Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control

Authors: Iñaki Iturrate[1,2], Ricardo Chavarriaga[2], Luis Montesano[1], Javier Minguez[1], and José del R. Millán[2*]

Affiliations:

[1]Instituto de Investigación en Ingeniería de Aragón, Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Spain

[2]Chair in Non-Invasive Brain-Machine Interface, Center for Neuroprosthetics & Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

*To whom correspondence should be addressed: E-mail: jose.millan@epfl.ch

Abstract:

Brain-machine interfaces (BMI) usually decode movement parameters from cortical activity to control neuroprostheses. This requires subjects to learn to modulate their brain activity to convey all necessary information, thus imposing natural limits on the complexity of tasks that can be performed. Here we demonstrate an alternative and complementary BMI paradigm that overcomes that limitation by decoding cognitive brain signals associated with monitoring processes relevant for achieving goals. In our approach the neuroprosthesis executes actions that the subject evaluates as erroneous or correct, and exploits the brain correlates of this assessment to learn suitable motor behaviours. Results show that, after a short user's training period, this teaching BMI paradigm operated three different neuroprostheses and generalized across several targets. Our results further support that these error-related signals reflect a task-independent monitoring mechanism in the brain, making this teaching paradigm scalable. We anticipate this BMI approach to become a key component of any neuroprosthesis that mimics natural motor control as it enables continuous adaptation in the absence of explicit information about goals. Furthermore, our paradigm can seamlessly incorporate other cognitive signals and conventional neuroprosthetic approaches, invasive or non-invasive, to enlarge the range and complexity of tasks that can be accomplished.

Introduction

Research on brain-machine interfaces (BMI) has demonstrated how subjects can voluntary modulate brain signals to operate neuroprosthetic devices[1-6]. These BMIs typically decode cortical correlates of movement parameters (velocity/position[1,5-8,9] or muscular activity[4]) in order to generate the sequence of movements for the neuroprosthesis. This control approach directly links neural activity to motor behaviour[10]. Mounting evidence, however, seems to suggest that motor control is the result of the combined activity of the cerebral cortex, subcortical areas and spinal cord. In fact, many elements of skilled movements, from manipulation to walking, are mainly handled at the brainstem and spinal cord level with cortical areas providing an abstraction of the desired movement such as goals and movement onset[11]. A BMI can mimic this principle, as studies have shown the feasibility to decode such a kind of cognitive information associated with voluntary goal-directed movements[3,12,13]. As an advantage of this approach over typical BMIs, and once the individual decoders are learnt, subjects do not need to learn to modulate their brain activity in order to generate all necessary movement parameters to operate the neuroprosthesis, which imposes natural limits on the complexity of tasks that can be solved. Nevertheless, this approach requires an intelligent neuroprosthesis, emulating the roles of the subcortical areas and spinal cord, capable to learn and generate the desired behaviours.

Here we demonstrate this alternative teaching paradigm (Fig. 1), where the neuroprosthesis learns optimal motor behaviours (or control policies) to reach a target location based on the decoding of human brain signals that carry cognitive information about the appropriateness of goal-directed movements —i.e., the error-related potential (ErrP)[14,15], a time-locked potential elicited when actions do not match users' expectations[16-21]. Error-related signals have been recently used to correct or adapt BMI decoders using both invasive[22,23] and non-invasive recordings[15,24]. In our paradigm error information is exploited to learn a motor behaviour that accomplishes the user's intended task from a set of basic pre-programmed actions. The user monitors the performance of the neuroprothesis as it executes a sequence of these actions. ErrPs are evoked by actions that the user considers wrong to achieve his desired goals, decoded online, and employed as a reward signal for a reinforcement learning algorithm (RL)[25] that improves the neuroprosthesis behaviour. We tested this approach in three closed-loop experiments of increasing real-life applicability involving twelve subjects (Fig. 2A). They ranged from 1D cursor movement, to a simulated robot, and, finally, a real robot arm—both robots operating in a 2D space.

Each experiment consists of two phases: training the ErrP decoder from the user's electroencephalogram (EEG) signals and online operation of the neuroprosthesis which, using the trained decoder, learns different reaching tasks. For training the decoder, each subject observed around 350 robot movements (or device actions) while it tries to reach predefined targets with 20% of wrong actions (i.e., movements away from the target location). During online operation, subjects, but not the neuroprosthesis, knew the target location and monitored the performance of the device. One run, lasting 100 device actions, was performed for each possible target (i.e., circles in Fig. 2A). The device controller was initialized to a random behaviour (i.e., equiprobable actions for all states) at the beginning of each run, and updated after each action based on the online decoding of the ErrPs. Whenever the device reached the target, the former was randomly reset to a new location. For experiments 2 and 3, there were two targets (practice targets) that were used during ErrP calibration and online operation; and two targets (new targets) that were only used during online operation.

Results

Decoding error-related EEG potentials

ErrPs elicited in all protocols were consistent with previous studies[20,21]. The difference event-related potential (ERP) for erroneous and correct actions of the device exhibited a characteristic waveform with prominent fronto-central positive and negative peaks at around 300 and 500 ms, respectively. Fig. 2B shows these ERPs at electrode FCz for all subjects in the three experiments.

Statistically significant effects were observed on the latency but not the amplitude of these ERPs[27,28] (see Supplementary Material). Accuracy of online single-trial decoding of the ErrPs was comparable for all experiments, independently of the of the task performed. Classification performance (73.8%, 72.5%, 74.3% on average for experiments 1 to 3 respectively) exceeded the chance level (except for one subject in experiment 2, see Fig. 3a) —a necessary condition for a reinforcement learning system to acquire an optimal control policy[25]. Remarkably, this decoding performance remained similar to the overall accuracy (FDR-corrected two-tailed independent t-test, $p>0.05$) during the whole experiment (see Fig. 3B) despite the fact that the neuroprotheses move randomly at the beginning of an experiment and the error rate decreases as the devices learn an optimal motor behaviour (i.e., control policy).

To discard the influence of artifacts on the ErrP decoding, the data used to train the classifier included all possible movements for each class, thus reducing the possibility that classification was biased by their directions. For instance, during experiment 1, both targets are used for the classifier training, thus the error and correct assessments are not likely to be correlated with left or right eye movements. Moreover, results obtained when testing new targets in experiments 2 and 3 further support the fact that ErrP classification depends on the movement evaluation and not on its direction. Indeed, the training set only contained samples where the target locations were Up and Down, while the BMI was also tested on targets Left and Right. Finally, to test whether the trained classifier discriminated different directions rather than assessments, we computed for each subject the accuracy of decoding the different pairs of movement directions (e.g., left versus right, up versus left, ...) from a fixed assessment (either correct or erroneous) with the same features and classifier used during the experiments. The mean accuracies obtained were of 52.16±5.22, 50.07±5.07, and 49.48± 6.41 for experiments 1 to 3, and thus did not exceed the chance levels of 56%, 54% and 54% (see Methods, 'ErrP classifier'), proving that the classifier was not trained to distinguish movement directions or associated ocular artifacts, but user's assessments. Additionally, statistical analysis of grand average ERPs shows significant differences only for assessment, but not for movement direction (see Methods and Supplementary Fig. S1).

In summary, our ERP analysis supports the hypothesis that ErrPs reflect a common phenomenon across all experiments, where the protocol mainly affects the temporal characteristics of the brain response.

ErrP-mediated acquisition of control policies

During online operation, the device converged to steady performance after 4 targets (Fig. 2C and Fig. S2A), thus rapidly acquiring (quasi) optimal policies in all experiments and reaching desired targets from any starting position. On average users reached 12.38±5.66, 12.46±5.40 and 12.75±6.63 targets per run for experiments 1, 2 and 3, respectively (see Fig. 4). In contrast, the number of targets reached following a random control policy is 2.27±1.56 for experiment 1, and 2.32±1.54 for the other experiments. Despite there was a large variability in the convergence rate among subjects, most of them reached a number of targets significantly greater than chance ($\alpha=0.05$). (see Fig. 2C, Fig. 4, Supplementary Fig. S2, and Supplementary movie). Figure 2C summarizes the performance in the three experiments when the corresponding device was tested on the same target locations used for training the ErrP classifier. It shows the number of actions required to reach each target within a run. Since the device was initialized at random positions, values were normalized to the initial distance to the target (i.e., a value of one corresponds to optimal performance). For illustration, data from all subjects and all the targets was fitted to an exponential function. For all subjects and targets there was a rapid decrease in the number of actions that converged towards values close to optimal performance; thus reflecting the acquisition of quasi-optimal behaviours (Fig. 4).

