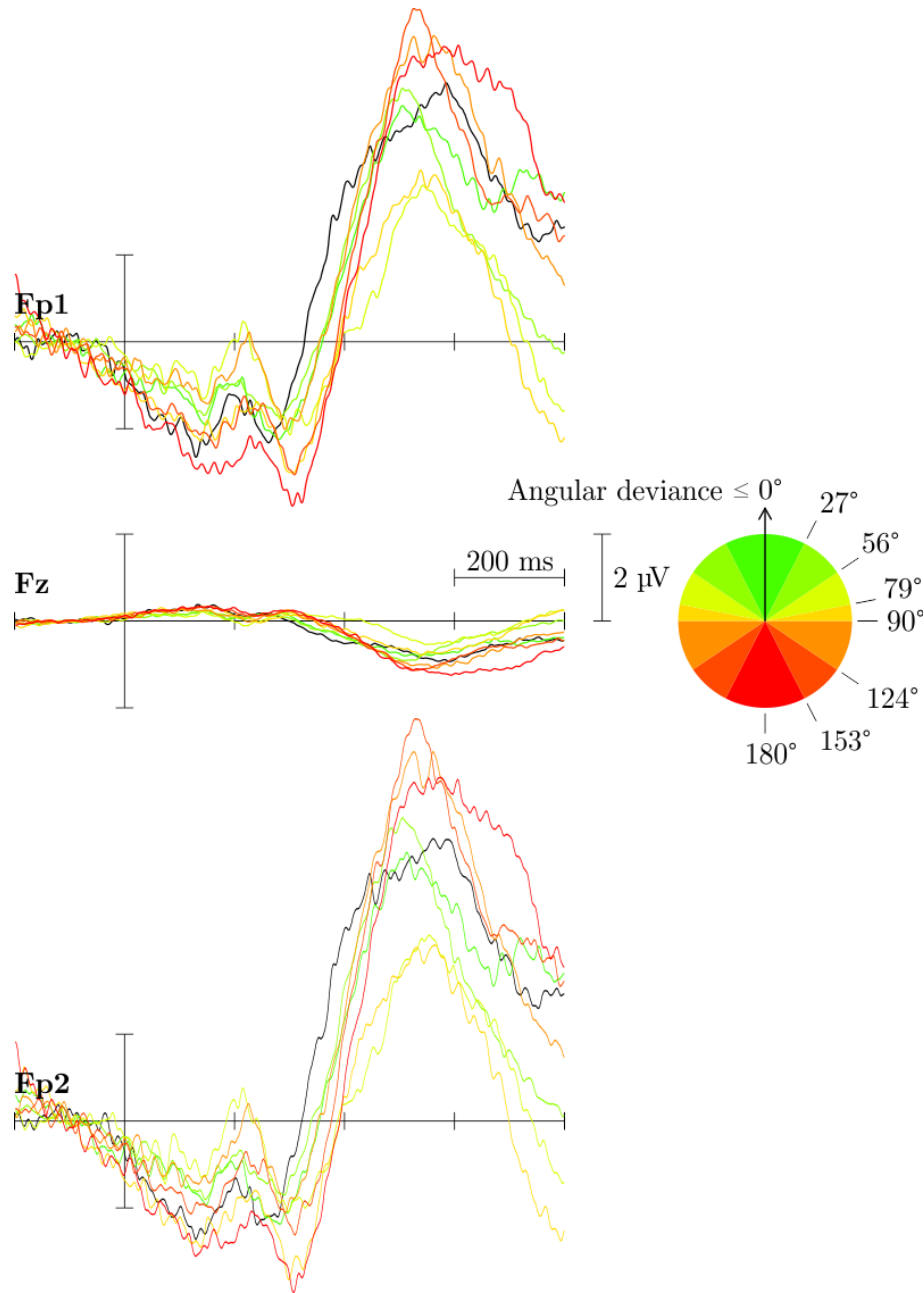


Fig. S6. Influence of Eye Movements on the ERP at Fz

The grand average ERP ($n = 19$) at Fp1, Fp2, and Fz for only those components that contained strong eye movements and were removed for neurophysiological analysis. No systematic response is visible.

