2023 IEEE International Symposium on Mixed and Augmented Reality (ISMAR) | 979-8-3503-2838-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISMAR59233.2023.00113

EEG-Based Error Detection Can Challenge Human Reaction Time in a VR Navigation Task

Michael Wimmer* Know-Center GmbH
Nicole Weidinger† Know-Center GmbH
Neven ElSayed‡ Know-Center GmbH
Gernot R. Müller-Putz§ Graz University of Technology
Eduardo Veas¶ Graz University of Technology

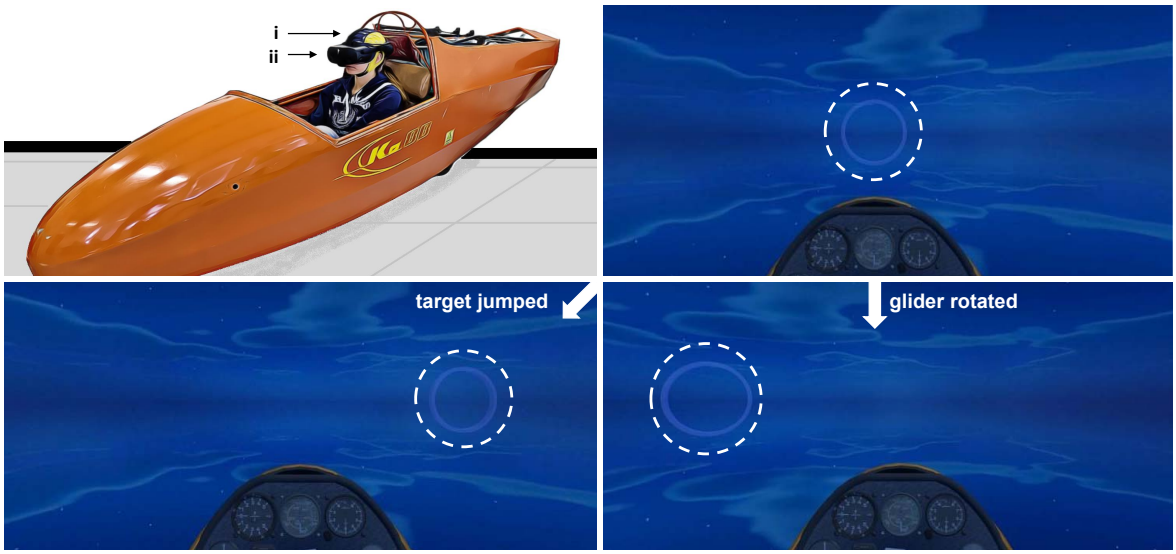

Figure 1: Experimental setup. Top-left: Participant in the glider wearing an EEG cap (i) and the HMD (ii). Top-right: Virtual glider approaching a target. Bottom-left: *Target* error: The target jumped to the right. Bottom-right: *Passive* error. The glider rotated to the right. Mind that in the *passive* and *active* conditions, the participant's whole field of view changed, whereas in the *target* condition, only the target was relocated. The dashed lines were added for illustration purposes only, they were not visible to the participants.

Abstract

Error perception is known to elicit distinct brain patterns, which can be used to improve the usability of systems facilitating human-computer interactions, such as brain-computer interfaces. This requires a high-accuracy detection of erroneous events, e.g., misinterpretations of the user's intention by the interface, to allow for suitable reactions of the system. In this work, we concentrate on steering-based navigation tasks. We present a combined electroencephalography-virtual reality (VR) study investigating different approaches for error detection and simultaneously exploring the corrective human behavior to erroneous events in a VR flight simulation. We could classify different errors allowing us to analyze neural signatures of unexpected changes in the VR. Moreover, the presented models could detect errors faster than participants naturally responded to them. This work could contribute to developing adaptive VR applications that exclusively rely on the user's physiological information.

*e-mail: mwimmer@know-center.at
†e-mail: nweidinger@know-center.at
‡e-mail: nelsayed@know-center.at
§e-mail: gernot.mueller@tugraz.at
¶e-mail: eveas@tugraz.at

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual Reality; Human-centered computing—Human computer interaction (HCI)—HCI design and evaluation methods—User models

1 Introduction

Human-computer interactions (HCIs) are prone to misinterpretations of the user's intentions by the system's interface. These errors in the interaction with a system can lead to a negatively perceived user experience. Methods aiming to improve the user experience in the presence of such errors could include real-time error detection, which would subsequently enable the system to take corrective actions. In this work, we present an approach to detect erroneous actions based on electroencephalographic (EEG) signals and additionally analyze the participant's corrective behavior in comparison.

Multiple studies investigated the physiological correlates of error processing, most of them using EEG signals, i.e., electrical brain signals recorded from the scalp [18, 23]. It has been shown that committing or observing errors elicits a change in the amplitude of the EEG signals. This event-related activity is commonly referred to as error-related potential (ErrP). ErrPs have already been used to detect errors online, i.e., in real-time, in non-immersive scenarios [22, 34] and realistic tasks [35, 37, 77].

Only a few studies explored error processing in virtual reality (VR). Previous works focused particularly on two areas, i.e., errors elicited by (i) violations of agency [51, 58] and (ii) unrealistic or erroneous interactions with the VR, such as tracking errors [23,

DOI 10.1109/ISMAR59233.2023.00113

63, 64]. Error detection was performed in some of these works [23, 63]. However, they did not attempt to find the time point when errors occurred. Nonetheless, this is necessary for the system to be reactive in the event of an error. Therefore, errors need to be detected asynchronously, i.e., independent of the error event [41].

Multiple studies analyzed the utility of incorporating the decoding of error processing in brain-computer interfaces (BCIs) [59, 74]. In essence, there are two existing approaches to utilize ErrPs in BCIs, i.e, (i) to correct erroneous actions [61], and (ii) to adapt the error detection model such that the model's classification accuracy improves over time [34]. However, none of these approaches have been studied in the context of VR. Possible goals for combining ErrPs and VR include a dynamic adaption of the visualization or interaction as a result of detected error events, or by providing visual aids for the user [46]. Further, error detection based on "fast" brain responses might even create a gain in time compared to the "slow" physical reactions to them. This could be of practical use, depending on the application, or simply foster a smoother interaction with the VR. Following these considerations, we formulated two research questions this work mainly addresses:

- **RQ1:** Can we estimate the time point of an erroneous event in VR using EEG data?
- **RQ2:** Can we detect erroneous events faster than participants would naturally react to them?

To answer these research questions, we performed asynchronous error detection in a VR navigation task. We presented different types of errors to the participants, allowing us to study the reaction to sudden changes in the environment. We designed three different error classification models that are trained following a generic approach, a personalized approach, and a combination of both. Personalized models usually require a calibration phase in which EEG signals are analyzed to train the classifiers before they can be used in BCIs. Alternatively, generic models use signals of other users, which allows for immediate application [36]. There is evidence that generic classifiers offer reasonable performance, which is, however, in general below the performance of personalized models [30, 35]. We hypothesize to see a similar effect in the results obtained with our models. The current work could contribute to the design of adaptive and corrective systems in VR applications, relying on the analysis of the users' physiological information.

2 Related Work

Virtual navigation techniques allow users to explore environments at various scales [42]. We concentrate on techniques that provide continuous control of direction, so-called steering-based locomotion techniques, incorporating physical steering props [6, 32]. These techniques are common in 3D first-person games [7, 32, 52], but also when it comes to controlling vehicles that are modulated by the simulation of the vehicle's physical properties [5, 62]. Interested readers are referred to Al Zayer and colleagues [76], who provided a survey on virtual locomotion.

In the present study, we investigated the neurophysiological response to error perception while navigating a virtual vehicle. Therefore, the following sections provide a brief overview of the research on error perception, its neurophysiological correlates, and its previous application in BCIs and VR.