In all experiments, the number of learned optimal actions consistently increased as more actions were performed. For experiment 1, the number of optimal actions learned for the visited states was significantly above chance level after 10 actions (false discovery rate (FDR)-corrected one-tailed unpaired t-tests, $p < 0.05$). Furthermore, it consistently increased as more actions were performed (correlation $r = 0.74$, $p < 1\times10^{-8}$). The number of normalized actions required to reach the target converges, to 1.19±0.52 after 9 targets reached, very close to the optimal value (Fig. 2C; red trace). For experiments 2 and 3, the convergence was slower due to the higher number of states and actions: whereas for experiment 1 the number of actions learnt was above chance after 10 actions, for experiments 2 and 3 it was necessary to execute 15 actions to surpass chance level. (FDR-corrected one-tailed unpaired t-tests, $p < 0.05$). As in experiment 1, there was also a high correlation between the amount of performed actions and optimal actions learned ($r = 0.84$, $p < 1\times10^{-8}$ for both experiments). The number of normalized actions required to reach the target for experiments 2 and 3 was 2.00±0.76 and 1.97±0.75, slightly worse than for Experiment 1 (Fig. 2C). In summary, all brain-controlled devices were operational almost from the beginning of the run (above chance results after a few actions), improving performance progressively over time (significant correlation between time and number of actions learnt) and approaching optimal behaviour at the end of each run (number of targets reached increasing throughout time, Fig. 4).

Learning control policies to reach new targets

Experiments with the robot arms demonstrate that control policies can be easily acquired to reach new targets, without the need of retraining the ErrP decoder. Figure 5A shows for the real robot experiment the number of actions required for practice and new targets. In both cases, the system improves its policy and approaches the optimal behaviour (see Fig. 4 and Supplementary Fig. S2). On average, users reached 14.42±7.81 and 12.54±6.44 targets for experiments 2 and 3 respectively, significantly similar to the ones reached during practice targets (two-tailed paired t-test, p=0.22 and p=0.90). Figures 5B and C illustrate the optimal policy for one practice and one new target, respectively. For experiments 2 and 3, there were no significant differences in the number of optimal actions learned between practice and new targets (FDR-corrected two-tailed paired t-test, $p = 0.47$ and $p = 0.37$, respectively).

Similarly to the practice targets, the number of optimal actions learned was significantly above chance level after 4 and 14 performed actions for experiment 2 and 3, respectively (FDR-corrected one-tailed unpaired t-tests, $p < 0.05$); and with a high correlation between the amount of performed actions and optimal actions learned ($r = 0.80$, $p < 1\times10^{-8}$ for both experiments). The final number of actions per target for these experiments was 1.81±0.48 and 1.47±1.12. This confirms that the ErrP does not depend on targets, as the ErrP classifier maintains its performance without needing to be retrained for unseen targets.

Discussion

These experiments illustrate a number of appealing properties associated with the use of error-related brain signals to allow a BMI to teach neuroprostheses suitable motor behaviours. First, we exploit a brain signal naturally elicited by the user, without requiring the explicit learning and execution of *ad-hoc* mental tasks. Moreover, user's training time is minimal —a calibration session is enough to model the user's ErrP decoder (25 minutes on average for each subject and experiment). Second, this paradigm makes it possible to achieve tasks in a user-specific manner —the learned control policy depends on the individual user's assessment. Third, single-trial decoding of ErrP does not need to be perfect to maneuver a neuroprosthesis —it suffices that the ErrP decoder performs statistically above random to learn the motor behaviour. Furthermore, the neuroprosthesis is operational as soon as the accuracy of the ErrP decoder is above chance level —which usually takes minutes as reported here— and keeps adapting indefinitely, as it is the case of human motor control. Finally, and perhaps more importantly, the ErrP is rather independent of the task (e.g., target or action type) —making control of neuroprostheses scalable to more complex tasks since the learning burden is on the robot side.

Scalability is indeed a crucial property of the teaching BMI approach since, as the experimental results demonstrate, ErrPs reflect a common error processing mechanism in the brain across tasks[18], and this was confirmed by our latest results, which showed that ErrP decoders can generalize across different tasks[27,28]. Importantly, error processing information from the brain can be observed using different recording signals such as human electroencephalogram (EEG)[17-21], electrocorticogram (ECoG)[23] and intracortical recordings[29]. ErrPs have also been reported and decoded in patients with severe motor disabilities[30]. Noteworthy, the development of adaptive mechanisms for BMIs is gaining increased attention[31-33]. ErrPs offer a natural alternative to drive adaptation in the absence of explicit information about goals for both invasive and non-invasive conventional control neuroprosthetic approaches.

As a first demonstration of the proposed paradigm, we have made use of one of the most straightforward and simple RL algorithms, Q-learning. However, this approach—as many other RL algorithms- suffers from two main problems: task generalization and scalability. It is then an open question how the proposed BMI paradigm may generalize across tasks or scale to more complex scenarios. Notwithstanding, current state of the art on reinforcement learning offers very promising alternatives for the tractability of generalization and high-dimensional spaces, such as the use of transfer learning or prior knowledge via demonstration among others[34].

ErrPs have been exploited to correct and adapt BMIs as well as to improve human-computer interaction[20,23,24,30]. Along this line, a possibility that has been explored in rodents is to extract information from the reward-processing brain areas (i.e., nucleus accumbens)[22] for RL-based adaptation of the BMI decoder[35]. Here we go beyond this BMI adaptation framework and, extending our previous 1D works[21,26], demonstrate for the first time in humans how the teaching BMI paradigm enables the acquisition of suitable control policies, scales to neuroprostheses with a task complexity similar to state-of-the-art BMIs, and generalizes across different targets.

In the experiments reported here, it is assumed that the neuroprosthesis owner wishes to initiate a voluntary, goal-directed movement whose low-level execution is delegated to subcortical, spinal cord and musculoskeletal structures. In our case, this lower level of motor control is emulated by an intelligent controller able to learn and to reuse control policies via ErrPs. Although a full demonstration of this extension to the teaching BMI approach remains to be proven, evidence suggests its feasibility. Firstly, cortical cognitive signals indicating goals[3,12], self-paced onset of movements[13], or anticipation of purposeful actions[36,37] can be decoded at the single-trial level. Secondly, several control policies can be learned as demonstrated here and stored to form a repertoire of motor behaviours (i.e., to reach different targets within the environment).

Note that, while in the current manuscript one control policy was associated with a specific target, one target could have several ways of being reached depending on the user's preferences. Once the control policies are stored, the decoding of the error-related signals can be exploited to infer the desired behaviour from this repertoire rather than being used to learn a new control policy. This possibility has been recently explored in[38].

We postulate that the combination of all these sorts of cognitive brain signals would be sufficient for chronic operation of neuroprostheses, whose range of tasks may change over time. Such a possibility is critical for patients —especially if suffering from neurodegenerative diseases— as they must rely upon neuroprostheses for extended periods of time. Despite remaining hurdles such as large clinical studies, further research will uncover additional cognitive brain signals that will enrich this initial basic set, thus enlarging the repertoire of decision-making processes available for natural, intuitive control of neuroprostheses to perform goal-directed movements and bringing BMI closer to therapeutic reality.

Methods

All experiments were carried out in accordance with the approved guidelines. Experimental protocols were approved by the Commission Cantonale (VD) d'éthique de la recherché sur l'être humain (protocole 137/10). Informed written consent was obtained from all participants that volunteered to perform the experiments.