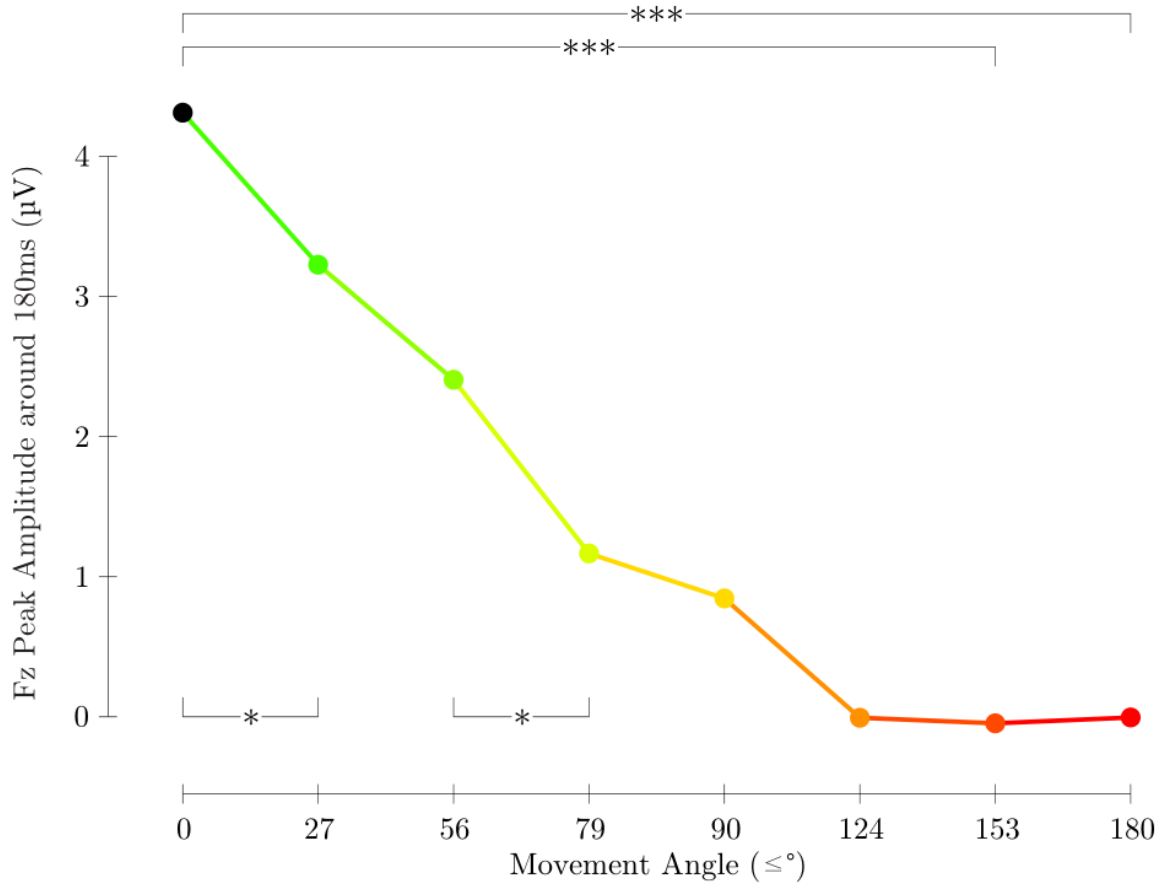


Fig. S7. Mean Peak Amplitudes at Fz

Peak amplitudes around 180 ms for the eight groups of cursor movements. The peak amplitudes differ significantly ($p < 0.001$ ***) between the classes used by the classifier. In between, the peak amplitudes scale linearly with angular deviance, as fitted by a linear regression model using each group's mean angular deviance as predictor ($b = -0.0035$, $F = 45.28$, $p < 0.001$; $R^2 = 0.33$). Statistically significant differences between adjacent groups are indicated as well ($p < 0.05$ *). See Supplementary Table S3 for exact figures of all comparisons. Classifier output followed a similar trend, as in Figure 4 of the manuscript.

