Journal of Neural Engineering

Towards error categorisation in BCI: single-trial EEG classification between different errors

Towards error categorisation in BCI: single-trial EEG classification between different errors

C Wirth [1,*], P M Dockree [2], S Harty [2], E Lacey [2], and M Arvaneh [1]

[1]Automatic Control and Systems Engineering Department, University of Sheffield, Sheffield, UK
[2]Institute of Neuroscience, Trinity College Dublin, Dublin, Ireland

E-mail: cwirth1@sheffield.ac.uk

Abstract.

Objective: Error-related potentials (ErrP) are generated in the brain when humans perceive errors. These ErrP signals can be used to classify actions as erroneous or non-erroneous, using single-trial electroencephalography (EEG). A small number of studies have demonstrated the feasibility of using ErrP detection as feedback for reinforcement-learning-based Brain-Computer Interfaces (BCI), confirming the possibility of developing more autonomous BCI. These systems could be made more efficient with specific information about the type of error that occurred. A few studies differentiated the ErrP of different errors from each other, based on direction or severity. However, errors cannot always be categorised in these ways. We aimed to investigate the feasibility of differentiating very similar error conditions from each other, in the absence of previously explored metrics.

Approach: In this study, we used two data sets with 25 and 14 participants to investigate the differences between errors. The two error conditions in each task were similar in terms of severity, direction and visual processing. The only notable differences between them were the varying cognitive processes involved in perceiving the errors, and differing contexts in which the errors occurred. We used a linear classifier with a small feature set to differentiate the errors on a single-trial basis.

Results: For both data sets, we observed neurophysiological distinctions between the ErrPs related to each error type. We found further distinctions between age groups. Furthermore, we achieved statistically significant single-trial classification rates for most participants included in the classification phase, with mean overall accuracy of 65.2% and 65.6% for the two tasks.

Significance: As a proof of concept our results showed that it is feasible, using single-trial EEG, to classify these similar error types against each other. This study paves the way for more detailed and efficient learning in BCI, and thus for a more autonomous human-machine interaction.

Keywords: ErrP, EEG, Classification, BCI, Human Machine Interaction, Neurophysiology, Error detection

Submitted to: *J. Neural Eng.*

1. Introduction

When a human recognises that an error has been committed, either by themselves or in actions that they are observing, characteristic signals known as error-related potentials (ErrP) are generated in the brain [1]. A number of studies have shown that it is possible to differentiate between errors and correct actions, by detecting ErrP using electroencephalography (EEG), on a single-trial basis [2, 3, 4, 5]. Interestingly, previous studies have confirmed the possibility of using single-trial error vs non-error classification as a feedback function for a reinforcement learning-based Brain Computer Interfaces (BCI) [2, 3, 4, 5]. This opens up the possibility of moving toward autonomous BCI systems, allowing the machine to learn appropriate low-level actions based on the human's perceptions of which actions are correct, and which are errors. Such systems are able to reduce human mental workload by learning quasi-optimal solutions in scenarios such as simple navigation tasks [3, 4]. However, when tasks increase in complexity, learning will become slower if the only available information is whether a given action was correct or erroneous. Hence, if a system can be given more detailed information about the type of error that occurred, it can correct its actions more appropriately, and learn more quickly.

More recently, a handful of studies have shown that, beyond classifying errors against correct actions, it is possible to distinguish different errors against each other based on their ErrP. In a study by Iturrate et al., participants observed a virtual robotic arm, which had the task of selecting a specific basket [6]. However, the arm also could erroneously select baskets 1 or 2 steps away from the target, to the left and to the right. The study showed that there were significant differences between the ErrP for errors to the left vs those to the right, and also between those of small vs large errors. In addition to this, a small number of studies have considered neurophysiological differences arising from varying sources of errors. Different ErrP and error types that have been discussed are as follows: "response ErrP" caused when a human recognised that they have responded incorrectly to a task [7, 8, 5], "feedback ErrP" caused when a human is informed that they have made an error of which they were previously unaware [7, 5], "observation ErrP occurring when a human observes an error committed by a machine or another human [7, 5], "execution errors" occurring when a machine fails to execute a command as instructed by the human [9, 5], and "outcome errors" appearing when a human experiences a task failure [9, 5]. A study by Spüler and Niethammer showed that it is possible to classify outcome errors (committed by a human) against execution errors (committed by a machine) on a single-trial basis [9].

Despite these recent advances, the vast majority of literature in the field concerns the classification of errors against correct actions, rather than the classification of different error types against each other. Where single-trial error categorisation has been explored in a few recent studies, metrics that have been considered to distinguish the error categories include direction, severity, and whether the error was committed by the human or the machine. However, different errors cannot always be categorised by such metrics. For example, if we are trying to navigate to a target location we could either take a wrong turn on the way, or we could reach the target but then pass it. These two errors could be of the same direction and magnitude, and therefore indistinguishable by currently explored metrics, but knowing which one had occurred would provide useful information. Therefore, it is important to consider whether there are significant neurophysiological distinctions in EEG signals between the brain's responses to very similar error conditions, even in cases where metrics explored in existing literature are not available.

To address this question, we evaluated data from two tasks. In the first task, users were presented with "go" and "no-go" stimuli and asked to respond to "go" stimuli, but withhold responses to "no-go" stimuli. All of the errors considered by this experiment were response errors committed by humans who failed to withhold responses to "no-go" stimuli, and then recognised their own errors. None of the errors had any direction associated with them, and participants were not instructed to consider any errors as more or less severe than any others. The key difference between the error conditions lay in the cognitive processes required to recognise them, with the recognition of one error condition being more memory-dependent than the other. In the second task, users observed a virtual robot attempting to navigate to, and grab, a target object. Here, we investigated users' EEG responses to two navigational errors: moving away from the target when in position and ready to grab it, and moving further away from the target object if not already in position. Errors were equally likely to be made to the left or the right. In this case, all errors were being committed by the machine. As with the first task, direction could not be used to distinguish the error conditions, and users were not told to consider either error to be more or less severe than the other. As such, the error conditions considered here could not be differentiated by metrics used in existing literature. However, the contexts in which the errors arose differed slightly: In one condition, the expected correct action would be a lateral movement towards the target. In the other condition, the expected correct action would be to grab the target. We aimed to use distinctions in the EEG signals, arising from these subtle differences of cognitive load and context, to classify the error conditions against each other.

To explore the neurophysiological distinctions between the responses to these error conditions, we used time domain data to compare the latency and amplitude of key ErrP features: the error-related negativity (ERN), and the error positivity (Pe). The ERN is a negative deflection, usually peaking fronto-centrally around 100ms after an error [10, 1, 2]. The Pe is a slower positive wave, often peaking centro-parietally between 200-400ms after the error [10, 11, 12, 2]. In contrast with the ERN, the Pe has been shown to depend on participants' awareness and confidence that an error has been committed [13, 14, 15, 16], suggesting that the Pe is linked to conscious processing of errors. In addition to amplitude, the "build-up rate" of the Pe (i.e. the steepness of the slope as amplitude increases to the peak) has also been identified as a marker of evidence accumulation for error detection [17]. Further to this, secondary Pe peaks have been identified, again being linked to conscious, evaluative processes [18, 19]. The ERN and

Pe have been displayed in a variety of previous single-trial error classification studies [2, 3, 7, 9].

We also investigated the spatial distribution of the brain's response to each error condition, using topographical maps. In order to distinguish between error conditions on a single-trial basis, we employed a stepwise linear discriminant analysis classification strategy, using a small, highly discriminative set of time domain features from 20 electrode sites. We tested the efficacy of this strategy using data from 20 young and 5 older adults performing one task, and 14 young adults performing the other task.

2. Methods

2.1. Participants

This study used data collected during two tasks, which we refer to as the "Error Awareness Dot Task" (EADT) and the "Claw Observation Task" (COT). Fifty-four healthy adults were recruited for the EADT. 28 of these were young (aged 18-34) and 26 were older (aged 65-80). Seventeen healthy adults were recruited for the COT.

All of these participants were included in neurophysiological analyses, but some were excluded from the single-trial classification phase of this study. 23 were excluded from the EADT (4 young, 19 older) due to not producing enough artefact-free trials for all conditions. A further 6 from the EADT (4 young, 2 older) were excluded as it may have been possible to classify their data based on motor signals, rather than ErrPs. The rationale for these exclusions is explained in further detail in section 2.4.1. This left 25 participants from the EADT (20 young, 5 older) to be included in the single-trial classification phase. 3 participants were excluded from the COT due to not producing enough artefact-free trials for all conditions. All COT participants used for single-trial classification were young (aged 18-35).

All participants for both tasks had normal or corrected-to-normal vision. They reported no history of psychiatric illness, head injury, or photosensitive epilepsy. Written informed consent was provided before testing began. All participants of the EADT also reported that they had no history of colour-blindness. All procedures for both tasks were in accordance with the Declaration of Helsinki. Procedures for the EADT were approved by the Trinity College Dublin Ethics Committee, and procedures for the COT were approved by the University of Sheffield Ethics Committee in the Automatic Control and Systems Engineering Department.