Subjects and data recording

Twelve able-bodied volunteers (four females, 23-24 years) participated in the study. EEG signals were recorded from 16 active electrodes located at Fz, FC3, FC1, FCz, FC2, FC4, C3, C1, Cz, C2, C4, CP3, CP1, CPz, CP2, and CP4 (10/10 international system). The ground was placed on the forehead (AFz) and the reference on the left earlobe. EEG was digitized at 256 Hz, power-line notch filtered at 50 Hz, and band-pass filtered at [1, 10] Hz. To reduce signal contamination, participants were also asked to restrict eye movements and blinks to indicated resting periods.

Experimental setup

All participants performed three experiments of different complexity, evaluated using the NASA-TLX questionnaire (28.44±14.01, 42.33±23.31 and 46.92±22.19 for experiments 1 to 3) in a different age-matched set of nine subjects. Each experiment was carried out on a different day, lasting around 2.5 hours. The time elapsed between two consecutive experiments was 17.58±10.09 days. In all experiments, subjects were instructed to monitor the device while it tried to reach a target (only known by the subject) and to assess whether the device actions were correct or incorrect. Each experiment was divided into two phases: training and online operation of the neuroprosthesis. Each phase was composed of several runs, each run consisting of 100 device actions. During each run the target location remained fixed and, whenever the device reached that location, its position was randomly reset to a location at least two positions away from the target, see Fig. 2A. Target location was randomly chosen between runs.

The training phase aimed at building a classifier able to detect error potentials. In the initial runs, the device performed erroneous actions with a fixed probability (20%). For all experiments, two target locations were used in this phase. After each run, all the collected data was used to train the ErrP classifier[27,28]. Once the decoding accuracy was above 80%, or four runs were elapsed, an additional run was executed where the output of the classifier was used to adapt the device controller using RL (see below). Thus, in this RL run the error probability was variable. If the accuracy in this RL run was below random, the classifier was retrained with additional RL runs until the criterion (accuracy above random) was reached. Subjects needed a median (± mean absolute deviation) of 1.00 ± 0.51 additional RL training runs. The mean duration of the entire training phase for all subjects and experiments was 25 minutes. The maximum length was 45 minutes.

In the online operation phase, the information decoded from the EEG (indicating whether the subject considered the action as correct or erroneous) was used as a reward signal to learn the behaviour through RL. One run was performed per target location (2 runs in the case of experiment 1, and 4 runs for experiments 2 and 3). In the last two experiments we tested the generalization capabilities of the proposed approach by including target locations that were not used in the training phase. In all RL runs, the device controller was initialized to a random behaviour where all actions at a given location are equiprobable.

Experiment 1: Moving Cursor[21] (Fig. 2A, Left). Participants faced a computer screen showing a horizontal grid with nine different positions (states; c.f., squares in Fig. 2A), including one blue moving cursor (device) and one red square (target). The cursor could execute two actions: move one position to the left or to the right. When the cursor was at the boundaries (i.e., at the left-or right-most states), actions that moved it out of the state space were not allowed. The time between two consecutive actions was drawn randomly from a uniform distribution within the range [1.7, 3.0] s. Only states at the left-most and right-most positions were used as targets.

Experiment 2: Simulated Robotic Arm (Fig. 2A, Center). Subjects faced a computer screen displaying a virtual robot (device). We simulated a Barrett whole arm manipulator (WAM) with 7 degrees of freedom using the RobotToolkit framework developed by the LASA laboratory at EPFL (http://lasa.epfl.ch). The robot could place its end-effector at 13 different positions (states; c.f., orange squares in Fig. 2A), with one position in green (target). It could perform four actions: moving one position to the left, right, up, or down. As before, when the device was at a boundary state, actions that moved the robot out of the state space were not allowed. In contrast to the first experiment, the robot movements between two states were continuous; lasting ~500 ms. The time between two consecutive actions was randomly distributed within the range [2.5, 4.0] s. During the training phase, the targets were located at up-and down-most positions (i.e., practice targets). For the online operation phase, the up-, down-, left-, and right-most positions were tested as targets.

Experiment 3: Real Robotic Arm (Fig. 2A, Right). This experiment followed the same design as experiment 2 but involving a real robotic arm (Barret WAM). The robot was two meters away from the user and was pointing at states on a Plexiglas transparent panel between the two. The distance between two neighbor states was 15 cm.

ErrP classifier

EEG signals were spatially filtered using common-average-reference and downsampled to 64Hz. Features were extracted as the signal from eight fronto-central channels (Fz, FCz, Cz, CPz, FC1, FC2, C1, and C2) within a time window of [200, 800] ms from the device movement onset and concatenated to form a vector of 312 features. These vectors were normalized and decorrelated using principal component analysis[39]. The most discriminant features were selected based on the r^2 score using a five-ten-fold cross-validation on the data of the training phase. On average, 36±13 features were selected per subject. ErrPs were decoded using a linear discriminant analysis (LDA) classifier.

To assess the statistical significance of the ErrP classifier accuracies during online operation, we compute the chance levels (α =0.05) according to the available number of trials using the binomial cumulative distribution[40]. The estimated chance levels were 56% for Experiment 1 and 54% for Experiments 2 and 3.

Reinforcement learning (RL) with ErrPs

The RL strategy[25] was modeled by a Markov decision process, denoted by the tuple $\{S, A, r, \gamma\}$ with S being the state space (the possible positions of the device), and A the action space (the possible actions of the device). The reward value r represented the goodness of the executed action at a given state and γ is a time discount factor. The goal of RL was to obtain a control policy $\pi{:}S\rightarrow A$ mapping the state space into the action space (i.e., which action had to be performed at each state) so as to maximize the expected return $R = \Sigma^{\infty}_{\mathrm{k}=0}\, \gamma^k r_{k+1}$ at time k. The RL implementation was the Q-learning iterative algorithm[25]:

$$Q_{k+1}(s_k, a_k) = Q_k\,(s_k, a_k) + \alpha[r_{k+1}(s_k, a_k) + \gamma \max_{a' \in \mathrm{A}} Q_k\,(s_{k+1}, a') - Q_k(s_k, a_k)], \qquad (1)$$

where k is the current step and α is the learning rate. Parameters γ and α were set empirically to 0.4 and 0.1, respectively. During online operation, at time k, the device executes an action a_k that takes it from state s_k to state s_{k+1}, according to its current policy. The output of the ErrP classifier is then used to obtain the reward value $r_{k+1}(s_k, a_k)$; it takes a value of -1 if the action is classified as error, otherwise is set to +1. This reward was used to update the RL policy after each action. All Q-values were set to zero at the beginning of each run (k=0), corresponding to a random control policy. At the end of the run, the final policy π was computed as the policy that, at each state s, always followed the action a' with the maximum Q-value, $\pi(s) = \arg\max_{a' \in \mathrm{A}} \mathrm{Q}(s, a')$.

At each step k, a ε-greedy strategy was used to select the next action a_k to be executed. This policy selected the action with highest Q-value (best action) for $(100-\varepsilon)$ % of the times, while a random action was selected the remaining times. The experiments started with a completely exploratory behaviour (ε = 100%), and every time an exploratory action was chosen ε was decreased by a constant factor (5%) until reaching a minimum value (20%) to always maintain a small percentage of exploration.

Acquisition of control policies (or behaviours)

We evaluated the acquisition of control policies using the number of optimal actions (i.e., those leading to the target location, c.f. arrows in Fig. 5B, and Fig. 5C) learned by the controller at a given time. Only states already visited were considered in this measure. We also compared the number of optimal actions learned to the chance level, computed as the number of actions learned with random rewards (i.e., ±1 with equal probabilities). Statistical tests were corrected with the false discovery rate, FDR[41].

Learning of the control policies was also assessed in terms of the number of actions required to reach the target location within a run. To account for the different initial states, the number of actions is divided by the initial distance to the target. For illustration purposes, we fitted the data of each experiment to an exponential curve, $y = a + be^{-cx}$, where y is the normalized number of actions required to reach the target for the x-th time (c.f., Figs. 2C, 5A and Supplementary Fig. S2).