2.1 Background on ErrPs

EEG is a non-invasive technique to study electrical brain activity with electrodes attached to the scalp. First experiments to study error-related brain activity date back to the beginning of the 1990s [18, 23]. Studies found a distinct event-related potential (ERP) in the EEG signals after participants committed errors in speed response choice tasks, commonly referred to as ErrP. ErrPs are typically found as a result of identifying erroneous actions, e.g., after participants committed errors. ErrPs are often characterized by an initial negative peak, i.e., the error-related negativity (ERN), followed by a positive peak, i.e., the error positivity (Pe). These components usually appear in the first 500 ms after the error event in the frontocentral and centroparietal regions of the brain [19]. Outcome errors are found when motor actions do not result in the desired goal [45]. Interestingly, similar patterns were found in the absence of motor response, e.g., while observing erroneous actions [60], or due to a sudden, unexpected change in the environment. The last type is referred to as target error and was first described by Dietrichsen and colleagues [17], who studied the influence of target jumps on participants who moved a cursor toward the target. Finally, it has been demonstrated that ErrPs can be elicited in interactions with agents, i.e., interaction errors. These errors are of special interest when interacting with BCIs and are elicited after a misinterpretation of the user's intention by the interface. Ferrez and Millán [20] studied errors in a simulated interaction with a robot, that could be navigated to one side by manually pressing the corresponding key. In 20% of the occasions, interaction errors were elicited by a movement in the opposite direction of the user's intention.

Notably, it has been shown that ErrPs are influenced by engagement. In an arrowhead flanker task, Hajcak and colleagues [26] motivated participants with monetary compensation for correct responses, i.e., a higher monetary compensation in 50% of the stimuli and a lower compensation in the remaining 50%. The ERN was significantly larger after mistakes in high-value stimuli.

2.2 ErrPs in BCIs

Until today, the main application of ErrPs is as control signals in BCIs [8, 74]. BCIs can use EEG signals to, e.g., decode the user's intentions and generate commands based on them. A misinterpretation of these intentions would lead to erroneous commands. Successful detections of these erroneous commands could enable BCIs to take corrective actions, e.g., stop the execution of the commands or perform counteractions [59, 61]. Error detection on a stimulus-locked, single-trial basis has been performed in numerous experiments with accuracies well above 80% [16, 20, 21, 63]. However, BCI research is developing to give users continuous control of machines. Incorporating ErrPs in such systems requires asynchronous error detection, which has been studied in offline [27, 37, 47–49, 69] and online [35] scenarios. There is evidence that error classifiers work stable over a period of several months [21], or when trained with data from other participants, i.e., generic classification [35, 36]. Even a transfer from able-bodied participants to participants with high spinal cord injury could be presented [35]. Such approaches reduce the time to train classifiers prior to using BCIs and hence improve usability [11].

Another approach to utilize ErrPs in BCIs, besides correcting erroneous commands, is adaptive learning [4]. Leera and colleagues [34] instructed participants to covertly shift their attention to either the left or right side of a screen. Identified misclassifications of the participant's attention were used to adapt the parameters of their binary error classification model. The offline classification accuracy could be significantly improved compared to the static error classification. Chavarriaga and colleagues [10] provided a comprehensive overview of the use of ErrPs in BCIs.

2.3 ErrPs in VR

Most experiments studying ErrPs focus on non-immersive settings [27, 28, 48, 56, 57, 69], some include real-world scenarios, such as driving a car [77] or controlling a robotic arm [35, 37]. However, there is increasing interest in ErrPs in the context of VR. Padrao and colleagues [51] were among the first to investigate neurophysiological signatures of error processing in VR using a head-mounted display (HMD). One of the experiment's objectives was to study the response to violating agency by providing wrong feedback from

the avatar, i.e., sometimes the avatar's hand moved in the opposite direction of the participant's hand [55]. Similarly, Raz and colleagues [58] studied, among others, the responses to unexpected hand bounces from the avatar. Other studies focused on errors in the interaction with the VR [24, 25, 63, 64]. Unrealistic interactions with the VR were investigated by giving visual [64] or vibrotactile [24, 25] feedback when tapping a virtual object, while the feedback was given prematurely on a few occasions. Si-Mohammed and colleagues [63] studied different error types, such as errors in tracking objects, giving wrong feedback after task completion, or anomalies in the environment.

Some studies [53, 54] used different visualization techniques, i.e., the CAVE automatic virtual environment (CAVE) system [12], studying the observation of erroneous actions in first- and third-person perspectives. Likewise, Spinelli and colleagues [68] instructed participants to observe an avatar in a reaching-to-grasp task that could perform correct or erroneous actions with either their left or right hand. A similar experiment was conducted with participants suffering from higher-order motor control disorders [67].

3 Materials and Methods

3.1 Participants

Nineteen volunteers (27.6 ± 2.3 years old, mean ± standard deviation (SD), seven female) participated in the study. All participants had normal or corrected-to-normal vision and were free of any known neurological diseases, 13 self-reported having only little or no experience with VR. After instructions, all participants gave written informed consent to take part in the experiment. The study was conducted according to the Declaration of Helsinki (1975) and approved by the ethical review committee of Graz University of Technology. Participation was honored with vouchers worth 20 euros.

3.2 Apparatus

The VR environment was shown with an HP Reverb G2 Omnicept HMD (HP, CA, USA), which includes a built-in eye tracking module (Tobii AB, Danderyd, Sweden) that records data at a sampling frequency of 120 Hz. The VR and the experimental paradigm were designed in Unity[1].

EEG signals were recorded using 63 actively shielded, gel-based Ag/AgCl electrodes (eego^{TM}sports, ANT Neuro, Hengelo, The Netherlands). The electrodes were positioned according to the 10-5 international system [50] at Fp1, Fp2, AF3, AFz, AF4, F7, F5, F3, F1, Fz, F2, F4, F6, F8, FT7, FC5, FC3, FC1, FCz, FC2, FC4, FC6, FT8, FFC3h, FCC1h, FCC2h, FCC4h, T7, C5, C3, C1, Cz, C2, C4, C6, T8, CCP3h, CCP1h, CCP2h, CCP4h, TP7, CP5, CP3, CP1, CP2, CP4, CP6, TP8, P7, P5, P3, P1, Pz, P2, P4, P6, P8, PO3, POz, PO4, O1, Oz, and O2. The ground and reference electrodes were placed at AFz and CPz, respectively. The signals were sampled at 512 Hz.

For the acquisition and synchronization of multiple data streams (EEG, gaze direction, controller movement, events) we utilized the lab streaming layer protocol[2].

We recorded further physiological data, i.e., electrocardiogram and pupil size [72, 73], which are not part of the analysis presented in this work and hence disregarded in the following deliberations.

3.3 Experimental procedure

Prior to the recordings, the 63 EEG electrodes were mounted such that the impedances between electrodes and scalp were below 10 kΩ to facilitate high signal quality. Additionally, the EEG signals were visually checked right before the recordings and monitored throughout the experiment.

Over the course of the experiment, participants sat comfortably in an immobile glider (Ka 8B, Alexander Schleicher, Germany), as shown in Fig. 1. The experiment consisted of two blocks, both consisting of three phases, i.e., (i) eye-tracking calibration of the HMD, (ii) acquisition of eye movement data for artifact removal (see Sect. 3.3.1), and (iii) four (first block) or five (second block) flight simulation runs (in total nine runs, see Sect. 3.3.2). Between each run, participants took breaks of one to five minutes without removing the HMD. The HMD was only removed for a break of approximately ten minutes between the two blocks. Removing the HMD only once reduced the possibility of electrode displacements or loss of contact with the scalp. However, we checked the impedances again before the second block. Participants completed up to two runs before the first block to familiarize themselves with the flight simulation.

To reduce electromyogenic artifacts in the EEG signals [44], participants were instructed to avoid non-task-related movements, e.g., head movements, during the recordings.

3.3.1 Acquisition of eye movement data

Eye movements and blinks are common sources of noise present in most EEG recordings [29, 70]. To correct eye movements and blinks, we used the sparse generalized eye artifact subspace subtraction algorithm (SGEYESUB), for which participants repetitively performed a series of eye-related artifacts. We refer to the original work from Kobler and colleagues [31] for a detailed description of the algorithm. To use the SGEYESUB with an HMD, we redesigned their proposed paradigm in Unity. Eye movement data were recorded for approximately five minutes in each block.

3.3.2 Flight simulation

The participants were instructed to navigate a virtual glider through targets, i.e., rings (see Fig. 1). As the glider moved at a constant speed by itself, participants could navigate (to the left, right, up, and down) using the control stick of the physical glider, to which one HMD controller was attached, as shown in Fig. 2. This should enhance the realism of the simulation.

Figure 2: Participant using the control stick of the glider, to which an HMD controller was attached.

