Received: 21 March 2023 | Revised: 25 July 2023 | Accepted: 8 August 2023

DOI: 10.1111/psyp.14434

PSYCHOPHYSIOLOGY SPR WILEY

ORIGINAL ARTICLE

Decoding agency attribution using single trial error-related brain potentials

Alba Gomez-Andres[1,2] | **Xim Cerda-Company**[1,2,3] | **David Cucurell**[1,2] | **Toni Cunillera**[2] | **Antoni Rodríguez-Fornells**[1,2,4,5]

[1]Cognition and Brain Plasticity Group [Bellvitge Biomedical Research Institute-IDIBELL], Barcelona, Spain

[2]Department of Cognition, Development and Educational Psychology, University of Barcelona, Barcelona, Spain

[3]Universitat Autonoma de Barcelona, Barcelona, Spain

[4]Institute of Neurosciences (UBNeuro), University of Barcelona, Barcelona, Spain

[5]Catalan Institution for Research and Advanced Studies, ICREA, Barcelona, Spain

Correspondence
Antoni Rodríguez-Fornells, Cognition and Brain Plasticity Group [Bellvitge Biomedical Research Institute-IDIBELL], L'Hospitalet de Llobregat, Barcelona, Spain.
Email: arfornells@gmail.com

Funding information
Ministerio de Economía y Competitividad, Grant/Award Number: BES-2016-078889, PSI2015-69178-P and PSI2016-79678-P

Abstract

Being able to distinguish between self and externally generated actions is a key factor influencing learning and adaptive behavior. Previous literature has highlighted that whenever a person makes or perceives an error, a series of error-related potentials (ErrPs) can be detected in the electroencephalographic (EEG) signal, such as the error-related negativity (ERN) component. Recently, ErrPs have gained a lot of interest for the use in brain–computer interface (BCI) applications, which give the user the ability to communicate by means of decoding his/her brain activity. Here, we explored the feasibility of employing a support vector machine classifier to accurately disentangle self-agency errors from other-agency errors from the EEG signal at a single-trial level in a sample of 23 participants. Our results confirmed the viability of correctly disentangling self/internal versus other/external agency-error attributions at different stages of brain processing based on the latency and the spatial topographical distribution of key ErrP features, namely, the ERN and P600 components, respectively. These results offer a new perspective on how to distinguish self versus externally generated errors providing new potential implementations on BCI systems.

KEYWORDS
decoding, EEG, ERPs, error-related potentials, sense of agency, support vector machine

1 | INTRODUCTION

In our changeable and uncertain world, the ability to monitor and evaluate our actions is crucial for self-regulation. Successful goal achievement requires the ability to distinguish between events caused by ourselves or by another agent, promoting behavioral adaptation whenever unintended action outcomes, such as errors, occur (Rabbitt, 1966). Voluntary actions are accompanied by the implicit and automatic feeling of being the authors of one's movements and their consequences, a feeling/experience known as the sense of agency (SoA). Several

Alba Gomez-Andres and Xim Cerda-Company contributed equally to this study.

behavioral [Sato & Yasuda, 2005; Tsakiris et al., 2006, see David et al. (2008) for a review] and neuroimaging (David et al., 2007; Farrer et al., 2003; Farrer & Frith, 2002, for a recent meta-analysis see Zito et al., 2020) studies have addressed the cognitive architecture and neural basis for the SoA, highlighting the influence of prediction-outcome consistency on the SoA attribution (Frith et al., 2000; Miall & Wolpert, 1996; Synofzik et al., 2008). Importantly, the ability to distinguish self versus externally generated actions lies at the heart of the concept of agency and individual responsibility in our society, concerning important issues such as moral and legal status of actions (Haggard & Tsakiris, 2009).