Analysis of ocular artifacts

We assessed the possibility of EEG signal contamination by movement-related ocular artifacts. We computed the grand average ERPs (correct and error) of all channels separately for each different action (moving left, right, up, or down). No substantial differences were found among these ERPs, suggesting little influence of eye movements. This is illustrated in Supplementary Fig. S1 that shows the averages of three fronto-central electrodes (FC3, FCz and FC4), separated by assessment (correct or error) and movement direction (left, right, up or down). As can be seen, the differences among assessments were larger than the differences among directions. This is consistent with previous studies that found no influence of this type for experiment 1[20,21].

To evaluate the existence of statistical differences due to both assessments and movement directions, we performed 2 (factor assessment: error or correct) x 4 (factor movement direction: left, right, up or down) within-subjects ANOVAs on the values of the most prominent positive and negative peak amplitudes of the grand averages of channel FCz (note that for experiment 1 the ANOVA was 2 x 2 since there were only two possible movement directions). When needed, the Geisser-Greenhouse correction was applied to assure sphericity. The assessment and direction main effects and the assessment x direction interaction were studied.

Regarding the main effects, statistical differences were found for the assessment for all the experiments, for the positive ($F_{1,11} = 17.277$, $p = 0.002$, $F_{1,11} = 15.567$, $p = 0.002$, and $F_{1,11} = 14.202$, $p = 0.003$ for experiments 1 to 3) and negative ($F_{1,11} = 10.087$, $p = 0.009$, $F_{1,11} = 14.658$, $p = 0.003$, and $F_{1,11} = 11.581$, $p = 0.006$) peaks. On the contrary, no significant differences were found for the direction main effect ($p > 0.1$). Regarding the assessment x direction interaction, significant differences were found during experiment 2 ($F_{3,33} = 3.721$, $p = 0.02$ and $F_{3,33} = 3.903$, $p = 0.02$ for the positive and negative peak); and during experiment 3 for the negative peak ($F_{3,33} = 3.461$, $p = 0.03$) but not for the rest of the cases ($p > 0.35$). These results indicated that the largest differences on the potentials were due to the different assessments (error / correct), whereas the movement directions of the device affected less the potentials.

References and Notes:

1. Carmena, J. M. et al. Learning to control a brain-machine interface for reaching and grasping by primates. *PLoS Biol.* **1**, 193–208 (2003).

2. Millán, J. d. R., Renkens, F., Mouriño, J. & Gerstner W. Noninvasive brain-actuated control of a mobile robot by human EEG. *IEEE Trans. Biomed. Eng.* **51**, 1026–1033 (2004).

3. Musallam, S., Corneil, B. D., Greger, B., Scherberger, H. & Andersen, R. A. Cognitive control signals for neural prosthetics. *Science* **305**, 162–163 (2004).

4. Ethier, C., Oby, E. R., M. J. Bauman, L. E. Miller. Restoration of grasp following paralysis through brain-controlled stimulation of muscles. *Nature* **485**, 368–371 (2012).

5. Hochberg, L. R. et. al. Reach and grasp by people with tetraplegia using a neurally controlled robotic arm. *Nature* **485**, 372–375 (2012).

6. Collinger, J. L. et al. High-performance neuroprosthetic control by an individual with tetraplegia. *The Lancet* **381**, 557–564 (2013).

7. Wolpaw, J. R. & McFarland, D. J. Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 17849–17854 (2004).

8. Santhanam G., Ryu, S. I., Yu, B. M., Afshar, A. & Shenoy, K. V. A high-performance brain-computer interface. *Nature* **442**, 195–198 (2006).

9. Aflalo, T., Kellis, S., Klaes, C., Lee, B., Shi, Y., Pejsa, K., Shanfield, K., Hayes-Jackson, S., Aisen, M., Heck, C., Liu, C. & Andersen, R.A. Decoding motor imagery from the posterior parietal cortex of a tetraplegic human. *Science* **348**, 906-910 (2015).

10. Scott, S. Optimal feedback control and the neural basis of volitional motor control. *Nat. Rev. Neurosci.* **5**, 534–546 (2004).

11. Courtine, G. et al. Transformation of nonfunctional spinal circuits into functional states after the loss of brain input. *Nat. Neurosci.* **12**, 1333–1342 (2009).

12. Ball, T., Schulze-Bonhage, A., Aertsen, A. & Mehring, C. Differential representation of arm movement direction in relation to cortical anatomy and function. *J. Neural Eng.* **6**, 016006 (2009).

13. Fried, I., Mukamel, R. & Kreiman, G. Internally generated preactivation of single neurons in human medial frontal cortex predicts volition. *Neuron* **69**, 548–562 (2011).

14. Schalk, G., Wolpaw, J. R., McFarland, D. J., & Pfurtscheller, G. EEG-based communication: Presence of an error potential. *Clin Neurophysiol*, **111**(12), 2138-2144 (2000).

15. Chavarriaga, R., Sobolewski, A. & Millán, J.d.R. Errare machinale est: The use of error-related potentials in brain-machine interfaces. *Front. Neurosci.* **8**, 208 (2014).

16. Cavanagh, J. F. & Frank, M. J. Frontal theta as a mechanism for cognitive control. *Trends*

Cogn. Sci. **18**, 414-421 (2014).

17. Falkenstein, M., Hoormann, J., Christ, S., Hohnsbein, J. ERP components on reaction errors and their functional significance: A tutorial. *Biol. Psychol.* **51**, 87–107 (2000).

18. Ullsperger, M., Fischer, A. G., Nigbur, R. & Endrass T. Neural mechanisms and temporal dynamics of performance monitoring. *Trends Cogn. Sci.* **18**, 259–267 (2014).

19. van Schie, H. T., Mars, R. B., Coles, M. G. H. & Bekkering, H. Modulation of activity in medial frontal and motor cortices during error observation. *Nat. Neurosci.* **7**, 549–554 (2004).

20. Ferrez, P. W. & Millán, J.d.R. Error-related EEG potentials generated during simulated brain-computer interaction. *IEEE Trans. Biomed. Eng.* **55**, 923–929 (2008).

21. Chavarriaga, R. & Millán, J.d.R. Learning from EEG error-related potentials in noninvasive brain-computer interfaces. *IEEE Trans. Neural Syst. Rehabil. Eng.* **18**, 381–388 (2010).

22. Mahmoudi, B. & Sanchez, J. C. A symbiotic brain-machine interface through value-based decision making. *PloS One* **6**, e14760 (2011).

23. Milekovic, T., Ball, T., Schulze-Bonhage, A., Aertsen, A. & Mehring C. Error-related electrocorticographic activity in humans during continuous movements. *J. Neural Eng.* **9**, 026007 (2012).

24. Ferrez, P. W. & Millán, J.d.R. Simultaneous real-time detection of motor imagery and error-related potentials for improved BCI accuracy. *Proc. 4th Int. BCI Workshop & Training Course,* Graz (Austria), 197–202. Graz: Verlag der TU Graz (2008, September).

25. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: An Introduction* (MIT Press, 1998).

26. Iturrate, I., Montesano, L. & Minguez J. Single trial recognition of error-related potentials during observation of robot operation. *Proc. 32nd Annual Int. Conf. IEEE Eng. Med. Biol. Soc.*, Buenos Aires (Argentina), 4181–4184, doi: 10.1109/IEMBS.2010.5627380 (2010, August).

27. Iturrate, I., Chavarriaga, R., Montesano, L., Minguez, J. & Millán, J.d.R. Latency correction of event-related potentials between different experimental protocols. *J. Neural Eng.* **11**, 036005 (2014).

28. Iturrate, I., Chavarriaga, R., Montesano, L., Minguez, J. & Millán, J.d.R. Latency correction of error-related potentials reduces BCI calibration time. *6th Brain-Computer Interface Conference 2014,* Graz (Austria), doi:10.3217/978-3-85125-378-8-64 (2014, September).

29. Brázdil, M. et al. Error processing— evidence from intracerebral ERP recordings. *Exp. Brain Res.* **146**, 460–466 (2002).

30. Spüler, M. et al. Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a P300-BCI. *Clin. Neurophysiol.* **123**, 1328–1337 (2012).