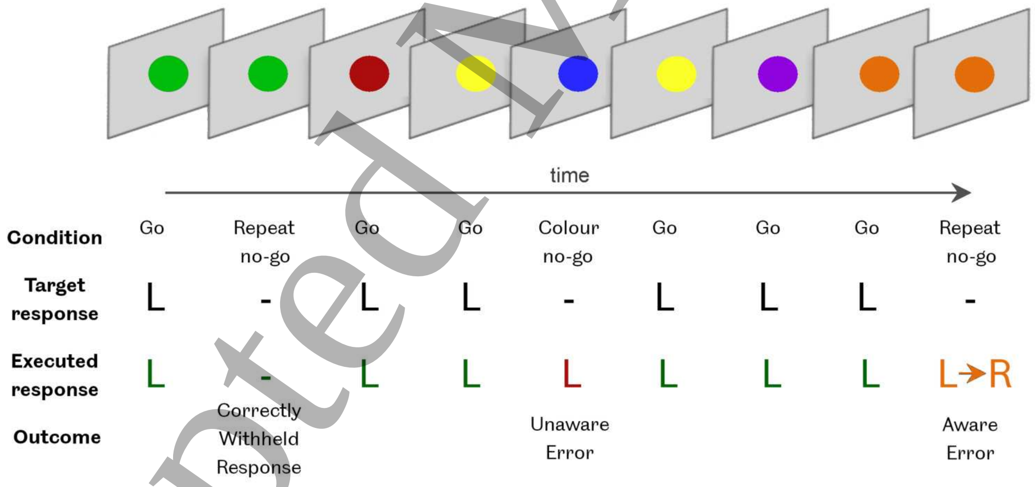
2.2. Experimental Setup

2.2.1. EEG Setup For the EADT, 64 channels of EEG were recorded at 512Hz, using the BioSemi ActiveTwo system. Electrodes were placed using the 10-20 system. Electrooculogram (EOG) electrodes were also placed at the outer cantus of each eye, and above and below the left eye. Reference electrodes were placed on the left and right mastoid.

For the COT, 20 channels of EEG were recorded at 500Hz, using an Enobio 20 5G headset. The electrode positions used were: F7, F3, Fz, F4, F8, FC1, FC2, T7, C3, Cz, C4, T8, CP1, CP2, P3, Pz, P4, PO7, PO8, and Oz. Reference electrodes were placed on the earlobe.

2.2.2. The Error Awareness Dot Task The EADT was a time-critical reaction task, requiring sustained attention. The task employed a "go/no-go" paradigm, requiring participants to react to "go" stimuli with a mouse click, but withhold their reaction in the case of "no-go" stimuli.

Participants were shown a succession of randomised, differently-coloured dots on a computer screen, with a blank grey screen shown between dots, as shown in **Figure 1**. Participants were asked to perform a left mouse click, in a timely manner, in response to the presentation of each new dot. However, in two "no-go" scenarios, they were asked to withhold their response. These scenarios were the presentation of a blue dot, or of a dot that was the same colour as the previous dot. These are known as the "colour condition" and "repeat condition", respectively. If participants did click in either of these scenarios, they were asked to perform a second click with the right mouse button, in order to indicate their awareness of the error.



Condition	Go	Repeat no-go	Go	Go	Colour no-go	Go	Go	Go	Repeat no-go
Target response	L	-	L	L	-	L	L	L	-
Executed response	L	-	L	L	L	L	L	L	L→R
Outcome		Correctly Withheld Response			Unaware Error				Aware Error

Figure 1. The Error Awareness Dot Task (EADT). Participants were asked to respond to "go" stimuli with a left mouse button click (L). They were asked to withold this response in the event of either a "colour no-go" stimulus (the stimulus is blue) or "repeat no-go" stimulus (the stimulus is the same colour as the previous stimulus). If participants performed a left mouse click following a no-go stimulus, they were asked to follow this with a right mouse button click (R), to register their awareness of their error.

Before testing began, a practice block took place, in which participants had to respond successfully to three consecutive no-go trials, either by withholding their initial response or, if they did click erroneously, by following up with an awareness click.

8 blocks of trials were collected from each participant, with the exception of five, for whom 4-6 blocks of trials were collected. Each block lasted approximately 6 minutes, and contained 176 "go" trials, 16 "repeat condition" trials, and 8 "colour condition" trials.

The duration for which each stimulus was shown varied throughout the task, depending on the accuracy of the participant in performing correct responses to go and no-go trials. Initially, stimuli were displayed for 750ms. However, if the participant's accuracy were below 50%, stimulus duration would increase to 1000ms. Conversely, if the participant's accuracy were above 60%, stimulus duration would decrease to 500ms. Accuracy between 50 and 60% would result in stimulus duration remaining at, or reverting to, 750ms. Stimulus duration was updated every 40 trials. An inter-stimulus gap, in which the screen was a blank grey, remained constant at 750ms. This meant that the time period between the onset of stimulus n and the onset of stimulus $n+1$ could vary between 1250ms and 1750ms.

2.2.3. The Claw Observation Task In the COT, the errors in question were committed by the machine and observed by the participants, as opposed to errors being committed by the participants themselves in the EADT. Thus, the COT is similar to error-driven BCI scenarios in which users observe actions made by a machine [6, 3].

Here, participants were asked to observe a computer-controlled simulation of an arcade 'claw crane' game. Participants were shown a screen with 8 coloured circles arranged in a row and, above the circles, a virtual robotic arm, as shown in **Figure 2**. A single circle, selected at random at the start of each run, was designated as the target. This circle was coloured blue and marked with a score of +25 points. Every other circle was coloured red. The red circles immediately adjacent to the target were marked with a score of -10 points, and the scores marked on each circle decreased by a further 5 points with each step further from the target. The robotic arm began each run directly above a circle either 2 or 3 steps away from the target. Every 1.5s, the robotic arm would either move 1 step to the left, move 1 step to the right, or extend downward to grab the circle beneath it. Movements occurred instantaneously. The probability of each type of action occurring depended on whether or not the arm was positioned directly above the target circle. A table of action probabilities is shown in **Table 1**.

A score was also displayed in the top left corner of the screen. When a "grab" action was performed, the score would be updated according to the score marked on the circle that had been grabbed. After each "grab" action the run would finish and the screen would become completely black. Nine of the COT participants were asked to silently count the number of times each movement error was made in each run, in an attempt to help them stay focused on the task. These participants were asked to write down the number of errors on a sheet provided at the end of each run. As such, the gap between the end of one run and the start of the next run was 10 seconds. The remaining eight COT participants were not asked to perform the counting. For these participants, the gap between runs was 5 seconds. In either case, a beep would sound

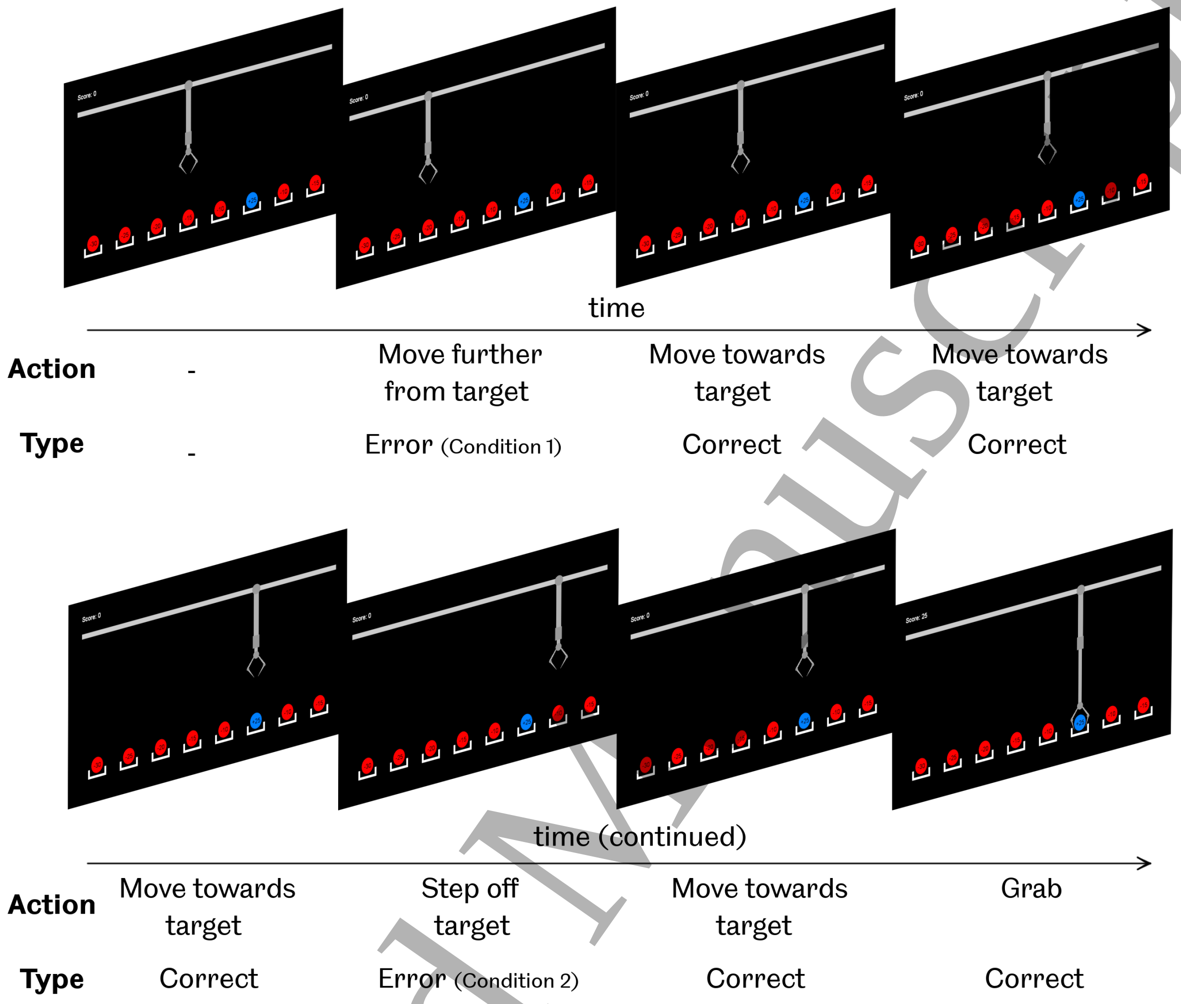


Figure 2. The Claw Observation Task (COT). Participants were asked to observe as a virtual robotic claw attempted to navigate towards, and grab, a blue target ball. If the claw was aligned over the target ball, possible actions were either to grab the ball or take 1 step away from the target. If the claw was not aligned over the target ball, possible actions were either to move 1 step towards the target, move 1 step further away from the target, or grab the red ball beneath the claw's current position.