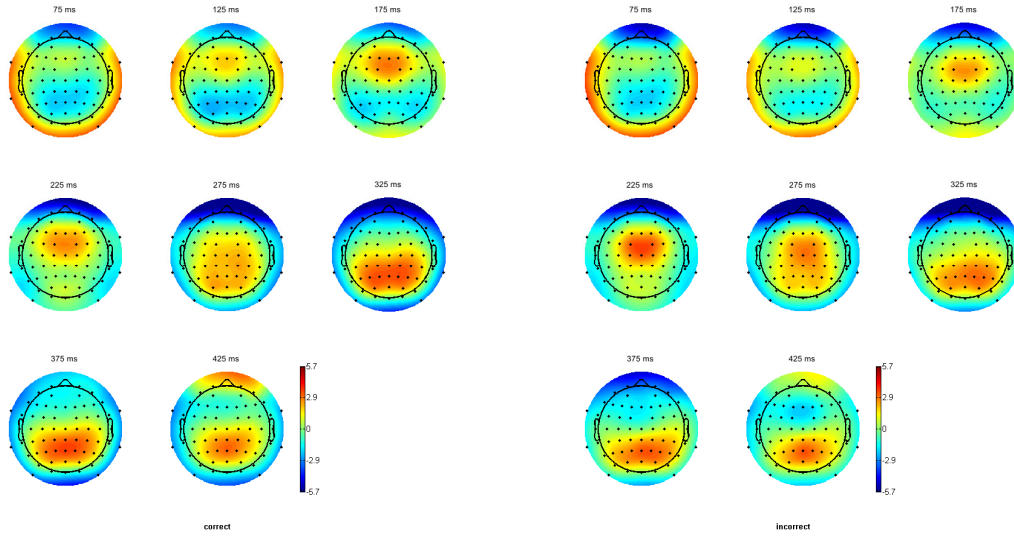


Fig. S8. Scalp Activity Series

Class-specific grand average scalp activities ($n = 19$), time-locked to cursor movement, at the middle of each time of the eight time windows used for classification. *Left*: Scalp activities of class 1, containing only cursor movements that went directly towards the target. *Right*: Scalp activities of class 2, containing cursor movements with an angular deviance of 135° or more. EEG data was first band-pass filtered from 0.1 to 5 Hz as per the classification approach and re-referenced to the common average. See Figure 1 in the main manuscript and Supplementary Movie S1 for the class-correlated scalp activity that the classifier focused on.