31. Orsborn, A. L., Dangi, S., Moorman, H. G. & Carmena, J. M. Closed-loop decoder adaptation

on intermediate time-scales facilitates rapid BMI performance improvements independent of decoder initialization conditions. *IEEE Trans. Neural Syst. Rehabil. Eng.* **20**, 468–477 (2012).

32. Gilja, V. et al. A high-performance neural prosthesis enabled by control algorithm design. *Nat. Neurosci.* **15**, 1752–1757 (2012).

33. Gürel, T. & Mehring, C. Unsupervised adaptation of brain-machine interface decoders. *Front. Neurosci.* **6**, 164 (2012).

34. Wiering, M. & van Otterlo, M. *Reinforcement Learning: State of the Art* (Springer, 2012).

35. DiGiovanna, J., Mahmoudi, B., Fortes, J., Principe, J. C. & Sanchez, J. C. Coadaptive brain-machine interface via reinforcement learning. *IEEE Trans. Biomed. Eng.* **56**, 54–64 (2009).

36. Walter, W. G., Cooper, R., Aldridge, V. J., McCallum, W. C. & Winter, A. L. Contingent negative variation: An electric sign of sensorimotor association and expectancy in the human brain. *Nature* **203**, 380–384 (1964).

37. Garipelli, G., Chavarriaga, R. & Millán, J.d.R. Single trial analysis of slow cortical potentials: A study on anticipation related potentials. *J. Neural Eng.* **10**, 036014 (2013).

38. Iturrate, I., Montesano, L. & Minguez, J. Shared-control brain-computer interface for a two dimensional reaching task using EEG error-related potentials. *Proc. 35th Annual Int. Conf. IEEE Eng. Med. Biol. Soc.*, Osaka (Japan), 5258–5262, doi: 10.1109/EMBC.2013.6610735 (2013, June).

39. Iturrate, I., Montesano, L., Chavarriaga, R., Millán, J.d.R & Minguez, J. Spatiotemporal filtering for EEG error related potentials. Proc. *5th Int Brain-Computer Interface Conf.*, Graz (Austria), 12–15. Graz: Graz: Verlag der TU Graz (2011, September).

40. Waldert, S. et al. Hand movement direction decoded from MEG and EEG. *J. Neurosci.* **28**, 1000–1008 (2008).

41. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.* **57**, 289–300 (1995).

Acknowledgments: Funding: This work has been partially supported by Spanish Ministry of Science projects HYPER-CSD2009-00067 (J.M.), DPI2009-14732-C02-01 (I.I. and J.M.) and DPI2011-25892 (L.M.), CAI Programa Europa (I.I.), DGA-FSE, grupo T04 (I.I., L.M. and J.M.), European ICT Programme Project FP7-224631 (J.d.R.M.), and Swiss NCCR “Robotics” (R.C. and J.d.R.M.).

Author contributions: J.d.R.M., J.M., R.C. and L.M. were responsible for the study conception; I.I. implemented and executed the experiments; all authors contributed to the methodology, data analysis, and manuscript preparation.

Competing financial interests: Authors declare that they have no financial interests that could be perceived as being a conflict of interest.

Figures:

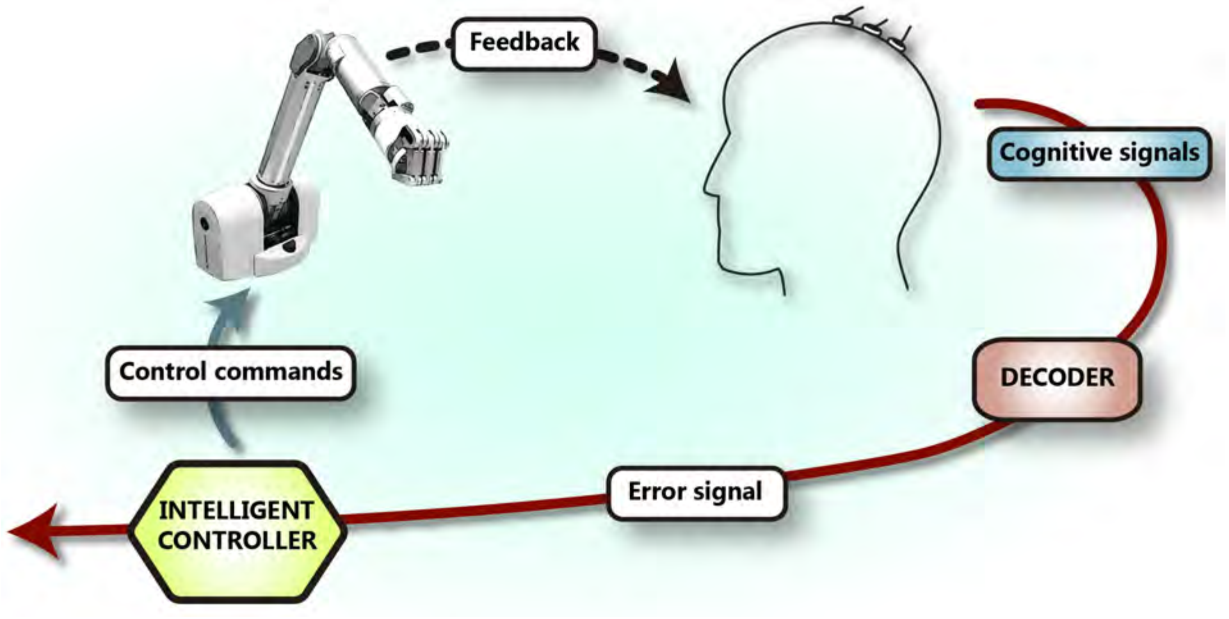

Fig. 1. **Teaching BMI paradigm.** In contrast with the standard control approach, in this paradigm users assess the actions performed by the neuroprosthesis as erroneous or correct. This information is decoded from the user's brain signals, and exploited by the reinforcement learning algorithm embedded in the neuroprosthesis controller to learn appropriate motor behaviours (or control policies) to perform different reaching tasks. See also Supplementary movie.