1 second before the next run began. Participants were asked to refrain from movement and blinking during each run, but told that they could move and blink freely between runs, while the screen was blank. This process repeated until the end of the block, with each block lasting approximately 4 minutes. The score was reset to 0 at the beginning of each new block.

The actions considered for this study were movement errors. Movements in which the virtual robot was aligned over one of the red non-target balls, and moved further away from the target, are hereafter referred to as "condition 1" errors. Movements in which the virtual robot was aligned over blue target ball, but stepped off it, are hereafter referred to as "condition 2" errors. A third error type was present in the task:

Arm location	Action	Type	Probability
Not above target	Move towards target	Correct	0.7
	Move further from target	Error (Condition 1)	0.2
	Grab	Error	0.1
Above target	Grab	Correct	0.65
	Step off target	Error (Condition 2)	0.35

Table 1. Action probabilities for the Claw Observation Task. Note that correct actions and grabbing errors were not considered as a part of this study, as the robot would always have information about whether it had performed a lateral movement or a grab action.

a "grab error", when the robot grabbed a non-target ball. These errors occurred from a different type of movement than condition 1 and 2 errors, which both occurred as a result of lateral movements. The robot would always have information about whether it had made a lateral movement or a grab action. As such, in a BCI application, there would be no need to differentiate grab errors against other error types using EEG. Standard error detection applied following a grab action would be enough to identify them. For this reason, grab errors were not considered as a part of this study. The score was only updated after a "grab" action, and not after lateral movements (including either "condition 1" or "condition 2" errors), therefore no points were directly gained or lost as a result of either error condition. Considering this, together with the fact that each error was of the same magnitude (1 step), we considered them to be of similar severity.

Participants were asked to observe blocks, with breaks of as long as they wished between blocks, until they reported their concentration levels beginning to decrease. Most participants observed 6 blocks of trials. However, four participants observed 3-5 blocks, and three participants observed 7-8 blocks.

2.3. Data Analysis

For both tasks, EEG data were first resampled to 64Hz. In order to do this trials were first upsampled, then filtered using a least squares linear phase anti-aliasing FIR filter with a lowpass cutoff of 32Hz. The filtered data were then downsampled by averaging across data points, and initial data points from the output of filtering were removed to compensate for the delay introduced by the linear phase filter. After resampling, data were band-pass filtered from 1Hz to 10Hz, as ErrP components have been shown to occur at low frequencies [1, 2]. Event related spectral perturbation plots confirmed that activity for these tasks occurred predominantly in low frequencies (see Supplementary Figure 1). For the EADT, trials were included in cases where the error was followed by a secondary mouse click to indicate the participant's awareness of their error. Trials were extracted from a time window of -300ms to 700ms, relative to the commission of each error (i.e. the initial, erroneous mouse click). Previous literature has shown evidence that participants' EEG may show signs of an error response before they commit the error [12]. As such, the EADT time window began before error commission. Errors of which the participants were unaware were not considered as part of the main investigations of this study. As the COT involved errors committed by the machine, rather than the human, it would not have been pertinent to consider signals prior to error commission. Therefore, for the COT, trials were extracted from a time window of 0ms to 1000ms, relative to the movement of the virtual robot. Each extracted error trial was baseline corrected relative to a period of 200ms immediately before the presentation of its related stimulus. Artefact rejection was performed by discarding any trials in which the range between the highest and lowest amplitudes, in any channel, was greater than 100μV. In EADT data, a mean of 1.9 colour condition trials and a mean of 3.0 repeat condition trials were rejected per participant, from overall means of 22.2 and 32.5 trials per participant for the two conditions respectively. In COT data, a mean of 2.0 trials from condition 1 and a mean of 0.7 trials from condition 2 were rejected per participant, from overall means of 48.8 and 23.4 trials per participant for the two conditions respectively. Further to this, independent component analysis (ICA) was performed on the pooled trials from all participants combined, for each task. Components resembling EOG artefacts, as identified by visual inspection of topographic maps, were filtered out of the data. Thus, one component was removed from the data related to each task, from a total of 64 components for the EADT and 20 components for the COT. The remaining components for each task were then recombined.

Grand average time domain ErrP data were plotted using the extracted trials, showing the mean voltage ± 1 standard error of the following comparisons: EADT colour condition vs repeat condition in young adults, EADT colour condition vs repeat condition in older adults, and COT condition 1 vs condition 2 in all participants. A small number of trials were excluded from the grand average time domain plots for the EADT, where the initial click had occurred at least 550ms after the presentation of the stimulus. This was due to the fact that longer reaction times could result in the presentation of stimulus *n+1*, which could occur 1250ms after stimulus *n* in the EADT, occurring within the time window (-300ms to 700ms, relative to the click) of stimulus *n*, and so the inclusion of these trials could have contaminated the late part of the grand average data with responses to these following stimuli. In total, 14 out of 717 colour condition trials and 12 out of 1181 repeat condition trials were excluded from these plots for this reason.

Peak analysis was performed in order to identify the latencies at which ERN and Pe occurred in the ErrP data. ErrP signals are known to be associated with midline electrodes [8]. Visual inspection of time domain ErrP and topographical plots showed high positive Pe activity around the central midline across all tasks and age groups, with the most notable amplitude difference between the classes being visible in Cz time domain data. As such, electrode site Cz was chosen as the most suitable channel for peak analysis for this study. In each task, this peak analysis was carried out on the grand average ErrP waveform related to each error condition, and also for the grand average ErrP of all trials of the two error conditions pooled together. In the EADT, the analysis was carried out seperately for each age group. For each group, the data were first averaged, and then peaks were identified in the resultant waveform. The ERN was identified as most prominent negative peak, and Pe as the highest positive peak, occurring in specific time windows. Time windows for ERN were -100ms to 200ms in the EADT, and 0ms to 300ms in the COT. Time windows for Pe were 0ms to 400ms in the EADT, and 100ms to 600ms in the COT. These time windows were selected based on a visual inspection of the time-domain data; ERN windows started slightly before the start of the negative deflection in grand average plots and centred on the negative peaks, and Pe windows began just before the start of the positive deflection and ended once amplitudes had returned approximately to baseline levels. As discussed earlier in this section, evidence has shown that some participants may show signs of an error response before they commit the error [12], hence the ERN time window in the EADT beginning 100ms prior to error commission. To check for statistically significant differences in peak latencies across error conditions, the same peaks were identified in the average time domain data for each individual participant with at least 12 trials per condition and at least 40 trials in total, as previous literature has suggested that a minimum of 12 trials are required to achieve a reasonable level of temporal stability of ERN and Pe, and that temporal stability increases with the number of trials [20]. Wilcoxon signed-rank tests were then carried out on these data, comparing the latencies identified in each of these participants' average time domain waveforms for the two conditions. To check for statistically significant differences in peak amplitude, the amplitude was calculated in each of these participants' average waveforms for each condition, in a 50ms window surrounding the ERN and Pe peaks identified in grand average data (from peak -25ms to peak + 25ms). Wilcoxon signed-rank tests were carried out to compare these amplitudes. Furthermore, the build-up rate of the Pe was calculated for the average waveform of each participant, in each error condition, for both tasks. This was achieved by performing a linear regression on a time window, 100ms in duration, ending at the identified Pe peak. This gives an indication of the rate at which the amplitude is increasing up to the peak. Wilcoxon signed-rank tests were carried out to check whether the build-up rates of the different error conditions varied in a statistically significant way.

Topographical maps were then plotted for each error condition, using the same time windows. All topographical maps for a given task used the same scale, from the minimum value to the maximum values across all grand averages.

While the main focus of this study was on errors of which the participants were aware, a brief analysis was carried out to compare the number of "aware errors" (errors followed by an awareness click) vs "unaware errors" (errors not followed by an awareness click) in the EADT. The percentage of errors of which each participant was aware was calculated for each error condition in each task. Wilcoxon signed-rank tests were carried out in order to check whether there was any significant difference between awareness rates for the various conditions.

2.4. Classification

Broadly, the same classification protocol was followed for all participants of both tasks. However, different time windows were used to extract features for the two tasks. The protocol is described in this section.

2.4.1. Preprocessing 20 electrode channels were available in the COT data (F7, F3, Fz, F4, F8, FC1, FC2, T7, C3, Cz, C4, T8, CP1, CP2, P3, Pz, P4, PO7, PO8, and Oz). As such, these 20 channels were used for single-trial classification of the both tasks. As with the neurophysiological analysis, data for classification were resampled to 64Hz and band-pass filtered between 1Hz and 10Hz. In the EADT, trials were extracted from -100ms to 400ms, relative to the commision of errors (i.e. the erroneous click), in cases where the participants showed awareness of the error. In the COT, trials were extracted from 100ms to 700ms, relative to the virtual robot's movement. These time windows were selected based on visual inspection of grand average time domain data for each task, aiming to encapsulate the areas which indicated differences between the amplitudes of responses to the two conditions. Trials were baseline corrected to a period of 200ms immediately before presentation of the stimulus, and artefact rejection was performed to remove any trials with a range of greater than 100μV between the highest and lowest amplitude in any of the channels being used for classification. After this, remaining EOG artefacts were cleaned using ICA, as previously described in section 2.3.