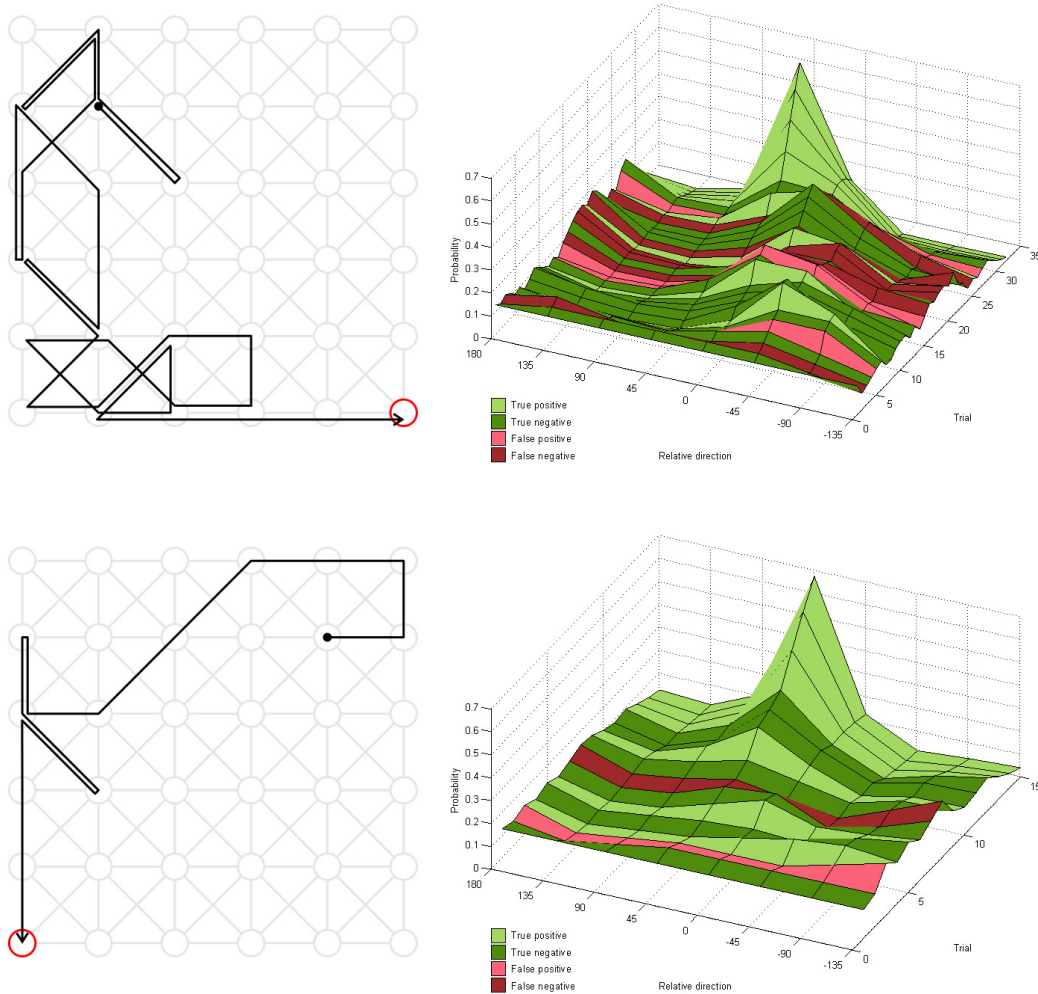


Fig. S9. Example Online Cursor Behavior

Visualization of two online target selections. *Left*: Trace of cursor movements over the grid. Each cursor movement only progressed one node over the grid; extended straight lines thus reflect multiple movements. *Right*: Progression of normalized directional probabilities relative to 0° , with 0° being the direction towards the target at the start of the grid. All directions started out with equal probabilities, indicated at trial 0. Each subsequent trial shows the directional probabilities after that trial's cursor movement was classified and the cursor was reinforced accordingly. The colors reflect the confusion matrix with the participant's button presses being used as ground truth. True positive (light green) indicates the cursor movement was correctly classified as 'correct'. True negative (dark green) indicates a correctly classified 'incorrect' movement; false positive (light red) an incorrectly classified 'correct' movement; false negative (dark red) an incorrectly classified 'incorrect' movement. The top row shows a relatively slow online run of 31 trials on a six-by-six grid. A number of misclassifications, especially near the middle, delay the probabilistic model's convergence towards the desired target direction. However, the erroneous trials do not lead to an adverse bias while the correct classifications do systematically point towards the target, and the correct direction is found. The lower row shows a shorter run of 14 trials with only two misclassifications.