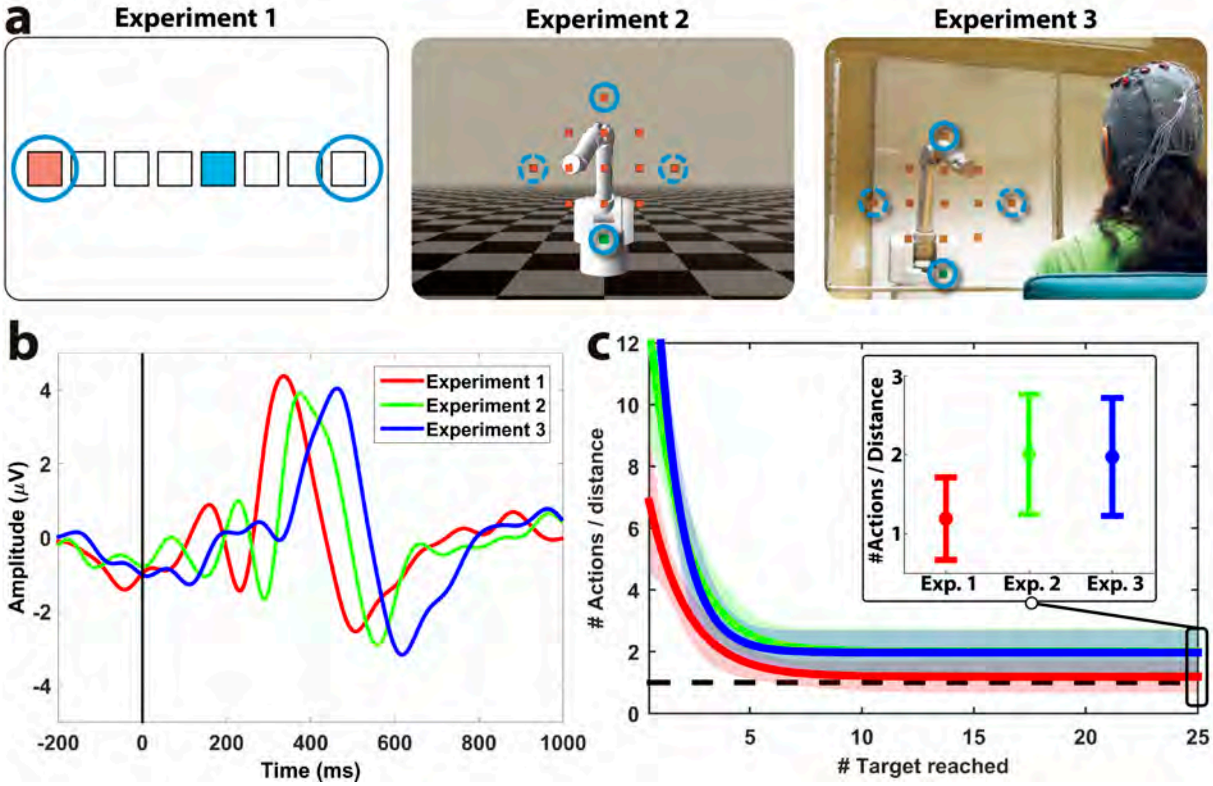

Fig. 2. Learning optimal behaviours from error-related brain activity. [A] Experimental setup. In experiment 1, the device (blue square) can move one position to the left or to the right in order to reach the target (red square). In experiments 2 and 3, the robot moves left, right, up, or down across 13 states (orange squares) to reach a target (green square). Solid and dashed circles denote practice and new targets, respectively. [B] Grand-average difference event-related potentials (ERP) for each experiment during the training phase at channel FCz (N = 12); t=0 ms represents the action onset. This difference ERP is computed as the difference of the subjects' evoked EEG response after erroneous and correct actions of the device. [C] Normalized number of actions needed to reach the targets within a run. Lines correspond to the fitting of an exponential function to the data of each experiment, with the 95% confidence interval shown as shadows of the fitting line (all subjects combined). The horizontal line (Y =1) indicates the optimal performance (See Methods). The inset shows the mean (± the 95% confidence) convergence value of the curve for each experiment.

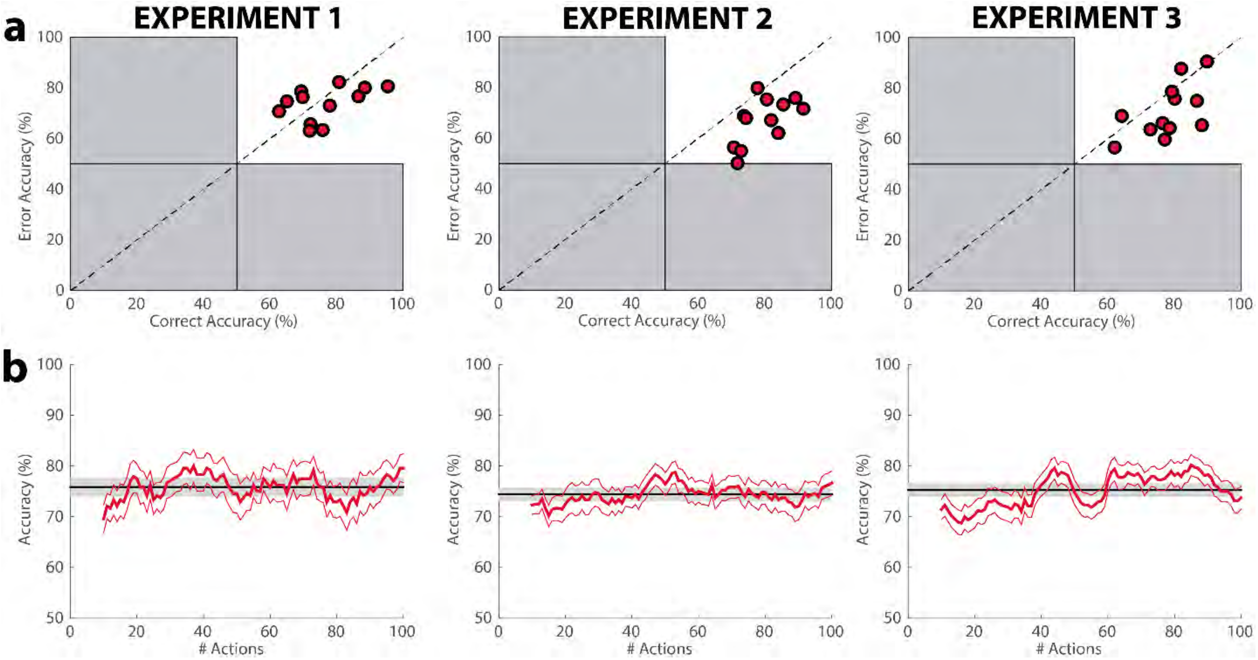

Fig. 3. Online classification accuracy. (a) For each subject and experiment, ErrP online classification accuracy. The x-axis and y-axis represent the correct and error accuracies, respectively. Each dot corresponds to the average online accuracy achieved by each subject. (b) Decoding performance throughout the RL execution. For each experiment, decoding performance (mean ± SEM, thick ± thin red lines) throughout one run, where x-axis represents the actions along the run. The performance is computed as the accuracy obtained in a sliding window of 10 actions. The black horizontal line indicates the accuracy for each experiment, with the SEM shadowed. Results are averaged across runs (2 and 4 for experiments 1 and, 2-3 respectively), and across subjects (N=12). Confidence intervals (α=0.05) for the accuracies throughout the run were of [71.30, 80.47], [70.93, 77.84], [68.67, 81.52] for experiments 1 to 3, respectively. The results showed no substantial differences in the accuracy variability throughout the runs.