As discussed in section 2.3, temporal stability of the ERN and Pe have been shown to increase with the number of trials, with a minimum of 12 trials being recommended to achieve a reasonable level of stability [20]. As such, for the purpose of single-trial classification, we only included participants who had generated at least 12 trials per error condition, and a minimum of 40 trials overall.

Due to the experimental setup of the EADT, which involved participants clicking a mouse to confirm error awareness, motor movements would sometimes occur less than 400ms after error commission, i.e. within the classification time window. As such, it was important to ensure that the classification was based on error responses rather than sensorimotor rhythms. To this end, two analyses were carried out on the latency between error commission and awareness confirmation in the various error conditions. Firstly, for each participant, a Fisher's exact test was carried out on the number of trials that did contain awareness confirmation within the time window used for classification vs the number that did not, in each of the two error conditions. This test was to check, for each participant, whether significant classification could feasibly be achieved based on the presense or absence of sensorimotor rhythms. Secondly, for each participant in each task, Welch's t-test was carried out, comparing the latencies at which participants confirmed their error awareness, between the two error conditions. The latencies of mouse clicks, confirming error awareness, were included in the t-test if they occurred within the classification time window (-100ms to 400ms). Clicks outside this window were ignored as they were not deemed to have a potential effect on classification. The t-test was automatically marked as not significant if there were no awareness confirmations within the classification epoch. The purpose of this test was to act as a guide, for each participant, as to whether significant classification could feasibly have been achieved based on differences in the time at which awareness-based sensorimotor rhythms occurred. We were mindful that the classification results of this study could have been unfairly biased if we had included any participants for whom classification may have been possible due to differences between motor signals across the two conditions. Therefore, participants for whom a significant result ($p < 0.05$) was recorded, in either the Fisher's exact test or the t-test, were discarded from the classification phase.

After preprocessing, 25 participants remained to be used in the classification phase from the EADT (20 young, 5 older), and 14 remained from the COT (8 asked to count errors, 6 not asked to count errors).

2.4.2. Feature Extraction Our EEG data, having been resampled at 64Hz, contained 33 time points per trial in the EADT and 40 time points per trial in the COT. If we were to consider all available time domain data, there would have been a total of 660 features (20 channels × 33 time points) or 800 features (20 channels × 40 time points) to describe each trial. Although we employed a minimum cutoffs of 12 trials per condition and 40 overall trials, many participants still had relatively few trials per class. With the number of features given by the full time domain data greatly outweighing the number of trials per condition, it was clear that the curse of dimensionality could cause problems if we attempted to classify based on all available time domain data [21].

Our classification was performed using stepwise linear discriminant analysis (SWLDA), as described in section 2.4.3. However, the feature selection inherent in SWLDA is relatively sophisticated, and less complex methods are known to be less susceptible to overfitting [22]. Therefore, we opted to reduce the dimensionality by using a simpler first step for preliminary feature extraction. This allowed the SWLDA to be applied to a small number of highly discriminative selected features.

For each participant, the preliminary step was carried out as follows: For each time-domain feature (i.e. each time point in each channel), there were a set of training data points. Each point had an amplitude and an associated class label. A linear correlation coefficient was calculated between these amplitudes and class labels, resulting in each feature having an associated correlation coefficient. The correlation coefficients acted as a simple indication of how strongly related the amplitude was to the class labels in a given feature, and thus how separable the classes may be based on the amplitude. In each channel, the feature with the largest absolute correlation coefficient was selected. This meant that each trial was represented by 20 features.

2.4.3. Stepwise Linear Discriminant Analysis Implementation In order to classify the data based on the most pertinent subset of the extracted features, SWLDA was chosen as our classification approach, since it has previously been shown to perform well in feature selection and classification of EEG data [23, 24, 25]. Stepwise regression was performed to select which features would be included in the model. Initially, an empty model was created. At each step, a regression analysis was performed on models with and without each feature, producing an F-statistic with a p-value for each feature. If the p-value of any feature was < 0.025, the feature with the smallest p-value would be added. Otherwise, if the p-value of any features already in the model had risen to > 0.075 at the current step, the feature with the largest p-value would be removed from the model. This process continued until no feature's p-value reached the thresholds for being added to, or removed from, the model. If no features were added to the model at all, a single feature with the smallest p-value would be selected. Training and test trials were then reduced to the selected features. The class with the fewest training trials was oversampled in order to ensure that training occurred with an equal number of trials per class. A linear classification model was then trained and tested.

All classifiers were trained and tested using leave-one-out cross validation. For each iteration, one trial was selected as the test sample, and all the other trials were used as the training samples. Feature extraction and training of the stepwise linear model were then performed on the training samples. The model was then tested on the test sample. This process was repeated until each trial had been selected as the test sample. To test statistical significance of the classification, a right-tailed Fisher's exact test was performed on the confusion matrix of each participant's results. As the individual participants were independent, no p-value adjustments were necessary [26]. Therefore, classification for an individual was deemed to be significant if the p-value was less than 0.05. In order to test the significance at a group level, individual p-values were combined into a group p-value using Fisher's method [27, 28]. To test whether there was any difference in the efficacy of the classification strategy across age groups, Welch's t-test was carried out comparing the overall accuracies of all young adults with those of older adults in the EADT.

3. Results

3.1. Neurophysiological Analysis of Error-Related Potentials

Peak analysis was used to identify ERN and Pe latencies based on the grand average Cz time domain waveform for each combination of task, condition, and age group. The identified latencies are shown in **Table 2**.

Wilcoxon signed-rank tests were carried out to check for statistically significant differences in the ERN and Pe amplitudes and latencies generated in response to the different error conditions, as discussed in section 2.3. The results of these tests are shown in **Table 3**.

Grand Average Peak Latency Identification			
ERN			
	Colour Condition	**Repeat Condition**	**Pooled Trials**
EADT, young	44ms	44ms	44ms
EADT, older	59ms	59ms	59ms
	Condition 1	**Condition 2**	**Pooled Trials**
COT	78ms	141ms	78ms
Pe			
	Colour Condition	**Repeat Condition**	**Pooled Trials**
EADT, young	216ms	247ms	231ms
EADT, older	200ms	247ms	215ms
	Condition 1	**Condition 2**	**Pooled Trials**
COT	281ms	344ms	328ms

Table 2. ERN and Pe latencies, relative to error commission, as identified by peak analysis on the grand average channel Cz time domain waveform. The most prominent negative peak, between -100ms and 200ms in the EADT, or between 0ms and 300ms in the COT, relative to error commission, was selected as the ERN. The highest positive peak, between 0ms and 400ms in the EADT, or between 100ms and 500ms in the COT, relative to error commission, was selected as the Pe.

3.1.1. Error Awareness Dot Task In the grand average ErrP of young adults in the EADT, responses to both conditions showed ERN with latencies of 44ms, as can be seen in **Figure 3** (blue and red lines). Wilcoxon signed-rank test showed no significant difference between the amplitudes of these ERNs (see **Table 3**), and showed no significant difference between the ERN latencies related to the two conditions, based on peaks identified in Cz data of each participant's average waveform ($p = 0.42$). However, there was a clear difference between the error conditions in the Pe. While the latencies of the Pe in response to the two conditions showed no significant difference ($p = 0.47$), the amplitudes of the Pe were increased in the colour condition, compared to the repeat condition ($p = 0.003$). The build-up rate of the Pe was also greater in the colour condition than the repeat condition, and a Wilcoxon signed-rank test showed that this distinction was statistically significant ($p = 0.001$). Topographical maps confirmed negative fronto-central activity during the ERN, and positive centro-parietal activity during the Pe, in response to both error conditions, as shown in **Figure 4 a, b, e, and f**.

Participants of all ages indicated awareness of a higher proportion of colour condition errors (mean 89.3%, SD 17.7%) than repeat errors (mean 76.4%, SD 23.5%). A Wilcoxon signed-rank test showed that this difference was significant ($p = 8.7 \times 10^{-8}$).

In the older adults' EADT data, early positivity stalls the ERN, and some

Condition Comparisons

	ERN Amplitude		**ERN Latency**	
	p-value	**Significant**	**p-value**	**Significant**
EADT, young	0.42	No	0.91	No
EADT, older	0.94	No	0.69	No
COT	0.22	No	0.72	No

	Pe Amplitude		**Pe Latency**	
	p-value	**Significant**	**p-value**	**Significant**
EADT, young	*0.003*	*Yes*	0.47	No
EADT, older	*0.016*	*Yes*	0.15	No
COT	0.19	No	*0.032*	*Yes*

Table 3. Wilcoxon signed-rank test results from comparisions of peak amplitudes and latencies of colour condition vs repeat condition (EADT) and condition 1 vs condition 2 (COT). Comparisons were performed at ERN and Pe sites, in young adults and older adults, using electrode site Cz. Amplitude comparisons were based on the mean amplitude recorded, for each subject, in ERN and Pe time windows 50ms in duration, from -25ms to 25ms relative to the peak latencies identified by grand average peak analysis. Latency comparisons were based on the peak latencies identified from each participant's average time domain data for each condition.

differences between the error conditions can be seen in the time domain data prior to error commission, as shown in **Figure 3** (green and brown lines). However, the difference between responses to the conditions was not found to be significant in older adults at the ERN. As with younger adults, the latencies of the ERN and Pe showed no significant difference ($p = 0.69$ and $p = 0.15$, respectively). While the build-up rate of the Pe was appeared to be steeper in response to the colour condition than the repeat condition, a Wilcoxon signed-rank test did not find this to be significant in older EADT participants ($p = 0.25$). Again, the most notable difference between the two error conditions was the greater amplitude of the Pe in the colour condition, as compared to the repeat condition ($p = 0.016$).