Each flight simulation run consisted of 70 targets, the first one always appeared straight ahead of the glider. As the glider passed a target, the next one appeared either straight ahead of it or displaced at a fixed angle horizontally or vertically. Only one target was visible at a time. Before 30% of the targets, an error event was triggered. Error events were evenly distributed among the following three conditions (seven errors per condition per run):

- ***target:*** The target unexpectedly jumped to its left or right, about 1.6 s before the glider would have reached it.
- ***passive:*** *Passive* errors were only triggered before targets that were straight ahead of the glider. With such targets, there was no need for active steering to reach the target, hence participants were in a passive state. In the *passive* condition, a torque

[1] https://unity.com/

[2] https://github.com/sccn/labstreaminglayer

was applied to the virtual glider, starting between 1.1 s and 1.8 s before reaching the next target. The torque led to a sudden and continuous rotation of the glider for 0.6 s.

- ***active:*** Like the *passive* error, but only triggered before targets with horizontal or vertical displacement. Reaching these targets required active navigation, hence participants were considered to be in an active state.

Participants were free to decide whether they wanted to correct for errors, e.g., try to reach suddenly displaced targets, or not. This was essential to investigate the natural reaction behavior of participants. Ten participants showed an apparent reaction to the error events, nine participants did not try to reach the targets after errors were triggered.

To prevent adjustments to the paradigm, the positioning of the targets and the timing of the error events were randomized. The average duration of one run was four minutes and 45 seconds, hence a new target appeared approximately every four seconds.

3.4 Signal processing

Behavioral and physiological data were processed and analyzed offline in Matlab R2022a (The MathWorks, MA, USA) utilizing the EEGLAB toolbox (v2022.0) [14].

3.4.1 Behavioral data processing

To study the participants' natural reactions to the errors, we recorded data related to the controller movement. Since target jumps (*target*) and glider rotations (*passive*, *active*) exclusively happened along the horizontal axis (e.g., targets never jumped upwards), corrective movements of the participants were only required in the horizontal direction, too. Hence, we only considered horizontal controller movements for the behavioral analysis.

The controller displacement data were recorded once per frame. First, we resampled the data from the horizontal displacement to 64 Hz to reduce computational effort. We then segmented the data into epochs of 1.5 s using a window of [-0.5 1] s relative to the error onsets. Then, we baseline-corrected the epochs by subtracting the mean value of the [-0.5 0] s window prior to the error onset. We only considered the absolute controller displacement values, and thus we disregarded information on whether a controller was moved to the left or right direction.

Next, to detect movement onsets, we defined an individual threshold for each participant and condition. Therefore, we calculated the first derivative of the controller displacement. From this data, we defined a vector b containing the maximum values of each trial in the baseline window and calculated the threshold as $mean(b)+2\cdot SD(b)$. The reaction time for each trial was defined as the time the first derivative of the displacement exceeded the threshold the first time. We only considered exceedances within a window of [0.2 0.7] s after the error onset to exclude non-error-related controller movements.

Ten of the participants showed corrective behavior to errors, i.e., a controller movement after error events. On average, these ten participants reacted in 78% of the error events.

3.4.2 EEG data processing

As a first step, we filtered the recorded EEG signals between 0.4 Hz and 30 Hz and suppressed power line noise with a notch filter at 50 Hz and 100 Hz, respectively (all filters were Butterworth, fourth order, zero-phase). Artifacts stemming from eye movements and blinks were removed using the SGEYSUB (see Sect. 3.3.2) [31]. After, we applied another bandpass filter between 1 Hz and 10 Hz (Butterworth, fourth order, zero-phase) since numerous studies found the lower frequency bands to have an essential role in error processing [39, 71, 75]. Subsequently, we resampled the signals to 64 Hz and removed the most frontal channels (Fp, AF) to minimize residual eye-related activity in the data. Another approach to dealing with eye-related artifacts is performing independent component analysis (ICA) [40]. This comes, however, with the disadvantage to be applicable only for offline data analysis. Since we wanted to provide tools for real-time error detection, we decided to use SGEYESUB, as comparable studies have before [35, 37].

Next, we segmented the signals of the remaining 58 channels into trials of 1.5 s. Trials belonging to one of the three error conditions were extracted [-0.5 1] s relative to the error onset. Correct trials were extracted [1.5 3] s after passing a target, no error was triggered in this window. We rejected artifact-contaminated trials based on amplitude threshold (exceeding ± 35 µV), kurtosis, and abnormal joint probability (threshold of $5\cdot SD$ for the last two) [15]. Further, we rejected trials based on visual inspection. Bad channels were identified based on visual inspection and variance. For the last, we calculated the first and third quartile ($Q1$, $Q3$) and the interquartile range ($IQR = Q3 - Q1$) of the channel variances. We interpolated channels with a variance greater than $Q3 + 1.5\cdot IQR$. We rejected on average 12% of the trials and interpolated 1.5 channels per participant. Finally, we re-referenced the signals to the common average [43].

3.4.3 Error detection

To compare the human reaction time to the detection time of our error detection models, we asynchronously classified errors based on a sliding window approach. We tested three different classification models, i.e., (i) a fully generic classification, (ii) a generic classification with personalized hyperparameters, and (iii) a personalized classification, each of them having two classes, i.e, *target* vs. *correct*, *passive* vs. *correct*, and *active* vs. *correct*. Models are described in detail in the following sections. All models are only evaluated on the ten participants that showed corrective behavior to errors. For all models, we used the true positive rate (TPR), the true negative rate (TNR), and the detection time as evaluation metrics. These metrics are described at the end of this section.

Fully generic error detection For each participant, we trained ErrP classifiers using the preprocessed EEG data from the remaining 18 participants (leave-one-out). As features for training the classifiers, we used all amplitude values from all 58 channels within a window of [0.14 0.4] s after error onset. For correct trials, we used the window [2.14 2.4] s after passing the previous target. In this segment, most errors were triggered in error trials. To reduce the number of features, we performed principal component analysis (PCA) and kept the components that explained 99% of the data variability as features, as suggested by other works [35, 37]. We used these features to train a shrinkage linear discriminant analysis (sLDA) classifier, which became a standard method in classification problems with ERPs [3, 38]. For each classifier, we randomly chose a subset of correct trials to maintain the ratio of 30% error trials to 70% correct trials from the initial experimental design (Sect. 3.3.2).

We tested each classifier asynchronously on the data of the participant which were not used for training. Therefore, we slid a 260 ms window with a leap of 16 ms through each trial. We used a softmax function to transform the linear scores of the classifier to a probability p_{err}. This probability denotes how likely an error event was triggered during the evaluated window. We consider an error as detected if p_{err} exceeded a certain threshold τ for a minimum of k consecutive windows. For the fully generic classification, we fixed $\tau = 0.5$ and $k = 2$ for all classifiers, similar to previous works [35, 37]. In other words, we considered an error as detected if $p_{err} > 0.5$ for at least two consecutive windows.

Generic error detection with personalized hyperparameters To improve the performance of the classifiers, we optimized the hyperparameters τ and k. The model and computation of p_{err} are like in the fully generic classifier. To optimize the hyperparameters, five-fold cross-validation (CV) was performed ten times on the probabilities p_{err}. To find the most suitable parameters, we varied τ

from 0 to 1 in fixed steps of 0.025 and k from 1 and 6 and evaluated the classifiers' performances for each combination on the training set. We used the optimal pair of parameters, i.e., the one that maximized the product of TPR and TNR, for the test set. For each classifier and participant, we computed the average TPR and TNR of the 50 folds.

Personalized error detection To evaluate the performance of the personalized classifiers, we performed ten times five-fold CVs on the data of one participant only. As training features, like in the generic classifiers, we used all amplitude values from the same windows, maintaining a 30% to 70% ratio for error trials to correct trials. Again, we performed PCA for feature dimensionality reduction. In each fold, we obtained a p_{err} for each window in each trial tested. We optimized the hyperparameters τ and k as described for the generic classifiers. Again, we calculated the average TPR and TNR for each of the 50 folds.

Evaluation metrics To evaluate the performance of the classifiers, we used TPR (sensitivity), TNR (specificity), and detection times. We considered a correct trial true negative (TN) if no error was detected during the entire trial duration. Subsequently, we computed the TNR as the fraction of correct trials that are TN. Similarly, we defined true positive (TP) trials as error trials in which errors are detected exclusively within a window of [0.1 0.6] s after the error onset. This is to ensure that we only consider error detections that are related to error events. The calculation of TPR is analogous to the TNR. Finally, we defined the detection time as the time of the first error detection in each TP trial.