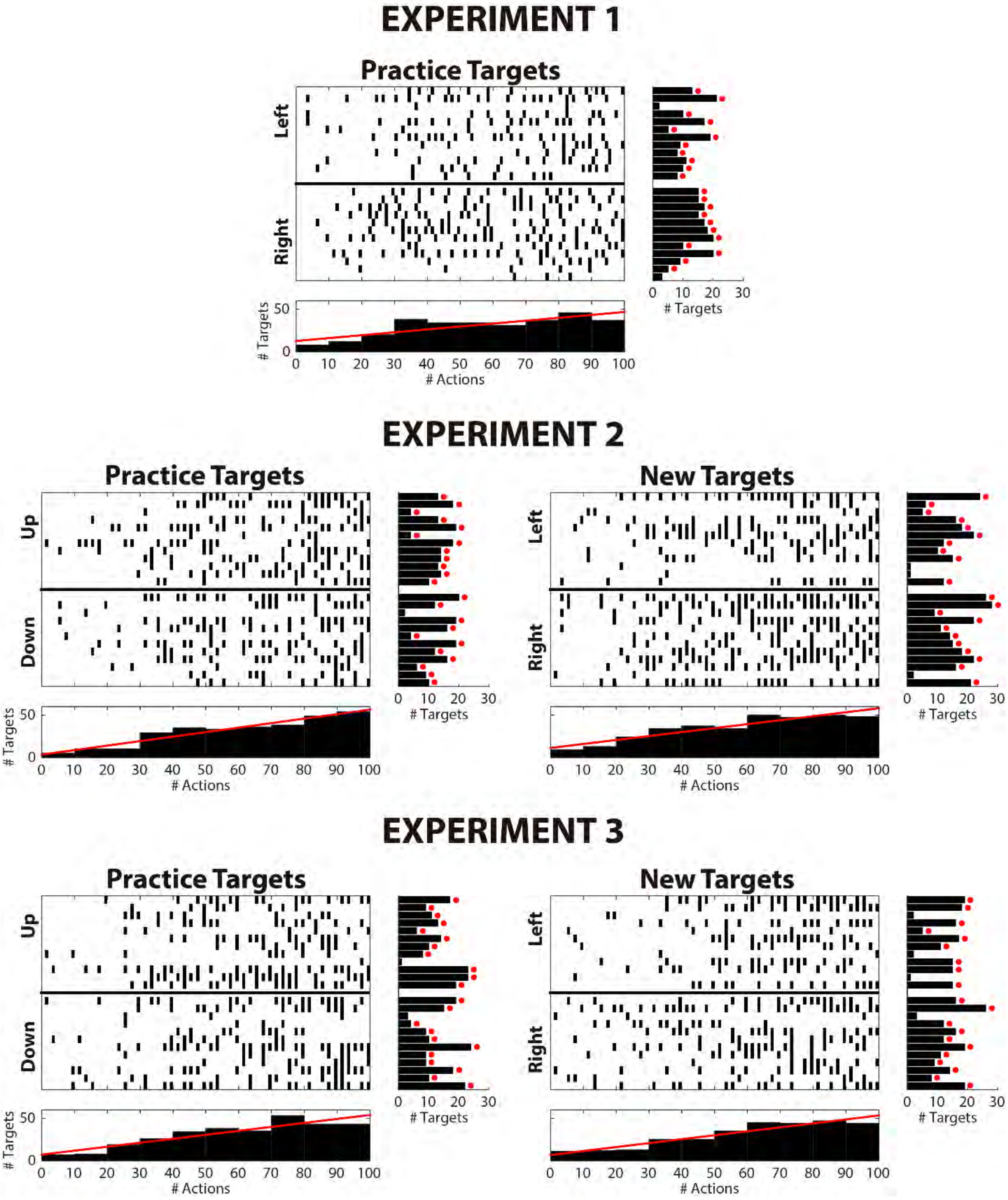

Fig. 4. Number of reached targets within a run. Each subfigure corresponds to either practice or new targets for each experiment. For each subfigure: (Top) Raster Plot, where x-axis represents time within a run (from 1 to 100 actions performed by the device), and each

row in the y-axis represents one of the 12 subjects for each of the two targets. Every tick corresponds to the moment a target is reached. (Bottom) Histogram (bin size of 10 actions) associated with the upper raster plot. Each bar represents the number of times a target was reached for the corresponding bin. Additionally, the trend line is also plotted in red. (Right) Histogram for each subject and target with the number of times a target was reached. Red dots indicate above chance results (confidence 95%). Results show how the time required for reaching the targets decreases as the run goes. Behaviour is similar across subjects, experiments and targets.

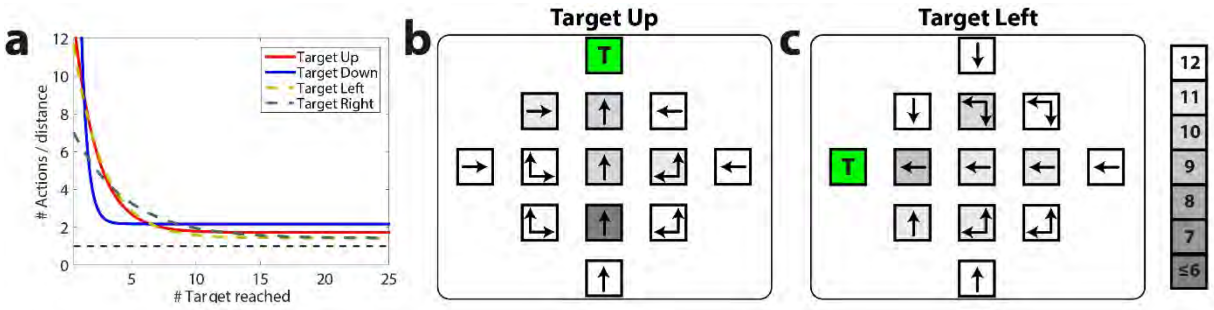

Fig. 5. Experiment 3, comparing performance between practice targets (Up and Down), and new targets (Left and Right). [A] Normalized number of actions required to reach each target as in Fig. 2. [B] Optimal actions per state (denoted by arrows) for a practice target (Up), and [C] for a new target (Left). The green square marks the target location. The number of subjects for which the action was correctly learned is color encoded in gray levels.

SUPPLEMENTARY INFORMATION

Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control

Authors: Iñaki Iturrate[1,2], Ricardo Chavarriaga[2], Luis Montesano[1], Javier Minguez[1], and José del R. Millán[2*]

Affiliations:

[1]Instituto de Investigación en Ingeniería de Aragón, Dpto. de Informática e Ingeniería de Sistemas, Universidad de Zaragoza, Spain

[2]Chair in Non-Invasive Brain-Machine Interface, Center for Neuroprosthetics & Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Supplementary Information

Methods

Analysis of error-related potentials waveforms

We analyzed variations in the grand average EEG potentials for both conditions, i.e., erroneous and correct device actions (see Fig. 2B and Supplementary Fig. S1, left panel). To this end we performed a statistical analysis on the difference ERP (error minus correct condition) for all electrodes in the time window [-200, 1000] ms, t =0 ms being the instant when the device starts to move. Only signals from the training runs —having a constant error-rate (20%)— are used in this analysis. These runs yielded between 200 and 400 trials for each subject.

A 3 (brain area: frontal, central or centro-parietal electrode locations) x 3 (left, midline or right locations) x 3 (experiments) within-subjects ANOVA was performed on the peak amplitudes and latencies of the difference average. Each group is summarized in Supplementary Table 1. When needed, the Geisser-Greenhouse correction was applied to assure sphericity. Pairwise post-hoc tests with the Bonferroni correction were computed to determine the differences between pairs of experiments.

We mainly found significant effects on the latency but not the amplitude of the difference potential. Latency increased with the complexity of the experiments. The type of experiment significantly affected the latencies of both the positive ($F_{2,22}$ = 41.594, p =3×10^{-8}) and negative peaks of the difference ERP ($F_{2,22}$ =7.522, p =0.003). The brain area also affected the latencies of the positive peak ($F_{1.32,14.55}$ = 14.175, p =0.001) but not the negative one $F_{2,22}$ =0.911, p =0.417). Similarly, the hemisphere affected the latency of the positive peak ($F_{2,22}$ =5.279, p = 0.013), but not the latency of the negative one ($F_{2,22}$ =1.711, p =0.204). No significant interactions were found.

For the positive peak latencies, post-hoc pairwise tests revealed significant differences between experiments 1 and 2 (p =0.0001), and between experiments 1 and 3 (p =0.0001), and close to significant difference between experiments 2 and 3 (p =0.068). For the negative peak latencies, there were significant differences between experiments 1 and 3 (p =0.009), but not between experiments 1 and 2 (p =0.357) or experiments 2 and 3 (p =0.122).

In contrast, the amplitudes of the positive and negative peaks were not significantly affected by the experiment ($F_{2,22}$ =0.124, p =0.884 and $F_{2,22}$ =2.304, p =0.123, respectively) nor the brain area ($F_{1.19,13.08}$ = 1.227, p =0.737 and $F_{1.32,14.47}$ =0.071, p =0.857). The laterality significantly affected the positive peak amplitude ($F_{2,22}$ =4.556, p =0.022), and was close to significance for the negative peak ($F_{2,22}$ =3.425, p =0.051). As in the case of the peak latencies, no significant interactions were found.