Both ERN and Pe peaks were observed to be more positive in older adults than young adults, in response to both error conditions. Welch's t-tests confirmed that that these age-related amplitude differences were statistically significant ($p = 2.1 \times 10^{-15}$ for colour condition related ERN amplitudes, $p = 5.4 \times 10^{-8}$ for colour condition related Pe amplitudes, $p = 3.1 \times 10^{-20}$ for repeat condition related ERN amplitudes, and $p = 5.4 \times 10^{-13}$ for repeat condition related Pe amplitudes).

The typical fronto-central negativity cannot be identified by visual inspection of the topographical maps of the ERN in response to either error condition for older adults' EADT data (**Figure 4 c-d**). A posterior-anterior shift in aging (PASA) has

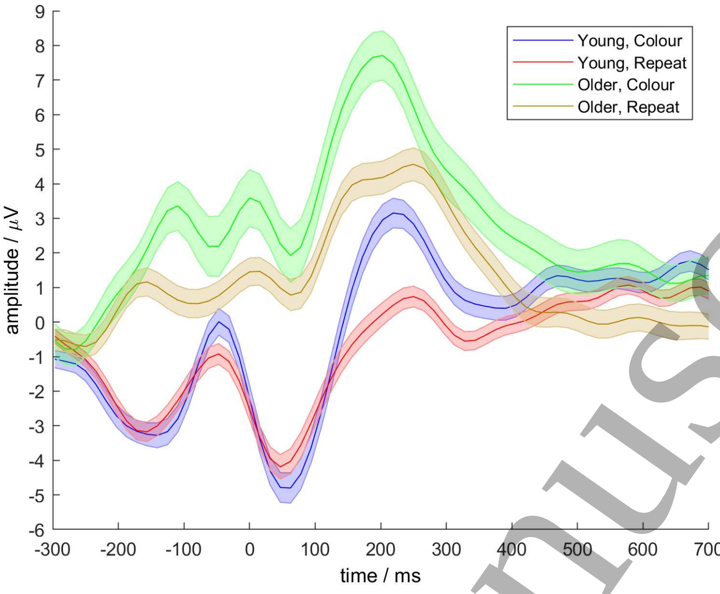


Figure 3. Grand average time domain EADT data at electrode site Cz. Time shown is relative to error commission. Central lines represent mean signals. Shaded areas cover 1 standard error. Blue lines show colour condition data from young adults. Red lines show repeat condition data from young adults. Green lines show colour condition data from older adults. Brown lines show repeat condition data from older adults.

been reported in previous literature [29, 30] and is evident here in the Pe related to both conditions of the EADT. As discussed previously, the most positively active areas during the Pe are centro-parietal in young adults, as shown in **Figure 4 e-f**. In older adults, this shifts toward more fronto-central activity, in both the colour condition and the repeat condition, as can be seen in **Figure 4 g-h**. Indeed, the electrode sites with the highest grand average Pe amplitudes in young adults were CPz & Cz for the colour condition, and CPz & Pz in the repeat condition. In older adults, the highest grand average Pe amplitudes were found at electrode sites FCz and FC1, for both error conditions.

Across all EADT participants, mean amplitudes for individual channels in the selected time windows ranged from $-11.1\mu V$ to $8.1\mu V$, and their associated standard deviations ranged from $0.04\mu V$ to $1.3\mu V$. Further topographical maps showing the standard deviation from the mean at each channel in the EADT are shown in Supplementary Figure 3a-h.

3.1.2. Claw Observation Task Time domain data related to responses to the COT can be seen in **Figure 5**. Here, no statistically significant difference was found between either the latency or amplitude of the ERN ($p = 0.72$ and $p = 0.22$, respectively). In contrast to the EADT, neither the amplitude of the main Pe peak, nor the build-up rate of the Pe showed signifigant differences ($p = 0.19$ and $p = 0.60$, respectively). However, the latencies of the Pe peaks, at their highest points, were found to be significantly

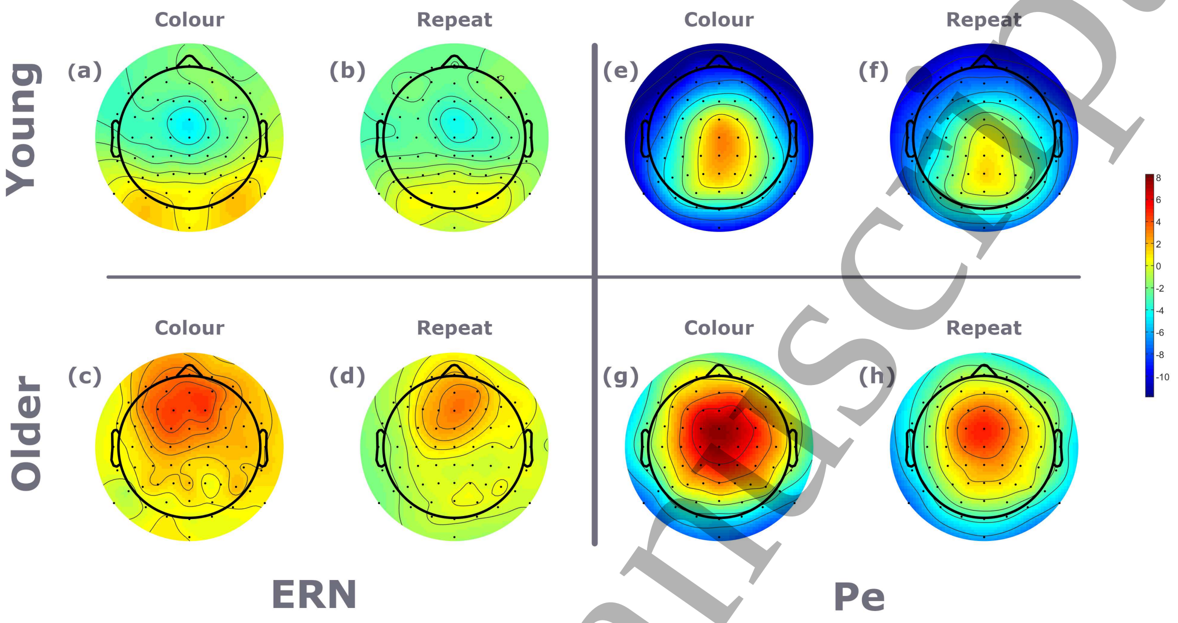


Figure 4. Grand average topographical maps of EADT data. Maps were plotted based on a 50ms window surrounding the peaks identified as ERN and Pe from grand average data across all participants. Plots shown represent (a) ERN in the colour condition in young adults, (b) ERN in the repeat condition in young adults, (c) ERN in the colour condition in older adults, (d) ERN in the repeat condition in older adults, (e) Pe in the colour condition in young adults, (f) Pe in the repeat condition in young adults, (g) Pe in the colour condition in older adults, and (h) Pe in the repeat condition in older adults.

different ($p = 0.032$), with the Pe in responses to condition 2 peaking later than that related to condition 1.

A secondary component of the Pe also appeared to be present in the grand average COT data, and appeared to be more prominent in response to condition 2 than condition 1, followed by a difference in grand average amplitudes. We identified that the maximum difference here occurred at 538ms (see Supplementary Figure 4 for illustration), and performed a further Wilcoxon signed-rank test on the amplitudes of the two conditions in the 50ms window surrounding this latency. The difference in amplitudes at this point was found to be statistically significant ($p = 6.1 \times 10^{-4}$).

Topographical maps showed broad, slightly negative amplitudes across the brain during the ERN of the COT, in response to both error conditions, as shown in **Figure 6 a and c**. Slightly more positive amplitudes can be seen in fronto-central regions in response to condition 1. During the Pe, strong positive activity can be seen in central and centro-parietal regions, as shown in **Figure 6 b and d**.

Mean amplitudes for individual channels in the time window ranged from $-1.1\mu V$ to $5.4\mu V$, and their associated standard deviations ranged from $0.01\mu V$ to $0.8\mu V$. Further topographical maps showing the standard deviation from the mean at each channel in

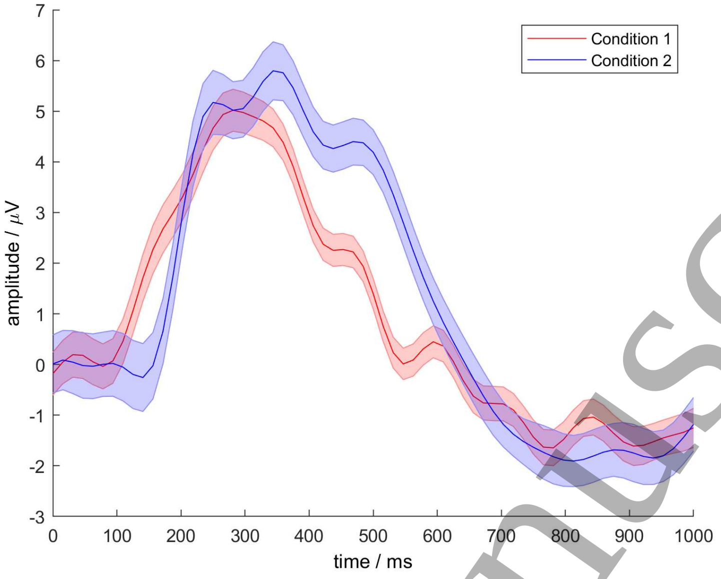


Figure 5. Grand average time domain COT data at electrode site Cz. Time shown is relative to the erroneous movement of the robot. Central lines represent mean signals. Shaded areas cover 1 standard error. Red line shows condition 1 data from all participants. Blue line shows condition 2 data from all participants.

524 the COT are shown in Supplementary Figure 3i-l.