3.4.4 Statistics

We performed sample-wise statistical tests on the 19 participant averages of the EEG data (see Fig. 4). For that purpose, we conducted Wilcoxon signed-rank tests and applied the false-discovery rate (FDR) procedure to correct for multiple comparisons [1,2]. Further, we performed Wilcoxon rank-sum tests with FDR correction to compare the reaction times between the first and the second block for each error condition using all detected reactions (see Fig. 3). Analogously, we compared the participant's reaction times to the error detection times of each model of the same condition.

4 Results

4.1 Behavioral analysis

The results of the behavioral analysis are shown in Fig. 3. Fig. 3a illustrates the averaged controller displacement for each error condition, i.e., the reactive behavior of the participants as a result of the error events. The average reaction times of the participants were 414 ms (*target*, in red), 427 ms (*passive*, in blue), and 419 ms (*active*, in black) after the error onset. The development of the participants' reaction times is shown in Fig. 3b. Here, the dashed lines are linear regressions of the corresponding mean reaction times for each run and indicate an individual trend for each block. The decline in the passive condition in the first block is mainly driven by the slow average reaction in the first run. However, statistical tests did not reveal a significant change (alpha = 0.05) in the reaction times from the first to the second block, which indicates that the participants had a consistent reaction time throughout the experiment (Fig. 3c). For **RQ2** we studied if we can detect errors faster than the participants would naturally try to correct them. In all conditions, the participants' reaction times were slower than the corresponding error detection time for all models. A comparison can be found in Table 1. Information about the participants' individual reaction times is depicted in Fig. 3d-f.

4.2 Electrophysiology

Fig. 4 presents the grand average EEG response to error processing, i.e., the mean of the participant averages. ERPs are depicted for the positions FCz, Cz, and Pz, the shaded areas indicate the standard error (SE) of the grand average ERPs. ErrPs are commonly defined as the differences *erroneous responses - correct responses* [59]. In our case, the grand average amplitudes of correct trials are very close to zero throughout the trial, hence the presented ERPs are equivalent to the respective ErrPs.

Table 1: Comparison of the grand average reaction times and error detection times (mean ± SD).

condition	natural reaction time (ms)	fully generic error detection time (ms)	optimized generic error detection time (ms)	personalized error detection time (ms)
target	414 ± 35	368 ± 17	375 ± 15	375 ± 6
passive	427 ± 41	361 ± 27	376 ± 22	384 ± 13
active	419 ± 29	345 ± 13	372 ± 16	385 ± 7

For electrode position FCz, which measures the activity of brain areas that are usually most associated with error processing, the ErrPs of all error conditions have an initial positive peak at approximately t = 190 ms after the error onset. *Target* errors (in red) elicit subsequent negative (ERN) and positive (Pe) peaks at 235 ms (amplitude of -0.4 µV) and 315 ms (4.16 µV), respectively. ErrPs caused by *active* errors (in black) are delayed compared to *passive* errors (in blue). For the *passive* condition, we found an ERN at 250 ms (-0.27 µV) and a Pe at 345 ms (1.3 µV), and for the *active* condition at 265 ms (-0.3 µV) and 390 ms (1.08 µV). Grand average ErrPs of all error conditions display a late negative component peaking at 470 ms (*target*), 485 ms *(passive*), and 500 ms (*active*) in electrode Cz. Trials without error event (*correct*) did not show any particular change in amplitude.

We performed sample-wise paired statistical tests for FCz, Cz, and Pz, each comparing the grand average ErrPs of two error conditions with one another. Significant samples, after FDR correction for multiple testing (alpha = 0.05), for the three comparisons (*target* vs. *passive*, *target* vs. *active*, and *passive* vs. *active*) are indicated by dots below the corresponding ErrPs in Fig. 4.

ERN and Pe are commonly found as components of ErrPs and appear mainly in the frontocentral and centroparietal regions of the brain. This is particularly visible in the topographical distributions of the averaged ErrPs of all 58 channels (see Fig. 4, bottom).

4.3 Error Detection

In **RQ1** we wanted to analyze whether we could asynchronously detect errors in this navigation task. Fig. 5 illustrates the classification results of the three investigated error detection models described in Sect. 3.4.3. As expected, classification accuracies, expressed as TPR and TNR, improve when using personalized models compared to generic models. We show the results for three binary classification problems, i.e., each error condition vs. *correct*. We obtained the lowest TPR and TNR with the fully generic classifier: average TPR and TNR for *target* are 70% and 87%, for *passive* 60% and 74%, and *active* 52% and 59%. Evaluation metrics improve when optimizing τ and k. The average TPR and TNR for the optimized generic model, computed as a mean of ten times five-fold CV are 76% and 91% for *target*, 63% and 81% for *passive*, and 53% and 79% for *active*. We achieved the highest TPR and TNR performing a personalized classification (with optimized hyperparameters) with 85% and 94% (*target*), 73% and 88% (*passive*), and 66% and 85% (*active*). Individual results for each participant are depicted in Fig. 5a-c. Table 2 summarizes the classification results.

However, on average errors were detected the fastest with the fully generic model, resulting in 368 ms for *target*, 361 ms for *passive*, and 345 ms for *active*. Error detec-

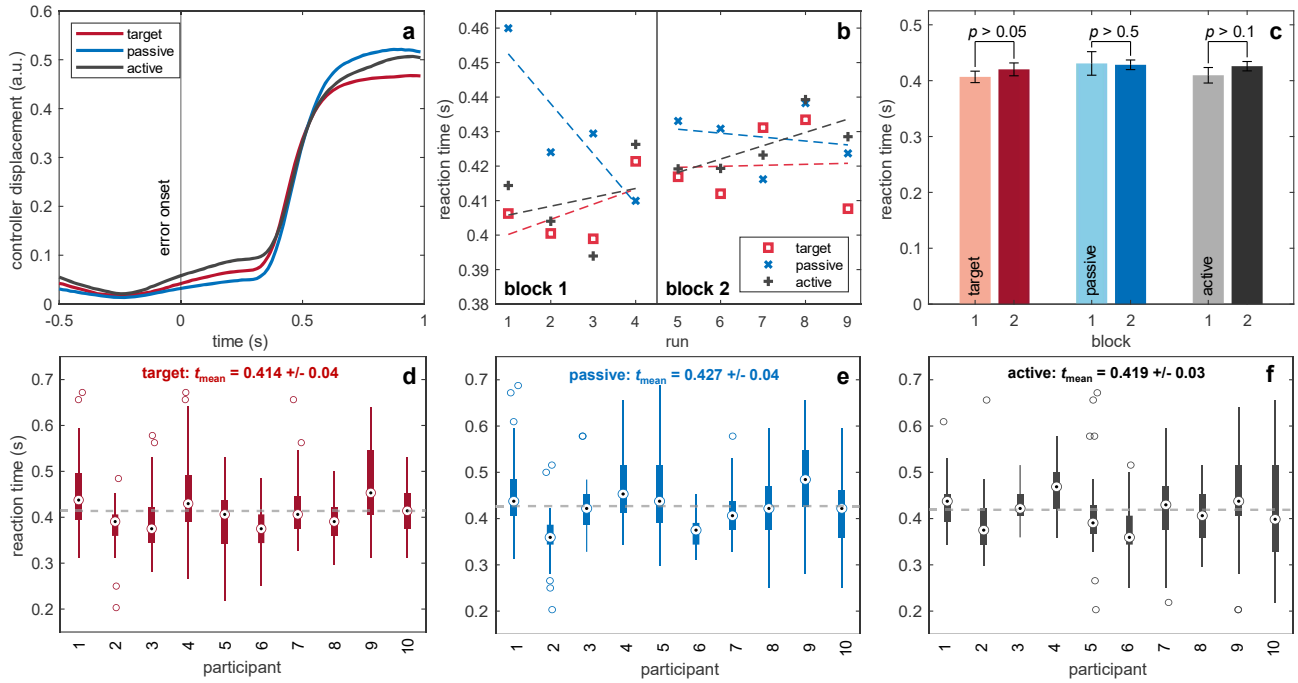

Figure 3: Behavioral results of ten participants. All times are relative to the error onset at $t = 0$ s. (a) Average controller displacement (to the left or right) at the error event. (b) Development of the reaction time. The mean reaction times of the participant averages of each run are indicated, the dashed lines are linear regressions of the average reaction times. (c) Comparison of the average reaction times of block 1 and block 2. Error bars show ± SD. (d)-(f) Boxplots of the individual reaction times of each participant. Average reaction times per error condition are indicated in each subfigure (grand average mean ± SD).