External factors, such as mechanical issues, other human agents' interventions, or environmental elements can be favorable or unfavorable for successful goal achievement, therefore, influencing the processing of action selection and adaptation. Intriguingly, as a result of the new improvements in biomedical and biotechnology research, we not only re-use body parts of other bodies for medical treatment but we also begin to experience new mind–body relationships as, for example, the possibility to control other bodies (e.g., avatars and robots). For example, BCI systems are being increasingly used as assistive devices in neurorehabilitation [see Tonin & Millán, 2021 for a recent review on BCI for robotic devices]. Stroke patients with motor impairments may benefit from these technological developments, using BCI-based exoskeletons (Liu et al., 2017; Zhang & Huang, 2018), prosthetics (Rotermund et al., 2006), spellers (Manyakov et al., 2012; Margaux et al., 2012), or robotic systems (Bhattacharyya et al., 2014; Rakshit et al., 2020). These new interactions confront us with new challenges regarding body(ies)–mind relationships that raise important issues concerning the moral and legal status of actions. From a classical legal definition (English Habeas Corpus Act, 1675), we are fully responsible and have authority and property rights over our own body. In this sense, the actions of our own body are attributed to our agency and we, therefore, have direct legal responsibility for those actions. It is, however, more questionable the extent to which our feeling of agency and legal consequences would be the same when governing different bodies, when our surrogate body might make independent decisions or carry out actions that cause errors with drastic consequences that were not planned or caused directly by you. Indeed, the possibility of hijacking your surrogate body and changing its actions might be easier than trying to influence directly in your brain and decision-making. How is it then possible for humans to distinguish whether an action of a non-human agent has been triggered by the mind that is supposed to be governing it? Is it distinguishable from the neural activity? Is our brain able to attribute agency in both types of errors, self- or external-induced errors? Is the agency attribution encoded in the neural activity?

In the present study, we address this issue implementing new decoding techniques in order to assess to which extent single-trial EEG activity contains reliable decodable information about the attribution of errors ("mine" vs. "external"). The possibility to access this information at the level of a trial (each real action) using decoding techniques will allow in the future to monitor brain activity while interacting with external agents and circumvent some of the problems regarding moral responsibility commented above.

Several ERP studies have reported already that erroneous actions elicit distinct neural responses (family of error-related potentials, i.e., ErrP), mostly using measures of the average time course of brain EEG activity. For example, the Error-Related Negativity (ERN) (Falkenstein et al., 2000; Gehring et al., 1993; Taylor et al., 2007), is a very robust and reliable negative deflection observed at frontomedial locations during response-locked averaged EEG recordings, and which is elicited immediately after an error has been committed (50–100 ms after error commission). Additionally, a similar ERN modulation has also been recorded during the observation of incorrect actions performed by another agent (i.e., 'observational' ERN), depicting a lower amplitude than the ERN for self-generated errors and a later occurrence (van Schie et al., 2004). Interestingly, our group also recently identified different ErrP signatures when participants observed external errors induced in their own body (embedded in an avatar, Padrao et al., 2016; a negative modulation at 400 ms) or observing their own hands committing an error that was not their own. In this last case, a positive modulation (referred here as P600) was observed for external versus self-errors (Gomez-Andres et al., 2023).

Crucially for the purpose of the present article, it has been previously observed that ErrP can be reliably decoded on a single-trial basis (Chavarriaga et al., 2014; Iturrate et al., 2015; Kim et al., 2017; Usama et al., 2021; Zander et al., 2016; see Kumar et al., 2019 for a review about decoding ErrP), thus allowing their implementation for BCI systems, decoding the users' intentions from his/her neural activity. Subsequently, the decoding of the user's perception that an error has occurred in the form of ErrP can allow the system to undergo corrective actions, for example, by preventing the erroneous action from being completely executed (Dal Seno et al., 2010; Ferrez & Millán, 2008; Schalk et al., 2000) and/or reducing the possibility of errors reappearing in the future through re-calibration of the system (Artusi et al., 2011; Llera et al., 2011). A handful of studies have shown that it is possible to not only classify errors against correct actions but also to distinguish among different types of errors

based on their ErrP [see, e.g., Iturrate et al., 2015]. Some studies have also addressed differences in ErrP in relation to the error source, for example, distinguishing between response ErrP caused by "me" versus observed errors, occurring when a human observes an error committed by a machine or another human (Ferrez & Millán, 2008). Nevertheless, despite these recent additions, most of the literature in the field concerns the classification of errors against correct actions or observing external agents committing errors, rather than the classification of the agency of errors regarding my "own actions" ("mine" vs. "externally" induced error). With this purpose in mind, we applied a linear SVM classifier to previously acquired EEG data, considering for the analysis the topographical distribution (considering the amplitude at all electrodes and time points) of the brain's response to each error condition. From a neurophysiological point of view, this approach is especially interesting because it permits us to explore the neurophysiological distinctions between the brain activity patterns associated with the different error conditions.