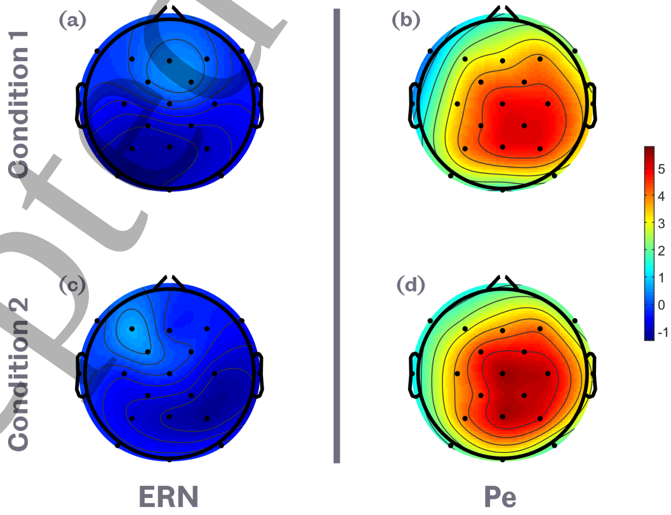


Figure 6. Grand average topographical maps of COT data. Maps were plotted based on a 50ms window surrounding the peaks identified as ERN and Pe from grand average data across all participants. Plots shown represent (a) ERN in the condition 1, (b) ERN in the condition 2, (c) Pe in condition 1, and (d) Pe in condition 2.

3.2. Classification of EADT Errors

The classification accuracies achieved for each individual participant in the EADT are shown in **Table 4**. The mean overall accuracy for all EADT participants was 65.2%. Amongst young adults, mean overall accuracy was 63.7%, and for older adults it was 71.3%. Mean colour condition accuracy was 60.4% for all participants, 59.4% for young adults, and 60.4% for older adults. The mean accuracy of the repeat condition was 67.6% for all participants, 66.0% for young adults, and 74.0% for older adults. Trained classification models for the EADT included a mean of 3.7 ± 1.3 features. Generally, more features were selected from posterior regions of the brain than anterior regions, echoing the heightened activity, varying in amplitude across the two classes, that was shown in these regions. A Wilcoxon signed-rank test was used to compare the average number of features selected per channel, for each participant, in more anterior channels (fronto-central channels and further anterior) against those in more posterior channels (centro-parietal channels and further posterior). The results showed the average number of selected features per channel was significantly higher in the posterior region compared to those in the anterior region ($p = 4.9 \times 10^{-4}$). At an individual level, features were often selected where the subject-average amplitude displayed a relatively large differences between the two classes. Supplementary Figure 5 contains a further breakdown of feature selection rates, including an example for an individual EADT participant.

Statistically significant separation of the error conditions ($p < 0.05$) was found, using Fisher's exact tests, for 17 of the 25 participants overall (68.0%). Statistical significance was achieved for 13 of the 20 young adults (65.0%), and 4 of the 5 older adults (80.0%). At a group level, the classification results were found to be statistically significant in each age group ($p = 1.6 \times 10^{-16}$ for young adults and $p = 3.2 \times 10^{-11}$ for older adults) and overall ($p = 2.7 \times 10^{-25}$).

The overall accuracies of young adults were compared with those of older adults using Welch's t-test. The result did not show any significant difference ($p = 0.16$). While Welch's t-test is considered to be reliable in dealing with unequal sample sizes [31, 32], it should be noted that only 5 older adults remained in the single-trial classification, which may mean that this finding should be treated with a measure of caution.

3.3. Classification of COT Errors

The classification accuracies achieved for each individual participant in the COT are shown in **Table 5**. The mean overall accuracy for all COT participants was 65.6%. Mean accuracy for condition 1 was 69.5%, and the mean accuracy for condition 2 was 57.4%. Welch's t-test showed no significant difference in participants accuracy depending on whether or not they were asked to keep count of the errors ($p = 0.80$, see Supplementary Table 1). Trained classification models for the COT included a mean of 2.9 ± 1.5 features. At a population level, it was difficult to discern clear patterns of which features were selected. However, as in the EADT, an individual level features were often selected where there was a relatively large difference between the subject-

Age Group	Subject	# Colour Trials	# Repeat Trials	Colour Accuracy	Repeat Accuracy	Overall Accuracy	Significant	p-value
Young	1	27	27	55.6%	48.1%	51.9%	No	0.5
	2	34	42	58.8%	69.0%	64.5%	Yes	0.014
	3	15	35	60.0%	74.3%	70.0%	Yes	0.024
	4	29	55	65.5%	70.9%	69.0%	Yes	0.0014
	5	21	26	57.1%	61.5%	59.6%	No	0.16
	6	30	38	50.0%	60.5%	55.9%	No	0.27
	7	14	31	42.9%	58.1%	53.3%	No	0.60
	8	17	57	58.8%	74.6%	71.1%	Yes	0.012
	9	41	53	64.3%	63.0%	63.5%	Yes	0.0071
	10	33	43	57.6%	65.1%	61.8%	Yes	0.041
	11	22	34	72.2%	70.6%	71.4%	Yes	0.0017
	12	26	42	50.0%	64.3%	58.8%	No	0.16
	13	32	51	75.0%	76.5%	75.9%	Yes	4.5×10^{-6}
	14	25	46	52.0%	76.1%	67.6%	Yes	0.017
	15	25	51	68.0%	72.5%	71.1%	Yes	8.7×10^{-4}
	16	20	29	55.0%	75.9%	67.3%	Yes	0.029
	17	30	30	46.7%	56.7%	51.7%	No	0.50
	18	42	45	61.9%	51.1%	56.3%	No	0.16
	19	33	58	66.7%	67.2%	67.0%	Yes	0.0017
	20	28	45	69.0%	64.4%	66.2%	Yes	0.0049
Older	21	17	47	41.2%	61.7%	56.3%	No	0.52
	22	*45*	*33*	*80.0%*	*81.8%*	*80.8%*	*Yes*	$\mathit{4.8 \times 10^{-8}}$
	23	21	47	76.2%	63.8%	67.6%	Yes	0.0024
	24	19	35	63.2%	80.0%	74.1%	Yes	0.0021
	25	13	46	61.5%	82.6%	78.0%	Yes	0.0034
Young	**Mean**	**27.2**	**41.9**	**59.4%**	**66.0%**	**63.7%**	**65.0%**	**Group p-value** $\mathbf{1.6 \times 10^{-16}}$
	SD	*7.7*	*10.3*	*8.6%*	*8.3%*	*7.2%*		
Older	**Mean**	**23.0**	**41.6**	**64.4%**	**74.0%**	**71.3%**	**80.0%**	**Group p-value** $\mathbf{3.2 \times 10^{-11}}$
	SD	*12.6*	*7.0*	*15.3%*	*10.3%*	*9.8%*		
All	**Mean**	**26.4**	**41.8**	**60.0%**	**67.6%**	**65.2%**	**68.0%**	**Group p-value** $\mathbf{2.7 \times 10^{-25}}$
	SD	*8.7*	*9.6*	*10.1%*	*9.1%*	*8.2%*		

Table 4. Single-trial classification results of EADT data. Overall accuracy calculated as the percentage of trials, of either class, correctly classified. SD refers to standard deviation. The participant for whom the highest overall accuracy was achieved is highlighted in italics. Group p-values were calculated by combining p-values using Fisher's method.

Subject	# Condition 1 Trials	# Condition 2 Trials	Condition 1 Accuracy	Condition 2 Accuracy	Overall Accuracy	Significant	p-value
1	42	27	66.7%	63.0%	65.2%	Yes	0.015
2	92	30	72.8%	53.3%	68.0%	Yes	0.0088
3	69	22	58.0%	40.9%	53.8%	No	0.63
4	43	23	65.1%	52.2%	60.6%	No	0.13
5	46	29	73.9%	65.5%	70.7%	Yes	8.2×10^{-4}
6	*30*	*14*	*86.7%*	*64.3%*	*79.5%*	*Yes*	*0.0011*
7	48	18	79.2%	72.2%	77.3%	Yes	1.8×10^{-4}
8	46	29	69.6%	62.1%	66.7%	Yes	0.0069
9	49	21	77.6%	47.6%	68.6%	Yes	0.036
10	33	19	63.6%	52.6%	59.5%	No	0.20
11	34	21	58.8%	42.9%	52.7%	No	0.56
12	39	26	64.1%	61.5%	63.1%	Yes	0.038
13	44	13	70.5%	61.5%	68.4%	Yes	0.040
14	32	22	65.6%	63.6%	64.8%	Yes	0.032
Mean	**46.2**	**22.4**	**69.4%**	**57.4%**	**65.6%**	**71.4%**	**Group p-value**
SD	*16.4*	*5.4*	*8.0%*	*9.2%*	*7.6%*		$\mathbf{1.9 \times 10^{-11}}$

Table 5. Single-trial classification results of COT data. Overall accuracy calculated as the percentage of trials, of either class, correctly classified. SD refers to standard deviation. The participant for whom the highest overall accuracy was achieved is highlighted in italics. The group p-values was calculated by combining p-values using Fisher's method.

average amplitudes of the classes. Supplementary Figure 5 contains a further breakdown of feature selection rates, including an example for an individual COT participant.

Statistically significant separation of the error conditions ($p < 0.05$) was found, using Fisher's exact tests, for 10 of the 14 participants (71.4%) in the COT. At a group level, the classification results were found to be statistically significant ($p = 1.9 \times 10^{-11}$).

4. Discussion

4.1. Distinctions in Responses by Condition and Age

Previous literature has shown that different tasks can elicit differing ErrP waveforms [33]. In some cases, distinctions have been shown in ErrPs even when the errors are committed during variants of the same task [34, 35]. Indeed, our findings are aligned with those of the previous literature on this point. Interestingly, when comparing the error conditions within each task, the key neurophysiological distinctions that we were able to identify were found in different components of the ErrP for the two tasks in this study.