Table 2: Comparison of the grand average classification results from the generic and personalized error detection models for three binary classification problems (each error condition vs. *correct*).

condition	fully generic		optimized generic		personalized	
%	TPR	TNR	TPR	TNR	TPR	TNR
target	70.2	87.0	76.4	90.9	85.1	94.1
passive	60.2	74.1	62.5	81.2	73.4	87.9
active	52.0	59.4	53.1	79.4	65.5	85.2

tion was slightly slower with the optimized generic model: 375 ms (*target*), 376 ms (*passive*), and 372 ms (*active*). With the personalized model, we detected errors on average after 375 ms in the *target* condition, after 384 ms for *passive*, and after 385 ms in the *active* condition. The times were computed from all correctly classified errors. The distributions of all correctly detected errors are shown in Fig. 5d. To answer **RQ2**, we compared the detection times with the participants' reaction times. A summary of the average detection times and the corresponding reaction times can be found in Table 1. We compared the reaction times of all conditions with the respective error detection times obtained with the three models. After performing Wilcoxon rank-sum tests and FDR correction (alpha = 0.05), we found the reaction times and their corresponding detection times to be significantly different for all models and in all conditions.

5 Discussion

We developed a study to analyze the behavioral and electrophysiological responses to erroneous events in a VR navigation task. Further, we wanted to asynchronously detect errors in a task with continuous control and continuous feedback, and investigate how different classification approaches, i.e., generic and personalized, compare. Finally, we intended to compare the error detection times and the natural reaction time of humans. Allowing for natural reaction enabled us to study an unbiased behavior, which is the closest proxy for a real-life scenario. To study these research questions, we designed a task in which participants were instructed to navigate a glider through targets in VR. To enhance the realism of the setting, participants were sitting in a physical glider, which was also the template for the virtual glider in the flight simulation. To further increase the sense of presence, i.e., the sense of *being in* the VR [65], participants could steer the virtual glider using the control stick of the physical glider. This should enhance the participants' engagement [13], as there is evidence that motivation influences error perception [26].

5.1 Electrophysiology

We could identify EEG correlates of different error types that are in alignment with existing literature [10], showing distinct negative and positive peaks that can be interpreted as ERN and Pe, respectively. Correct trials did not show any error-related responses. However, grand average results show that errors perceived while participants actively steered the glider elicited a delayed response compared to passive error events. This is particularly visible in the time-domain signals and the corresponding statistically different samples in the segment between the ERN and the Pe. A similar effect has been described by Lopes-Dias and colleagues [37], who found that a masked (jittered) error onset delays the succeeding ErrP. Interestingly, we found significant differences between *target* and the other error classes before the error event. This difference is visible as a negative deflection in the ErrP in *target* and is likely related to the participants' expectation to reach the target. However, investigations

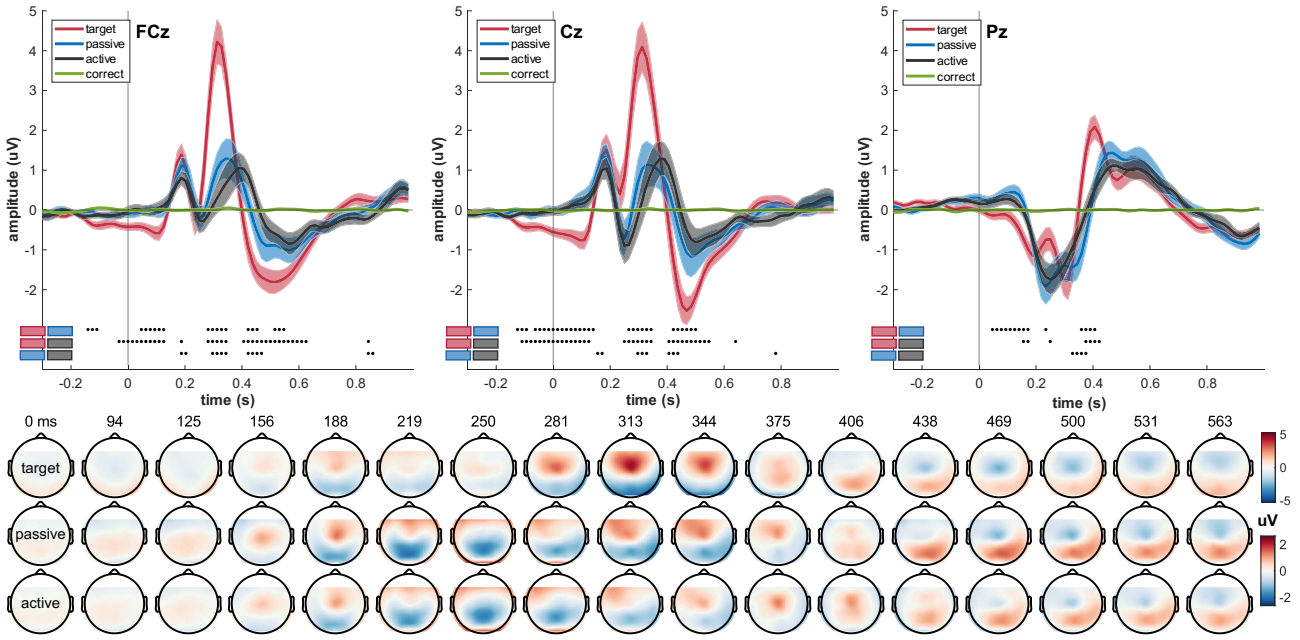

Figure 4: Grand average neurophysiological results calculated from 19 participants. All times relative to the error onset at $t = 0$ s. Top: Grand average ErrPs at electrode positions FCz (left), Cz (middle), and Pz (right). Shaded areas indicate the SE of the participant averages. The dots show significantly different time points, comparisons are indicated by the rectangles at the bottom left of each subfigure (e.g., red-blue means *target* vs. *passive*). Bottom: Topographical distributions of the ErrPs of all 58 channels for the conditions *target* (top row), *passive* (middle row), and *active* (bottom row). The corresponding time points are shown above the top row.

on this observation lie outside the scope of this work.

5.2 Error detection

We implemented generic and personalized approaches for error detection. Previous studies have shown the performance advantage of personalized over generic classifiers [30, 35]. Nevertheless, generic models are considered more applicable because they do not require any training with data from the end user before employing them in HCI systems [35]. Another possibility to reduce the training time for ErrP classifiers is to use models trained for the detection of different brain patterns, which are expected to be similar to ErrPs. This has been attempted at the cost of poorer performance [27, 30].

For error detection, we created three binary classification problems, i.e., each error condition vs. *correct*. On average, all classifiers could detect errors with reasonable to high accuracy, with the *active* fully generic detection as an exception. With the personalized models, we achieved average TPRs between 65% and 85%, while the average TNRs were above 85% for all classification problems. Notably, we designed the classifier to optimize the hyperparameters, i.e., the threshold τ and k consecutive windows, such that the product $TPR \cdot TNR$ is maximized. This optimization criterion can be adapted to fit different scenarios, e.g., for applications in which mistakenly detected errors are more problematic than missed errors.

We decided not to take into consideration other features than the ones obtained from the time-domain signals, as proposed in related works [27, 47–49], including frequency-domain features that are not commonly considered in studies with ErrPs [9]. However, Spüler et al. [69] and Lopes-Dias et al. [37] did not report classification improvements. Moreover, other classification methods potentially yield better accuracies and could be tested in future research. Recently, Sosulski and Tangermann [66] suggested using block-Toeplitz covariance matrices for LDA classification and outperformed sLDA on multiple data sets.

5.3 Reaction time vs. detection time

We did not instruct the participants on how to react to errors. This allowed us to analyze their natural reactions to errors. Ten of the 19 participants showed corrective behavior, i.e., tried to reach the targets despite the error events. We found that the reaction time of these ten participants was slower than the respective error detection times with all classifiers. Error detections in real-time could be used, e.g., to provide aid for the user or to allow systems to autonomously correct erroneous actions without interrupting the user's flow in performing tasks [46]. Incorporating such systems would make the interaction with the VR more smoothly and arguably improve the user experience. This could be of particular interest in navigation tasks, where lack of depth and dimensionality perception in the presented visualization inherently leads to more frequent erroneous actions by the user [32]. However, this work should provide a methodological foundation to address error detection in VR, without concentrating on a particular application. On the contrary, the presented approach is not limited to flight simulations but can be expanded to virtually any interaction with potentially erroneous interactions. Concrete examples that are applicable to numerous scenarios include errors in the interaction with the VR, such as object selection. Given two objects are close to each other, one object could be falsely selected. Implementing our proposed approach, the system might be able to automatically select the intended one, even faster than the user could correct the mistake. However, the particular use of such gains in time are highly dependent on the application and should be subject to future research.