2 | METHOD

2.1 | Participants

A sample of 25 healthy participants (graduate students) were paid to participate in the original study (Gomez-Andres et al., 2023). Two participants were excluded from the analysis due to a low number of self-generated (SE) error trials (24 and 23, respectively). The final sample consisted of 23 participants (15 females, $M = 24.2$ years ± 4.2, mean $\pm SD$) with normal or corrected-to-normal vision. All participants were naïve with respect to the aim of the experiment. The procedures of the experiment were approved by the Biomedical Research Institute of Bellvitge (IDIBELL) and Hospital Universitari de Bellvitge ethics committee (CEIC, Ref. PR254/15). Informed consent in accordance with the Declaration of Helsinki was obtained from all participants prior to the commencement of the study.

2.2 | Apparatus and experimental task

An apparatus inspired by the Rubber Hand Illusion (Kalckert & Ehrsson, 2012, 2014) and the Nielsen's (1963) paradigm was built (see Figure 1). The experiment was performed inside a Faraday chamber with a full HD 24.5-inch monitor displaying the experimental task at 200 Hz refresh rate. The monitor was mounted on a wooden stand and adjusted to the subjects' body, oriented with an inclination of 30° on the horizontal plane (Figure 1a). Participants were asked to put on a pair of white latex gloves to remove any morphological cues that could affect self-identification and to place their hands on top of the wooden stand surface (hidden from their view due to the monitor overlap), where two fixed joysticks with a button at the top were attached. After general instructions were given, the EEG cap was set up and the state of each electrode was checked. Finally, the room lights were turned off for the realization of the experiment.

The experimental task consisted of a modified version of the Eriksen Flanker task (Padrao et al., 2016; Rodriguez-Fornells et al., 2002) (see Figure 1b and Step 1 in Figure 2). Stimulus presentation was controlled with EPrime (Psychology Software Tools Inc., Pittsburgh, PA). Participants were instructed to focus on a central target arrow from a visual array of three arrows allocated vertically and to respond as fast and accurately as possible with the right or left hand (i.e., thumb press on button placed on the high-end of the fixed joystick), depending on the directionality of the target arrow. The flanker arrows located above and below the target arrow were either compatible (i.e., all arrows in the same direction) or incompatible (i.e., flanker arrows pointing in the opposite direction than the target arrow) with respect to the target arrow. To optimize the number of errors, a proportion of 40–60% of compatible and incompatible trials were presented, respectively, in a randomized order.

For each trial, a pair of life-sized hands (i.e., real hand photographs—adult size—wearing the same white latex gloves as the participants') mimicked the participants' hands actions at a coherent position with respect to the participants' hands and body posture. The duration of the target presentation was fixed to 150 ms, followed by a response threshold of 1000 ms (Reaction Time, RT). At the same time the participants were responding, the 'virtual' hands provided the participants with real-time, online feedback (Observed Response) for 100 ms. A variable fixation slide (depending on the RT) appeared at the end of the trial (see Figure 1b for an example trial).

The experimental task was divided in two conditions: (i) a *standard* condition (one block of 160 trials), during which congruent feedback was presented in all cases (i.e., the 'virtual' hand response was always the same as the participant's response) and (ii) an *error induction* condition (two blocks of 320 trials each, 640 trials in total), where incongruent feedback was provided in 10% of the trials (64 trials in total). During the incongruent trials, when the participants pressed the button with one hand, the opposite 'virtual' hand performed the response movement, causing an external error (EE). We avoided introducing EE in incompatible trials to avoid the pre-activation of