In the EADT, the clearest distinction shown between the error conditions was in the amplitude of the Pe. We witnessed greater amplitudes of Pe in the colour condition than the repeat condition for both young and older adults. Previous studies, including some which were based on error awareness tasks, have shown a diminished Pe in errors of which participants are unaware, compared to errors of which they are aware [13, 14, 15, 16]. Here, in the case of the colour condition, all necessary information for the participant to know whether they have committed an error is present, on-screen, in the current stimulus. With the repeat condition, however, participants are relying on their memory of the previous stimulus to determine whether or not they have committed the error. Indeed, Wilcoxon signed-rank tests found that participants were significantly more likely to be aware that they had committed a colour condition error than a repeat condition error. While this study was focused on trials in which participants signified awareness of their errors, it is possible that participants could be more confident in their assertion of the error for some trials than others. It is possible, therefore, that the higher amplitude of the Pe in the colour conditions, compared to the repeat conditions, is due to greater certainty and confidence that an error was committed. Previous studies have also identified the build-up rate of the Pe as a marker of evidence accumulation for error detection [17]. In young adults, the build-up rate to the Pe was found to be significantly greater in the colour condition than the repeat condition. This is a further indication that a greater degree of awareness may be present in the case of colour condition errors than repeat condition errors.

Some distinctions were also noted between the different age groups in the EADT. Older participants' responses were found to generate more positive amplitudes at both the ERN and Pe latencies, for both error conditions. A posterior-anterior shift in aging was also identified in the spatial distribution of the Pe.

In the COT, the most notable difference in time domain data appeared to result from a secondary component of the Pe. This occurred at around 500ms, causing an increase in the amplitude of responses to condition 2 compared to those of condition 1 in the grand average signals. This gap remained until beyond 600ms. A Wilcoxon signed-rank test found the amplitude difference, at its widest point (538ms) to be statistically significant ($p = 6.1 \times 10^{-4}$). As discussed in section 1, secondary Pe components have previously been identified, and have been linked to conscious, evaluative processes [18, 19]. This suggests that condition 2, in which the virtual robot steps off the target, having been aligned above it, elicits stronger responses in the aware aspect of the error response.

4.2. Single-Trial Classification

Across all participants who were included in the classification stage, we achieved a mean overall accuracy of 65.2% for the EADT data, and 65.6% for the COT data. The associated standard deviations were relatively high (8.2% and 7.6% respectively) as, although statistically significant classification was not possible for some participants, high classification rates were achieved for others. Indeed, in the best cases, for both tasks, the error conditions were classified against each other with around 80% overall accuracy. Group p-values calculated using Fisher's method showed that, at a population level, statistically significant separation of the error conditions was achieved

($p = 2.7 \times 10^{-25}$ for the EADT and $p = 1.9 \times 10^{-11}$ for the COT). As a proof of concept, these classification accuracies show that it is possible to classify these subtly different error conditions, which could not be differentiated by previously explored metrics such as direction or severity, against each other using single-trial EEG.

A Welch's t-test, comparing the results of young adults with those of older adults, returned non-significant results. Though this finding should be taken tentatively, due to the small number of older participants included in the classification phase, it suggests that our chosen classification strategy is robust across different age groups, despite some age-related neurophysiological differences.

In previous literature regarding error decoding, a wide variety of classification accuracies have been reported. When classifying errors against non-errors, some studies have been able to achieve very high single-trial classification rates. For example, SVM-based classification models have been used to achieve average accuracies of 80% [36] or even above 90% [5], deep learning approach achieved average accuracy of 84% [37], and Gaussian models have been reported to achieve a high of around 90% [38].

Classification of different error conditions against each other can be considered more challenging than error vs non-error classification as the EEG signals in response to errors are expected to be more similar to each other than to the signals of non-errors. Nonetheless, some errors have been classified against each other on a single trial basis with a high level of success. In a virtual robot reaching task, performed by 2 participants, Iturrate et al. reported correct classification of left vs right sided errors with an impressive 90% accuracy [6]. Furthermore, in the same study, they were able to distinguish small vs larger errors with around 75% accuracy. Spüler and Niethammer reported an overall accuracy of 75.5% for the classification of execution errors against outcome errors (i.e. errors committed by a machine vs errors committed by a human) during a computer game task [9]. However, they did not find significant differences between movement errors occurring at different angles, highlighting the potential difficulty of differentiating errors based on subtle differences.

One of the challenges in error decoding is that data sets for error trials may be small, as errors often occur more rarely than correct actions, both in real-world scenarios [5] and experimental paradigms [39]. Small sample sizes are known to be challenging in classification problems [40, 41]. This is exacerbated when attempting error vs error classification, as the error trials are divided into still smaller groups. Indeed, for both tasks of the present study, we were able to achieve higher classification accuracy for the class with more training samples, on average.

Given the challenges of comparing such similar error conditions as the ones in this study, we believe that the results are encouraging. Separation of the error conditions was above chance level for most participants across both tasks. While mean overall classification rates did not reach the accuracy of the most successful studies discussed above, this study has shown that it is indeed feasible to classify ErrPs of different error conditions against each other based on differences in cognitive process, or in the context of differing expected actions. The fact that overall accuracy of around 80% was achieved for some participants is particularly encouraging. In future, it may be interesting to investigate the use of other classification techniques such as those discussed above, especially if larger training sets are available, with the aim of increasing classification accuracy further.

4.3. Implications for BCI

Error detection is becoming an increasingly useful aspect of BCI [2]. It has proven to be utilisable in increasing the accuracy of existing BCI control techniques, such as motor imagery [42] and P300 [43], by performing immediate error correction [44]. Furthermore, error detection has been successfully integrated into various BCI systems as feedback for reinforcement learning (RL) strategies, allowing the systems to gradually improve over time [45, 3, 4, 5]. As discussed in section 1, this creates the possibility of BCI becoming more autonomous [3, 4]. RL-based systems such as these can work effectively as long as the classification accuracy exceeds chance level [2, 3].

It has been shown, in previous literature, that different errors can elicit different ErrP waveforms [46, 47]. Recently, a few studies have begun to classify different errors using single-trial EEG, based on aspects such as the direction of the error [6], the severity of the error [6], or whether the error was committed by the human themselves or by a machine [9].

In the COT, we presented a scenario in which a virtual robot was attempting to navigate towards, and grab, a target object among several non-target objects. This scenario could be used in an error-driven BCI. Each robot action would be followed by single-trial EEG classification, to tell the robot what kind of action the human had observed. If we employed simple error detection, we would be able to tell the robot when it had made an incorrect move. However, with the error categorisation displayed in this study, an extra layer of detail could be switched on for participants with statistically significant separation. In the case of condition 1 errors, we could tell the robot that the target is in the other direction, but is not in the adjacent location. In the case of condition 2 errors, we could tell the robot precisely that the target is in the location it just stepped away from. These principles could be applied to a number of BCI-based navigation or target selection scenarios.

Investigating the EADT allowed us to provide further evidence that errors can be categorised in the absence of previously used metrics, with only subtle difference between error conditions.

Statistically significant classification accuracy was achieved for the vast majority of the participants included in the classification phase in our study. Thus, the error categorisation displayed here is accurate enough to be utilised in a BCI, for immediate and specific error correction, or as an integral part of a learning system. This opens up the potential for more detailed information to be garnered about the category of error that has occurred, thus allowing for a BCI with more effective error correction and more efficient error-driven learning.

5. Conclusion

The error conditions considered in this study were very similar to one another. Nevertheless, due to the different cognitive processes required to recognise the errors in the EADT, and the different contexts in which the errors occurred in the COT, we were able to identify differences between the grand average ErrP waveforms of the different error conditions. In the EADT, the clearest distinction between the error conditions was found in the amplitude of the Pe. The colour conditions generally elicited greater amplitudes than the repeat conditions, leading us to speculate that the increased Pe in these conditions could be due to greater certainty that an error had been committed. In the COT we found distinctions in the ERN, and in a secondary component of the Pe. These distinctions led us to speculate that participants may have had a heightened anticipation of a correct action when the virtual robot was aligned above the target, ready to grab it.

Interestingly, we were able to classify the error conditions of both the EADT and the COT, the latter of which could be directly applied in a BCI, with over 65% mean overall accuracy, and around 80% in the best cases. Classification rates were above chance level ($p < 0.05$) for most participants, of those included in the classification phase of the study, for both tasks, and group-level analysis showed the single-trial separation of the different error conditions to be highly significant overall ($p = 2.7 \times 10^{-25}$ for the EADT and $p = 1.9 \times 10^{-11}$ for the COT). The ability to classify such similar errors using single-trial EEG, as we have shown here, is very promising for the future prospect of making error-driven BCI more efficient through the acquisition of more detailed information.

We believe that the findings of this study uncover new opportunities in brain-machine interaction, pushing towards a more autonomous BCI.

Acknowledgments

This research was supported by University of Sheffield, United Kingdom, and the Engineering and Physical Sciences Research Council (EPSRC). Dr Paul Dockree was supported by an IRC Laureate grant: IRCLA/2017/306.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The authors would like to thank Dr Kianoush Nazarpour of the University of Newcastle for helpful discussions about the classification of the data, and Dr Peter R Murphy of the University Medical Center Hamburg - Eppendorf, whose work influenced the design of the EADT.

References

[1] William J. Gehring, Brian Goss, Michael G. H. Coles, David E. Meyer, and Emanuel Donchin. A neural system for error detection and compensation. *Psychol Sci*, 4:385–390, 1993.