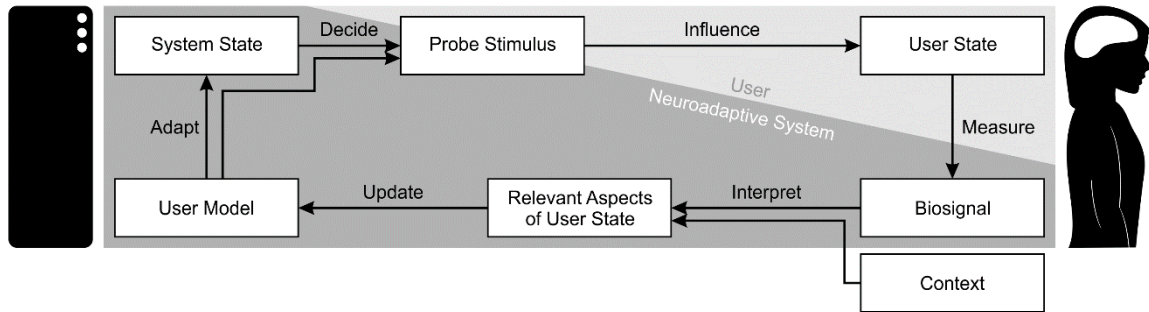


Fig. S10. The neuroadaptive control loop

A stimulus or parameter change provokes an automatic response from the user, whose biosignals are monitored. The user's response can be classified from the gathered data, allowing the system to interpret the response in light of the previously gathered user information and the current context. With this information, the system updates its user model. Based on the user model and the current system status, the system may decide on a new probe stimulus.

Tables

Participant	4 × 4 Grids			6 × 6 Grids		
	TP	TN	CR	TP	TN	CR
07	.91	.86	.89	.	.	.
08	.91	.60	.77	.	.	.
09	.81	.53	.71	.	.	.
10	.83	.79	.81	.76	.75	.76
11	1.00	.83	.93	.88	1.00	.93
12	.83	.63	.76	.82	.64	.73
13	.67	.47	.59	.87	.38	.67
14	.75	.63	.70	.84	.36	.69
15	.82	.69	.75	.71	.82	.75
16	.52	.50	.52	.39	.10	.29
17	.58	.62	.59	.50	.68	.58
18	.67	.62	.65	.44	.53	.48
19	.80	.67	.75	.92	.65	.80
20	.71	.80	.74	.70	.62	.67
21	.88	.56	.76	.74	.45	.63
22	.60	.67	.63	.40	.50	.43
Mean	.77	.65	.72	.69	.58	.65

Table S1. Online Classification Accuracy

Online classification rates for the 4×4 grid, and the 6×6 grid. TP = True Positives, TN = True Negatives, CR = Classification Rate (combined TP and TN).

Supplementary Table S1 lists the classification rates for the online blocks, i.e. what percentage of classifier output agreed with the given definitions of ‘correct’ and ‘incorrect’ movements. These definitions were the same as the ones used for calibration: Movements with an angular deviance to the target of 0° were defined to be ‘correct’, and those with a deviance of 135° or more were taken to be ‘incorrect’. The labels of all other movements were subject to individual interpretation, and have therefore not been included in this analysis.

The table lists these rates for all participants that completed online blocks. From the total of nineteen, this excludes the first three participants, who only performed offline calibration blocks. A further three participants only performed 4×4 grids online.

Pairwise Comparison							
	180°	135°	90°	45°	0°	-45°	-90°
135°	1.000
90°	0.004	0.003
45°	0.000	0.000	0.000
0°	0.000	0.000	0.000	1.000	.	.	.
-45°	0.000	0.000	0.304	1.000	0.000	.	.
-90°	0.000	0.000	1.000	0.000	0.000	0.006	.
-135°	1.000	1.000	0.005	0.000	0.000	0.000	0.000

Table S2. Post Hoc: Final Direction Probabilities

Bonferroni-adjusted pairwise comparisons of the mean directional probabilities upon reaching the target.

Supplementary Figure S2 shows the mean directional probabilities upon reaching the target in both online grids, sorted by their angular deviance to the target, which was fixed relative to the cursor's starting position. Supplementary Table S2 lists the Bonferroni-adjusted results of pairwise post hoc tests from a one-way ANOVA ($F(7,105) = 57.520, p < 0.001$) on this data. On average, the classifier has been able to reinforce the 'correct' directions significantly more strongly than 'incorrect' directions.