5.4 Limitations

Due to the limited number of trials available, we assessed the performance of our classifiers on unseen test sets using a CV approach. Alternatively, employing validation sets separate from the test sets could reduce the risk of overfitting the models and might impact the

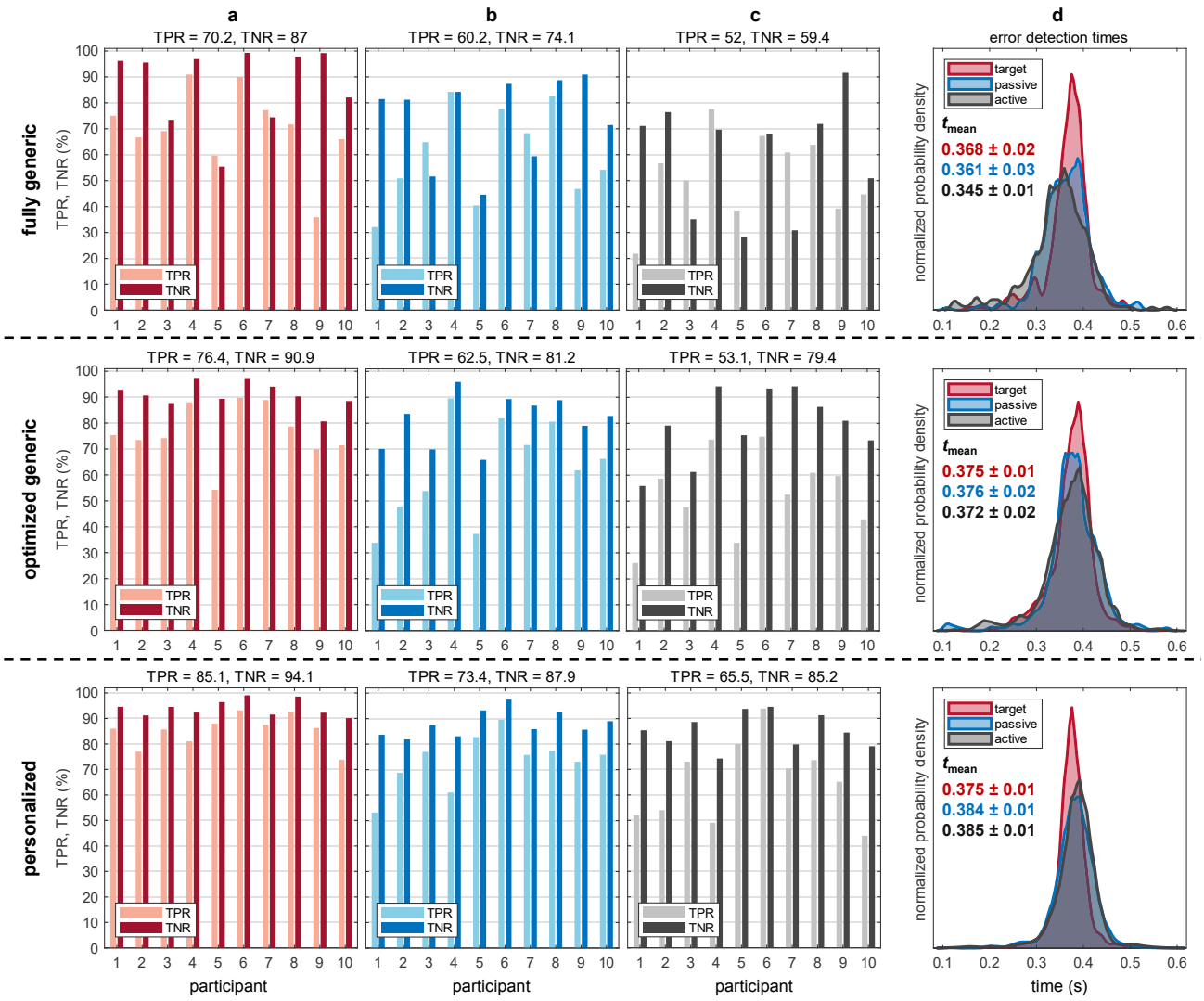

Figure 5: Error detection results for the fully generic model (top row), the generic model with optimized hyperparameters (middle row), and the personalized model (bottom row). (a)-(c) Classification results for the error conditions (a) *target*, (b) *passive*, and (c) *active*. Shown are TPR (fraction of correctly classified error trials) and TNR (fraction of correctly classified correct trials). Average TPR and TNR for each classification model and condition are given on top of the respective subfigures. (d) Distribution of the error detection times computed from all correctly classified error trials. Grand average error detection times are given (mean ± SD). Times are relative to the error onset at $t = 0$ s.

accuracies. Similarly, a transfer to continuous real-time detection would naturally increase the number of false positive detections compared to the presented asynchronous classification of trials of 1.5 s. Finally, utilizing standardized measures for presence, like the Multimodal Presence Scale [33], would allow for validation if affordances to create realistic environments were successful. However, the participants' feedback after the experiments was predominantly positive in this regard.

6 Conclusion

In this work, we explored different errors and the subsequent corrective human behavior in a VR navigation task. We asynchronously decoded errors with different classifiers, i.e., generic and personalized, that could detect erroneous events with reasonable to high accuracy. Further, we could show that error detection that solely relies on EEG signals can be faster than the natural reaction of humans. The approaches and findings presented in this work could contribute to developing systems that can detect errors and are applicable in various VR scenarios that profit from an autonomous adaptation of the presented visualization or error correction. Topics for future research in this context should include real-time error detection in VR and investigations on what means of corrective actions can improve the user experience when using such systems.

Acknowledgments

The Know-Center is funded within the Austrian COMET Program - Competence Centers for Excellent Technologies - under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

REFERENCES

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[2] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29:1165–1188, 2001.

[3] B. Blankertz, S. Lemm, M. Treder, S. Haufe, and K. R. Müller. Single-trial analysis and classification of ERP components - a tutorial. *NeuroImage*, 56:814–825, 2011. doi: 10.1016/j.neuroimage.2010.06.048

[4] J. Blumberg, J. Rickert, S. Waldert, A. Schulze-Bonhage, A. Aertsen, and C. Mehring. Adaptive classification for brain computer interfaces. In *Proc. EMBC*, pp. 2536–2539, 2007. doi: 10.1109/IEMBS.2007.4352845

[5] G. Bouyer, A. Chellali, and A. Lécuyer. Inducing self-motion sensations in driving simulators using force-feedback and haptic motion. In *Proc. VR*, pp. 84–90. IEEE Comput. Soc. Press, Los Alamitos, CA, USA, 2017. doi: 10.1109/VR.2017.7892234

[6] D. A. Bowman, D. Koller, and L. F. Hodges. A methodology for the evaluation of travel techniques for immersive virtual environments. *Virtual Reality*, 3:120–131, 1998.

[7] F. P. Brooks. What's real about virtual reality? *IEEE Computer Graphics and Applications*, 19:16–27, 11 1999. doi: 10.1109/38.799723

[8] C. Brunner, N. Birbaumer, B. Blankertz, C. Guger, A. Kübler, D. Mattia, J. D. R. Millán, F. Miralles, A. Nijholt, E. Opisso, N. Ramsey, P. Salomon, and G. R. Müller-Putz. BNCI horizon 2020: towards a roadmap for the BCI community. *Brain-Computer Interfaces*, 2:1–10, 1 2015. doi: 10.1080/2326263X.2015.1008956

[9] A. Buttfield, P. Ferrez, and J. D. R. Millán. Towards a robust BCI: error potentials and online learning. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 14(2):164–168, 2006. doi: 10.1109/TNSRE.2006.875555

[10] R. Chavarriaga, A. Sobolewski, and J. D. R. Millán. Errare machinale est: The use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*, 8, 2014. doi: 10.3389/fnins.2014.00208

[11] K.-J. Chiang, D. Emmanouilidou, H. Gamper, D. Johnston, M. Jalobeanu, E. Cutrell, A. Wilson, W. W. An, and I. Tashev. A closed-loop adaptive brain-computer interface framework: Improving the classifier with the use of error-related potentials. In *Proc. NER*, pp. 487–490, 2021.