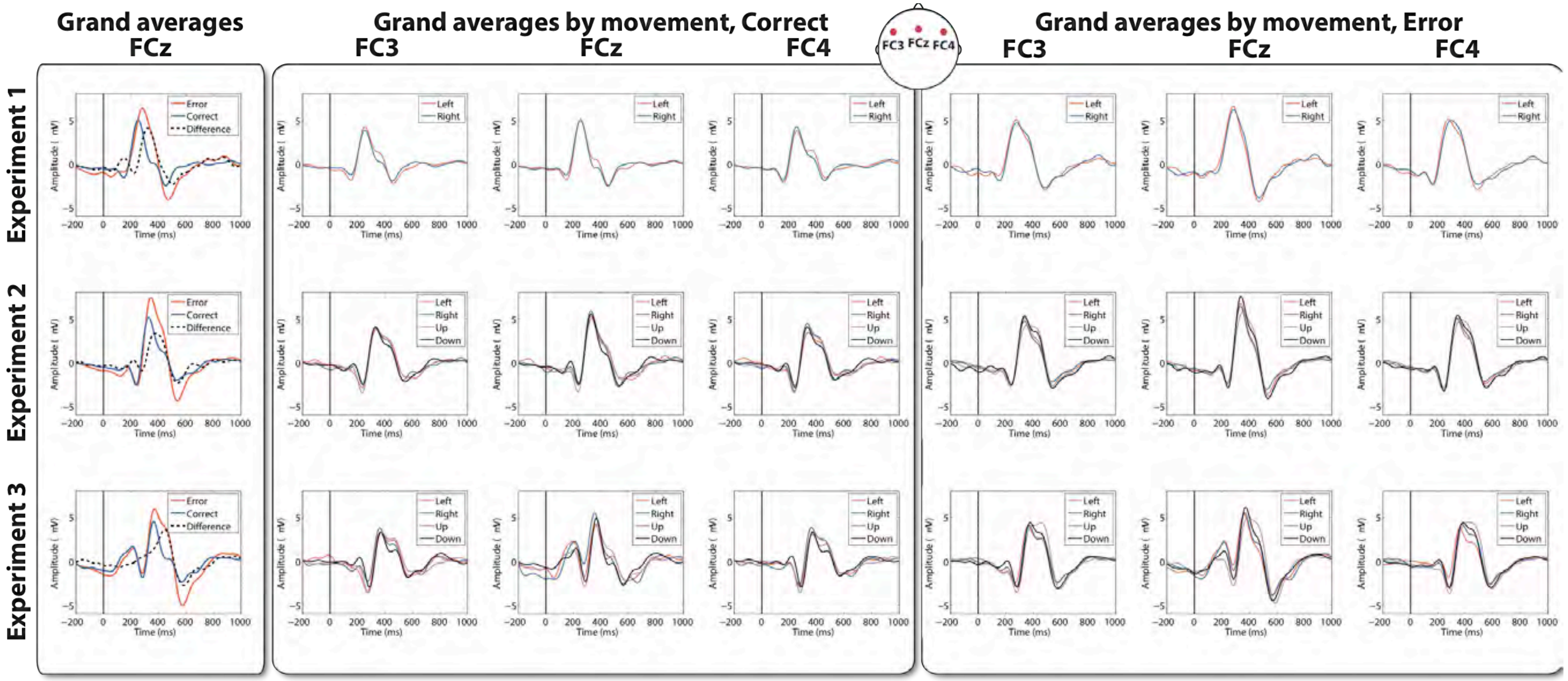

Fig. S1. Grand averaged ERPs. Grand averaged ERPs in channels FC3, FCz, and FC4 over all subjects (N=12). Each row corresponds to a different experiment. Left panel displays the grand averages at FCz of correct (blue), error (red), and difference (black). Center and right panels illustrate the grand averages for correct and erroneous assessments of each movement direction (left, right, up or down), respectively. These two panels show that the averaged signals are very similar across all directions, reducing the possibility that the direction of robot movements may have a systematic influence on the ErrP classification process.

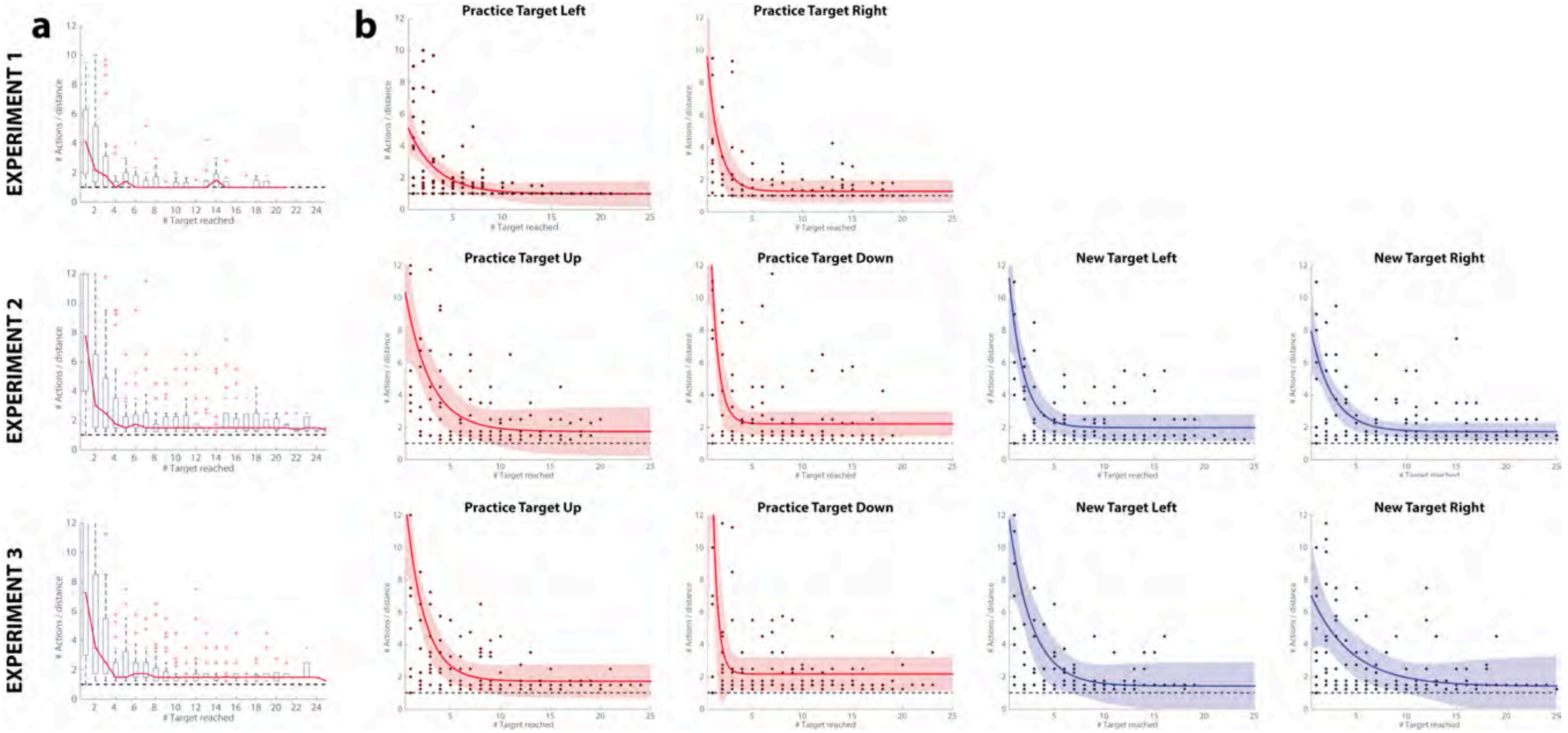

Fig. S2. Acquisition of control policies. (a) Acquisition of control policies for all the targets, and each experiment separately, as a bar plot each time a target was reached. Additionally, the median line is also shown in red. (b) Acquisition of control for each experiment and target. Dots represent the normalized number of actions required to reach a target location during an online run (all subjects together). The continuous line corresponds to an exponential fitting ($y = a + be^{-cx}$) of the data. The shadowed area corresponds to the 95% confidence interval of the fitting. A consistent decrease in the number of actions required to reach the goal is observed for all targets. Moreover, no difference is observed in experiments 2 and 3 between practice (up and down, red) and new targets (left and right, blue).

Table S1. **Electrode locations used as factors for the ERP statistical analysis**. Each row corresponds to the brain areas, while columns correspond to the laterality.

	Left	**Midline**	**Right**
Frontal	FC3, FC1	Fz,FCz	FC2,FC4
Central	C3,C1	Cz	C2,C4
Centro-parietal	CP3,CP1	CPz	CP2,CP4

Movie S1. **Demonstration of the teaching BMI paradigm.** The movie demonstrates the online operation of our teaching BMI paradigm with a real arm robot executing a reaching task. The user monitors the performance of the robot arm that has no knowledge about the target location (green square). Error-related potentials (ErrPs) are elicited whenever an executed action does not match the user's expectations. These ErrPs are decoded online as shown in the bottom-left part. The outcome of this decoding is used as a reward signal for a reinforcement learning algorithm that updates the control policy. This policy is shown in the upper-right panel, where arrows show the estimated optimal action for each state. As observed, acquisition of (quasi-)optimal actions is very fast. Trajectories are initially random and progressively become straight throughout the run (100 actions).