[2] Ricardo Chavarriaga, Aleksander Sobolewski, and José del R. Millán. Errare machinale est: the use of error-related potentials in brain-machine interfaces. *Front Neurosci*, 8:208, 2014.

[3] Iñaki Iturrate, Ricardo Chavarriaga, Luis Montesano, Javier Minguez, and José del R. Millán. Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control. *Sci Rep*, 5:13893, 2015.

[4] Thorsten O. Zander, Laurens R. Krol, Niels P. Birbaumer, and Klaus Gramann. Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proc Natl Acad Sci U S A*, 113:14898–14903, 2016.

[5] Su Kyoung Kim, Elsa Andrea Kirchner, Arne Stefes, and Frank Kirchner. Intrinsic interactive reinforcement learning - using error-related potentials for real world human- robot interaction. *Sci Rep*, 7:17562, 2017.

[6] Iñaki Iturrate, Luis Montesano, and Javier Minguez. Robot reinforcement learning using eeg-based reward signals. In *Proc. IEEE International Conf. on Robotics and Automation (ICRA10)*, pages 4822–4829, Anchorage, AK, USA, 2010.

[7] P. W. Ferrez and J. del R. Millán. Error-related eeg potentials generated during simulated braincomputer interaction. *IEEE Trans. Biomed. Eng.*, 55(3):923–929, March 2008.

[8] Michael Falkenstein, Jörg Hoormann, Stefan Christ, and Joachim Hohnsbein. Erp components on reaction errors and their functional significance: a tutorial. *Biol Psychol*, 51:87–107, 2000.

[9] Martin Spüler and Christian Niethammer. Error-related potentials during continuous feedback: using eeg to detect errors of different type and severity. *Front Hum Neurosci*, 9:155, 2015.

[10] 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. *Electroencephalogr Clin Neurophysiol*, 78:447–455, 1991.

[11] Thérèse J. M. Overbeek, Sander Nieuwenhuis, and K. Richard Ridderinkhof. Dissociable components of error processing: On the functional significance of the pe vis-a-vis the ern/ne. *J Psychophysiol*, 19:319–329, 2005.

[12] Siobhán Harty, Peter R. Murphy, Ian H. Robertson, and Redmond G. O'Connell. Parsing the neural signatures of reduced error detection in older age. *Neuroimage*, 161:43–55, 2017.

[13] Redmond G. OConnell, Paul M. Dockree, Mark A. Bellgrove, Simon P. Kelly, Robert Hester, Hugh Garavan, Ian H. Robertson, and John J. Foxe. The role of cingulate cortex in the detection of errors with and without awareness: a high-density electrical mapping study. *Eur J Neurosci*, 25:2571–2579, 2007.

[14] Peter R. Murphy, Ian H. Robertson, Darren Allen, Robert Hester, and Redmond G. O'Connell. An electrophysiological signal that precisely tracks the emergence of error awareness. *Front Hum Neurosci*, 6:65, 2012.

[15] Sander Nieuwenhuis, K. Richard Ridderinkhof, Jos Blom, Guido P.H. Band, and Albert Kok. Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, 38:752–760, 2001.

[16] Tanja Endrass, Cosima Franke, and Norbert Kathmann. Error awareness in a saccade countermanding task. *J Psychophysiol*, 19:275–280, 2005.

[17] Peter R Murphy, Ian H Robertson, Siobhán Harty, and Redmond G O'Connell. Neural evidence accumulation persists after choice to inform metacognitive judgments. *Elife*, 4:e11946, 2015.

[18] Yael Arbel and Emanuel Donchin. Parsing the componential structure of posterror erps: A principal component analysis of erps following errors. *Psychophysiology*, 46:1179–1189, 2009.

[19] Tanja Endrass, Benedikt Reuter, and Norbert Kathmann. Erp correlates of conscious error recognition: aware and unaware errors in an antisaccade task. *Eur J Neurosci*, 26:1714–1720, 2007.

[20] Michael J Larson, Scott A Baldwin, Daniel A Good, and Joseph E Fair. Temporal stability of the error-related negativity (ern) and post-error positivity (pe): The role of number of trials. *Psychophysiology*, 47:1167–1171, 2010.

[21] Eamonn Keogh and Abdullah Mueen. Curse of dimensionality. In Claude Sammut and Geoffrey I.

Webb, editors, *Encyclopedia of Machine Learning*, pages 257–258. Springer, Boston, MA, USA, 2010.

[22] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: Model selection and overfitting. *Nat. Methods*, 13:703, 2016.

[23] D. J. Krusienski, E.W. Sellers, D.J. Mcfarland, T.M. Vaughan, and J.R. Wolpaw. Toward enhanced p300 speller performance. *J Neurosci Methods*, 167:15–21, 2008.

[24] Eric W. Sellers and Emanuel Donchin. A p300-based braincomputer interface: Initial tests by als patients. *Clin Neurophysiol*, 117:538–548, 2006.

[25] E. Donchin, K. M. Spencer, and R. Wijesinghe. Toward enhanced p300 speller performance. *IEEE Trans. Rehabil. Eng.*, 8:174–179, 2000.

[26] Kenneth J. Rothman. No adjustments are needed for multiple comparisons. *Epidemiology*, 1:43–46, 1990.

[27] Thomas M Loughin. A systematic comparison of methods for combining p-values from independent tests. *Comput. Stat. Data Anal.*, 47:467–485, 2004.

[28] N A Heard and P Rubin-Delanchy. Choosing between methods of combining p-values. *Biometrika*, 105:239–246, 2018.

[29] CL Grady, JM Maisog, B Horwitz, LG Ungerleider, MJ Mentis, JA Salerno, P Pietrini, E Wagner, and JV Haxby. Age-related changes in cortical blood flow activation during visual processing of faces and location. *J Neurosci*, 14:1450–1462, 1994.

[30] Simon W. Davis, Nancy A. Dennis, Sander M. Daselaar, Mathias S. Fleck, and Roberto Cabeza. Qu pasa? the posterioranterior shift in aging. *Cereb Cortex*, 18:12011209, 2008.

[31] Greame D Ruxton. The unequal variance t-test is an underused alternative to student's t-test and the mannwhitney u test. *Behav Ecol*, 17:688690, 2006.

[32] Ben Derrick, Deirdre Toher, and Paul White. Why welchs test is type i error robust. *Quant Methods Psychol*, 12:30–38, 2016.

[33] Rozhin Yousefi, Alborz Rezazadeh Sereshkeh, and Tom Chau. Exploiting error-related potentials in cognitive task based bci. *Biomed Phys Eng Express*, 5:015023, 2019.

[34] Iñaki Iturrate, Luis Montesano, and Javier Minguez. Task-dependent signal variations in eeg error-related potentials for braincomputer interfaces. *J Neural Eng*, 10:026024, 2013.

[35] Jason Omedes, Andreas Schwarz, and Luis Montesano. Factors that affect error potentials during a grasping task: toward a hybrid natural movement decoding bci. *J Neural Eng*, 15:046023, 2018.

[36] Xavier Artusi, Imran Khan Niazi, Marie-Françoise Lucas, and Dario Farina. Performance of a simulated adaptive bci based on experimental classification of movement-related and error potentials. *IEEE Trans. Emerg. Sel. Topics Circuits Syst.*, 1:480–488, 2011.

[37] Martin Völker, Robin T. Schirrmeister, Lukas D. J. Fiedere, Wolfram Burgard, and Tonio Ball. Deep transfer learning for error decoding from non-invasive eeg. In *Proc. 2018 6th International Conf. on Brain-Computer Interface (BCI)*, pages 1–6, GangWon, South Korea, 2018.

[38] Lucas C. Parra, Clay D. Spence, Adam D. Gerson, and Paul Sajda. Response error correctiona demonstration of improved human-machine performance using real-time eeg monitoring. *IEEE Trans. Neural Syst. Rehabil. Eng.*, 11:173–177, 2003.

[39] 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. *J Neurophysiol.*, 120:1107–1118, 2018.

[40] A. K. Jain and B. Chandrasekaran. Dimensionality and sample size considerations in pattern recognition practice. In P.R. Krishnaiah and L.N. Kanal, editors, *Handbook of Statistics*, volume 2, pages 835–855. North Holland, Amsterdam, Netherlands, 1982.

[41] Sarunasj Raudys and Anilk Jain. Small sample size effects in statistical pattern recognition - recommendations for practitioners. *IEEE Trans. Pattern Anal. Mach. Intell.*, 13:252–264, 1991.

[42] P. W. Ferrez and J. del R. Millán. Simultaneous real-time detection of motor imagery and error-related potentials for improved bci accuracy. In *Proc. 4th International Brain-Computer*

Interface Workshop and Training Course, pages 197–202, Graz, Austria, 2008.

[43] Nico M Schmidt, Benjamin Blankertz, and Matthias S Treder. Online detection of error-related potentials boosts the performance of mental typewriters. *BMC Neurosci.*, 13:19, 2012.

[44] Thorsten O. Zander, Christian Kothe, Sabine Jatzev, and Matti Gaertner. Enhancing human-computer interaction with input from active and passive brain-computer interfaces. In Desney S. Tan and Anton Nijholt, editors, *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*, pages 181–199. Springer London, London, 2010.

[45] Anna Buttfield, Pierre W. Ferrez, and José del R. Millán. Towards a robust bci: Error potentials and online learning. *IEEE Trans. Neural Syst. Rehabil. Eng*, 2:164–168, 2006.

[46] Peter S Bernstein, Marten K Scheffers, and Michael G H Coles. Where did i go wrong? a psychophysiological analysis of error detection. *J Exp Psychol Hum Percept Perform*, 21(6):1312–1322, 1995.

[47] 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, 2018.