[12] C. Cruz-Neira, D. J. Sandin, and T. A. Defanti. Surround-screen projection-based virtual reality: The design and implementation of the CAVE. In *Proc. SIGGRAPH*, pp. 135–142, 1993.

[13] J. J. Cummings and J. N. Bailenson. How immersive is enough? a meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, 19:272–309, 4 2016. doi: 10.1080/15213269.2015.1015740

[14] A. Delorme and S. Makeig. EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134:9–21, 3 2004. doi: 10.1016/j.jneumeth.2003.10.009

[15] A. Delorme, T. Sejnowski, and S. Makeig. Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *NeuroImage*, 34:1443–1449, 2 2007. doi: 10.1016/j.neuroimage.2006.11.004

[16] C. L. Dias, A. I. Sburlea, and G. R. Müller-Putz. Masked and unmasked error-related potentials during continuous control and feedback. *Journal of Neural Engineering*, 15, 4 2018. doi: 10.1088/1741-2552/aab806

[17] J. Diedrichsen, Y. Hashambhoy, T. Rane, and R. Shadmehr. Neural correlates of reach errors. *Journal of Neuroscience*, 25:9919–9931, 10 2005. doi: 10.1523/JNEUROSCI.1874-05.2005

[18] M. Falkenstein, J. Hohnsbein, J. Hoormann, and L. Blanke. Effects of crossmodal divided attention on late ERP components. II. error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology*, 78:447–455, 1991. doi: 10.1016/0013-4694(91)90062-9

[19] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein. ERP components on reaction errors and their functional significance: a tutorial. *Journal of Experimental Psychology: Human Perception and Performance*, 51:42–54, 2000.

[20] P. W. Ferrez and J. D. R. Millán. You are wrong!-automatic detection of interaction errors from brain waves. pp. 1413–1418, 2005.

[21] P. W. Ferrez and J. D. R. Millán. Error-related EEG potentials generated during simulated brain-computer interaction. *IEEE Transactions on Biomedical Engineering*, 55:923–929, 3 2008. doi: 10.1109/TBME.2007.908083

[22] P. W. Ferrez and J. D. R. Millán. Simultaneous real-time detection of motor imagery and error-related potentials for improved BCI accuracy. In *Proc. 4th Int. Brain-Comput. Interface Workshop Training Course*, 2008.

[23] W. J. Gehring, B. Goss, M. G. Coles, D. E. Meyer, and E. Donchin. A neural system for error detection and compensation. *Psychological science*, 4(6):385–390, 1993.

[24] L. Gehrke, S. Akman, P. Lopes, A. Chen, A. K. Singh, H.-T. Chen, C.-T. Lin, and K. Gramann. Detecting visuo-haptic mismatches in virtual reality using the prediction error negativity of event-related brain potentials. In *Proc. CHI*, pp. 1–11, 2019.

[25] L. Gehrke, P. Lopes, M. Klug, S. Akman, and K. Gramann. Neural sources of prediction errors detect unrealistic VR interactions. *Journal of Neural Engineering*, 19, 6 2022. doi: 10.1088/1741-2552/ac69bc

[26] G. Hajcak, J. S. Moser, N. Yeung, and R. F. Simons. On the ERN and the significance of errors. *Psychophysiology*, 42:151–160, 3 2005. doi: 10.1111/j.1469-8986.2005.00270.x

[27] F. Iwane, I. Iturrate, R. Chavarriaga, and J. D. R. Millán. Invariability of EEG error-related potentials during continuous feedback protocols elicited by erroneous actions at predicted or unpredicted states. *Journal of Neural Engineering*, 18, 8 2021. doi: 10.1088/1741-2552/abfa70

[28] S. Jalilpour and G. R. Müller-Putz. Balance perturbation and error processing elicit distinct brain dynamics. *Journal of Neural Engineering*, 2023. doi: 10.1088/1741-2552/acc486

[29] T. Katila, R. Maniewski, T. Poutanen, T. Varpula, and P. J. Karp. Magnetic fields produced by the human eye. *Journal of Applied Physics*, 52:2565–2571, 1981. doi: 10.1063/1.329000

[30] S. K. Kim and E. A. Kirchner. Handling few training data: Classifier transfer between different types of error-related potentials. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24:320–332, 3 2016. doi: 10.1109/TNSRE.2015.2507868

[31] R. J. Kobler, A. I. Sburlea, C. Lopes-Dias, A. Schwarz, M. Hirata, and G. R. Müller-Putz. Corneo-retinal-dipole and eyelid-related eye artifacts can be corrected offline and online in electroencephalographic and magnetoencephalographic signals. *NeuroImage*, 218, 9 2020. doi: 10.1016/j.neuroimage.2020.117000

[32] J. J. LaViola Jr, E. Kruijff, R. P. McMahan, D. Bowman, and I. P. Poupyrev. *3D user interfaces: theory and practice*. Addison-Wesley Professional, 2017.

[33] K. M. Lee. Presence, explicated. *Communication theory*, 14(1):27–50, 2004.

[34] A. Llera, M. A. van Gerven, V. Gómez, O. Jensen, and H. J. Kappen. On the use of interaction error potentials for adaptive brain computer interfaces. *Neural Networks*, 24:1120–1127, 12 2011. doi: 10.1016/j.neunet.2011.05.006

[35] C. Lopes-Dias, A. I. Sburlea, K. Breitegger, D. Wyss, H. Drescher, R. Wildburger, and G. R. Müller-Putz. Online asynchronous detection of error-related potentials in participants with a spinal cord injury using a generic classifier. *Journal of Neural Engineering*, 18, 8 2021. doi: 10.1088/1741-2552/abd1eb

[36] C. Lopes-Dias, A. I. Sburlea, and G. R. Muller-Putz. A generic error-related potential classifier offers a comparable performance to a personalized classifier. In *Proc. EMBC*, vol. 2020-July, pp. 2995–2998, 7 2020. doi: 10.1109/EMBC44109.2020.9176640

[37] C. Lopes-Dias, A. I. Sburlea, and G. R. Müller-Putz. Online asynchronous decoding of error-related potentials during the continuous control of a robot. *Scientific Reports*, 9, 12 2019. doi: 10.1038/s41598-019-54109-x

[38] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger. A review of classification algorithms for EEG-based brain-computer interfaces: A 10 year update. *Journal of Neural Engineering*, 15, 4 2018. doi: 10.1088/1741-2552/aab2f2

[39] P. Luu, D. M. Tucker, and S. Makeig. Frontal midline theta and the error-related negativity: Neurophysiological mechanisms of action

regulation. *Clinical Neurophysiology*, 115:1821–1835, 8 2004. doi: 10.1016/j.clinph.2004.03.031

[40] S. Makeig, A. Bell, T.-P. Jung, and T. J. Sejnowski. Independent component analysis of electroencephalographic data. *Advances in neural information processing systems*, 8, 1995.

[41] S. G. Mason and G. E. Birch. A brain-controlled switch for asynchronous control applications. *IEEE Transactions on Biomedical Engineering*, 47:1297–1307, 10 2000. doi: 10.1109/10.871402

[42] J. McCrae, I. Mordatch, M. Glueck, and A. Khan. Multiscale 3D navigation. In *Proc. I3D*, pp. 7–14, 2009. doi: 10.1145/1507149.1507151

[43] D. J. McFarland, L. M. McCane, S. V. David, and J. R. Wolpaw. Spatial filter selection for EEG-based communication. *Electroencephalography and clinical Neurophysiology*, 103:386–394, 1997.