	Pairwise Comparison						
	0°	≤ 27°	≤ 56°	≤ 79°	≤ 90°	≤ 124°	≤ 153°
≤ 27°	0.032
≤ 56°	0.001	0.075
≤ 79°	0.000	0.000	0.020
≤ 90°	0.000	0.000	0.004	0.309	.	.	.
≤ 124°	0.000	0.000	0.000	0.023	0.071	.	.
≤ 153°	0.000	0.000	0.000	0.021	0.066	0.507	.
≤ 180°	0.000	0.000	0.000	0.023	0.071	0.520	0.531

Table S3. Post Hoc: Peak Amplitude Differences at Fz

Pairwise comparisons, adjusted for false discovery rate, of the mean peak amplitudes at Fz for the eight angular categories.

A one-way ANOVA indicated a significant influence of angular deviance on peak amplitude ($F(7,126) = 47.24, p < 0.001$). In Supplementary Table S3 are listed the post-hoc comparisons—one-sided t-tests with pooled standard deviations, corrected for false discovery rate—between the eight individual groups of cursor movement.

	Pairwise Comparison						
	0°	≤ 27°	≤ 56°	≤ 79°	≤ 90°	≤ 124°	≤ 153°
≤ 27°	0.328
≤ 56°	0.000	0.002
≤ 79°	0.000	0.000	0.023
≤ 90°	0.000	0.000	0.000	0.115	.	.	.
≤ 124°	0.000	0.000	0.000	0.006	0.125	.	.
≤ 153°	0.000	0.000	0.000	0.308	0.311	0.829	.
≤ 180°	0.000	0.000	0.000	0.311	0.829	0.979	0.952

Table S4. Post Hoc: Classifier Output

Pairwise comparisons, adjusted for false discovery rate, of the mean peak amplitudes at Fz for the eight angular categories.

A one-way ANOVA indicated a significant influence of angular deviance on classifier output ($F(7,105) = 28.32, p < 0.001$). Below are listed the post-hoc comparisons—one-sided t-tests with pooled standard deviations, corrected for false discovery rate—between the eight individual groups of cursor movement.

	Pairwise Comparison						
	0°	≤ 27°	≤ 56°	≤ 79°	≤ 90°	≤ 124°	≤ 153°
≤ 27°	0.004
≤ 56°	0.000	0.007
≤ 79°	0.000	0.000	0.009
≤ 90°	0.000	0.000	0.000	0.093	.	.	.
≤ 124°	0.000	0.000	0.000	0.168	0.599	.	.
≤ 153°	0.000	0.000	0.000	0.014	0.161	0.134	.
≤ 180°	0.000	0.000	0.000	0.004	0.046	0.404	0.227

Table S5. Pairwise Comparisons: Projected ERP Peak Amplitude Differences

Pairwise comparisons of the mean amplitudes in the third time window (150-200 ms) of the LDA-projected ERP.

Supplementary Figure S5 shows LDA-projected ERPs. At Fz, significant amplitude differences were found at around 180 ms following cursor onset (see Figure 1 of the main manuscript and Supplementary Figure S7). This falls within the third time window used by the classification system (150 to 200 ms following cursor movement). This table shows the results of pairwise comparisons using one-tailed permutation tests of the mean amplitudes of the LDA-projected ERPs in that time window, between the eight individual groups of cursor movement.

References

1. Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87(0), 96–110.
2. Palmer, J. (2012). AMICA—Adaptive Mixture ICA. Retrieved 2014-04-25, from http://sccn.ucsd.edu/~jason/amica_web.html.
3. Oostenveld, R. & Delorme, A. (2003). DIPFIT 2.x. Retrieved 2014-04-25, from http://sccn.ucsd.edu/wiki/EEGLAB_Extensions_and_plugin-ins.
4. Miyakoshi, M. (2003). dipoleDensity. Retrieved 2014-04-25, from http://sccn.ucsd.edu/wiki/EEGLAB_Extensions_and_plugin-ins.