[44] B. W. McMenamin, A. J. Shackman, L. L. Greischar, and R. J. Davidson. Electromyogenic artifacts and electroencephalographic inferences revisited. *NeuroImage*, 54:4–9, 1 2011. doi: 10.1016/j.neuroimage.2010.07.057

[45] T. Milekovic, T. Ball, A. Schulze-Bonhage, A. Aertsen, and C. Mehring. Error-related electrocorticographic activity in humans during continuous movements. *Journal of Neural Engineering*, 9, 4 2012. doi: 10.1088/1741-2560/9/2/026007

[46] M. H. Miraz, M. Ali, and P. S. Excell. Adaptive user interfaces and universal usability through plasticity of user interface design. *Computer Science Review*, 40, 5 2021. doi: 10.1016/j.cosrev.2021.100363

[47] J. Omedes, I. Iturrate, R. Chavarriaga, and L. Montesano. Asynchronous decoding of error potentials during the monitoring of a reaching task. In *Proc. SMC*, pp. 3116–3121, 1 2015. doi: 10.1109/SMC.2015.541

[48] J. Omedes, I. Iturrate, J. Minguez, and L. Montesano. Analysis and asynchronous detection of gradually unfolding errors during monitoring tasks. *Journal of Neural Engineering*, 12, 7 2015. doi: 10.1088/1741-2560/12/5/056001

[49] J. Omedes, I. Iturrate, and L. Montesano. Asynchronous detection of error potentials. In *Proc. 6th Int. Brain-Comput. Interface Conf.*, 2014. doi: 10.3217/978-3-85125-378-8-65

[50] R. Oostenveld and P. Praamstra. The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, 112:713–719, 2001.

[51] G. Padrao, M. Gonzalez-Franco, M. V. Sanchez-Vives, M. Slater, and A. Rodriguez-Fornells. Violating body movement semantics: Neural signatures of self-generated and external-generated errors. *NeuroImage*, 124:147–156, 1 2016. doi: 10.1016/j.neuroimage.2015.08.022

[52] R. Pausch, J. Snoddy, R. Taylor, S. Watson, and E. Haseltine. Disney's aladdin: first steps toward storytelling in virtual reality. In *Proc. SIGGRAPH*, pp. 193–203, 1996.

[53] E. F. Pavone, G. Tieri, G. Rizza, E. Tidoni, L. Grisoni, and S. M. Aglioti. Embodying others in immersive virtual reality: Electro-cortical signatures of monitoring the errors in the actions of an avatar seen from a first-person perspective. *Journal of Neuroscience*, 36:268–279, 1 2016. doi: 10.1523/JNEUROSCI.0494-15.2016

[54] R. Pezzetta, V. Nicolardi, E. Tidoni, and S. M. Aglioti. Error, rather than its probability, elicits specific electrocortical signatures: a combined EEG-immersive virtual reality study of action observation. *Journal of Neurophysiology*, 120:1107–1118, 2018. doi: 10.1152/jn.00130.2018

[55] T. Porssut, F. Iwane, R. Chavarriaga, O. Blanke, J. d. R. Millán, R. Boulic, and B. Herbelin. EEG signature of breaks in embodiment in VR. *PLOS One*, 18(5), 2023. doi: 10.1371/journal.pone.0282967

[56] H. S. Pulferer, K. Kostoglou, and G. R. Müller-Putz. Getting off track: Cortical feedback processing network modulated by continuous error signal during target-feedback mismatch. *NeuroImage*, 274, 2023. doi: 10.1016/j.neuroimage.2023.120144

[57] H. S. Pulferer and G. R. Müller-Putz. Continuous error processing during a closed-loop 2D tracking task. *Current Directions in Biomedical Engineering*, 8(2):173–176, 2022.

[58] G. Raz, G. Gurevitch, T. Vaknin, A. Aazamy, I. Gefen, S. Grunstein, G. Azouri, and N. Goldway. Electroencephalographic evidence for the involvement of mirror-neuron and error-monitoring related processes in virtual body ownership. *NeuroImage*, 207, 2 2020. doi: 10.1016/j.neuroimage.2019.116351

[59] G. Schalk, J. R. Wolpaw, D. J. McFarland, and G. Pfurtscheller. EEG-based communication: presence of an error potential. *Clinical Neurophysiology*, 111:2138–2144, 2000.

[60] H. T. V. Schie, R. B. Mars, M. G. Coles, and H. Bekkering. Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, 7:549–554, 5 2004. doi: 10.1038/nn1239

[61] B. D. Seno, M. Matteucci, and L. Mainardi. Online detection of P300 and error potentials in a BCI speller. *Computational Intelligence and Neuroscience*, 2010, 2010. doi: 10.1155/2010/307254

[62] D. Serje and E. Acuña. Driving and flying simulators: A review on relevant considerations and trends. *Transportation Research Record*, 2676(3):551–570, 2022. doi: 10.1177/03611981211052963

[63] H. Si-Mohammed, C. Lopes-Dias, M. Duarte, F. Argelaguet, C. Jeunet, G. Casiez, G. R. Müller-Putz, A. Lécuyer, and R. Scherer. Detecting system errors in virtual reality using EEG through error-related potentials. In *Proc. VR*, pp. 653–661. IEEE Comput. Soc. Press, Los Alamitos, CA, USA, 2020. doi: 10.1109/VR46266.2020.00088

[64] A. K. Singh, H. T. Chen, Y. F. Cheng, J. T. King, L. W. Ko, K. Gramann, and C. T. Lin. Visual appearance modulates prediction error in virtual reality. *IEEE Access*, 6:24617–24624, 5 2018. doi: 10.1109/ACCESS.2018.2832089

[65] M. Slater and S. Wilbur. A framework for immersive virtual environments (FIVE): Speculations on the role of presence in virtual environments. *Presence: Teleoperators and Virtual Environments*, 6:603–616, 1997. doi: 10.1162/PRES.1997.6.6.603

[66] J. Sosulski and M. Tangermann. Introducing block-toeplitz covariance matrices to remaster linear discriminant analysis for event-related potential brain–computer interfaces. *Journal of Neural Engineering*, 19(6), 2022. doi: 10.1088/1741-2552/ac9c98

[67] G. Spinelli, R. Pezzetta, L. Canzano, E. Tidoni, and S. M. Aglioti. Brain dynamics of action monitoring in higher-order motor control disorders: The case of apraxia. *eNeuro*, 9, 3 2022. doi: 10.1523/ENEURO.0334-20.2021

[68] G. Spinelli, G. Tieri, E. F. Pavone, and S. M. Aglioti. Wronger than wrong: Graded mapping of the errors of an avatar in the performance monitoring system of the onlooker. *NeuroImage*, 167:1–10, 2 2018. doi: 10.1016/j.neuroimage.2017.11.019

[69] M. Spüler and C. Niethammer. Error-related potentials during continuous feedback: Using EEG to detect errors of different type and severity. *Frontiers in Human Neuroscience*, 9, 3 2015. doi: 10.3389/fnhum.2015.00155

[70] W. O. Tatum, B. A. Dworetzky, and D. L. Schomer. Artifact and recording concepts in EEG. *Journal of Clinical Neurophysiology*, 28:252–263, 6 2011. doi: 10.1097/WNP.0b013e31821c3c93

[71] L. T. Trujillo and J. J. Allen. Theta EEG dynamics of the error-related negativity. *Clinical Neurophysiology*, 118:645–668, 3 2007. doi: 10.1016/j.clinph.2006.11.009

[72] M. Wimmer, N. Weidinger, E. Veas, and G. R. Müller-Putz. Toward Hybrid BCI: EEG and Pupillometric Signatures of Error Perception in an Immersive Navigation Task in VR. In *Proc. 10th Int. Brain-Comput. Interface Meeting*, p. 142, 2023. doi: 10.3217/978-3-85125-962-9-137

[73] M. Wimmer, N. Weidinger, E. Veas, and G. R. Müller-Putz. Neural and Pupillometric Correlates of Error Perception in an Immersive VR Flight Simulation. In *Proc. EMBC*, 2023, to be published.

[74] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G. Pfurtscheller, and T. M. Vaughan. Brain-computer interfaces for communication and control. *Clinical Neurophysiology*, 113:767–791, 2002.

[75] J. Yordanova, M. Falkenstein, J. Hohnsbein, and V. Kolev. Parallel systems of error processing in the brain. *NeuroImage*, 22:590–602, 6 2004. doi: 10.1016/j.neuroimage.2004.01.040

[76] M. A. Zayer, P. MacNeilage, and E. Folmer. Virtual locomotion: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 26:2315–2334, 6 2020. doi: 10.1109/TVCG.2018.2887379

[77] H. Zhang, R. Chavarriaga, Z. Khaliliardali, L. Gheorghe, I. Iturrate, and J. D. R. Millán. EEG-based decoding of error-related brain activity in a real-world driving task. *Journal of Neural Engineering*, 12, 2015. doi: 10.1088/1741-2560/12/6/066028