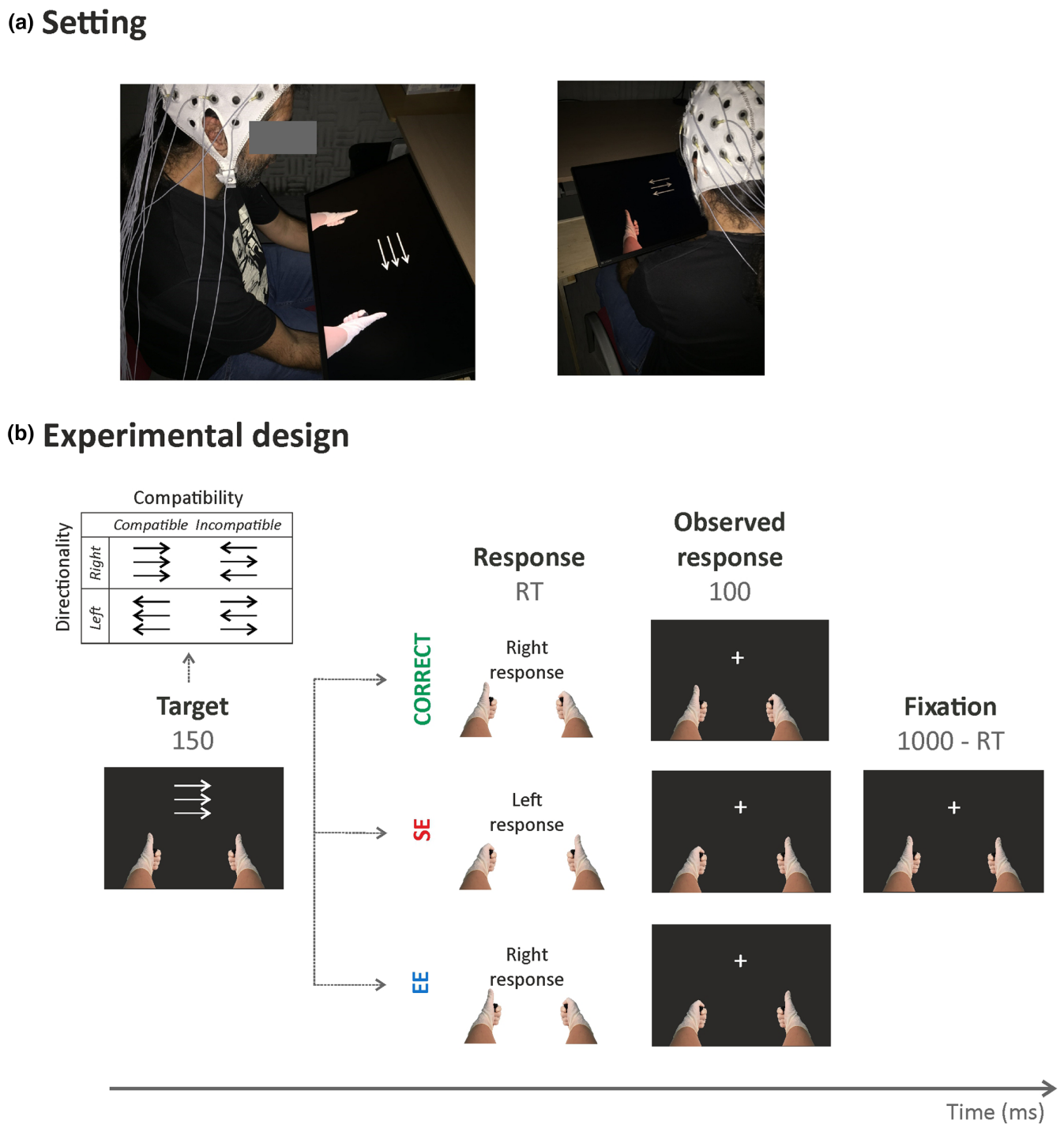

FIGURE 1 Experimental setting and task. (a) Schematic representation of the experimental apparatus. Lateral (left) and bird-view (right) perspective of the participants' position with respect to the wooden-mounted monitor. The participants hands are hidden underneath the monitor, in a coherent position with respect to the 'digital' hand displayed, holding the response joysticks, and the visual feedback displayed on the monitor. (b) Experimental paradigm depicting a modified version of the Eriksen Flanker task (Padrao et al., 2016; Rodriguez-Fornells et al., 2002). All trials started with the target presentation (150 ms) followed by a response threshold (<1000) in which participants had to respond to the target arrow as fast as possible (RT: Reaction Times). The visual feedback (Observed Response, OR) corresponding to the virtual hand response movement was displayed for 100 ms, showing either congruent correct, congruent error or incongruent correct visual displays.

incorrect motor channels responsible for the increase in error rates in the incompatible trials (compatibility effect). Every 80 trials a block pause of 10 s was presented. A training phase (20 trials) was always performed before the experiment began to ensure an adequate speed-accuracy rate.

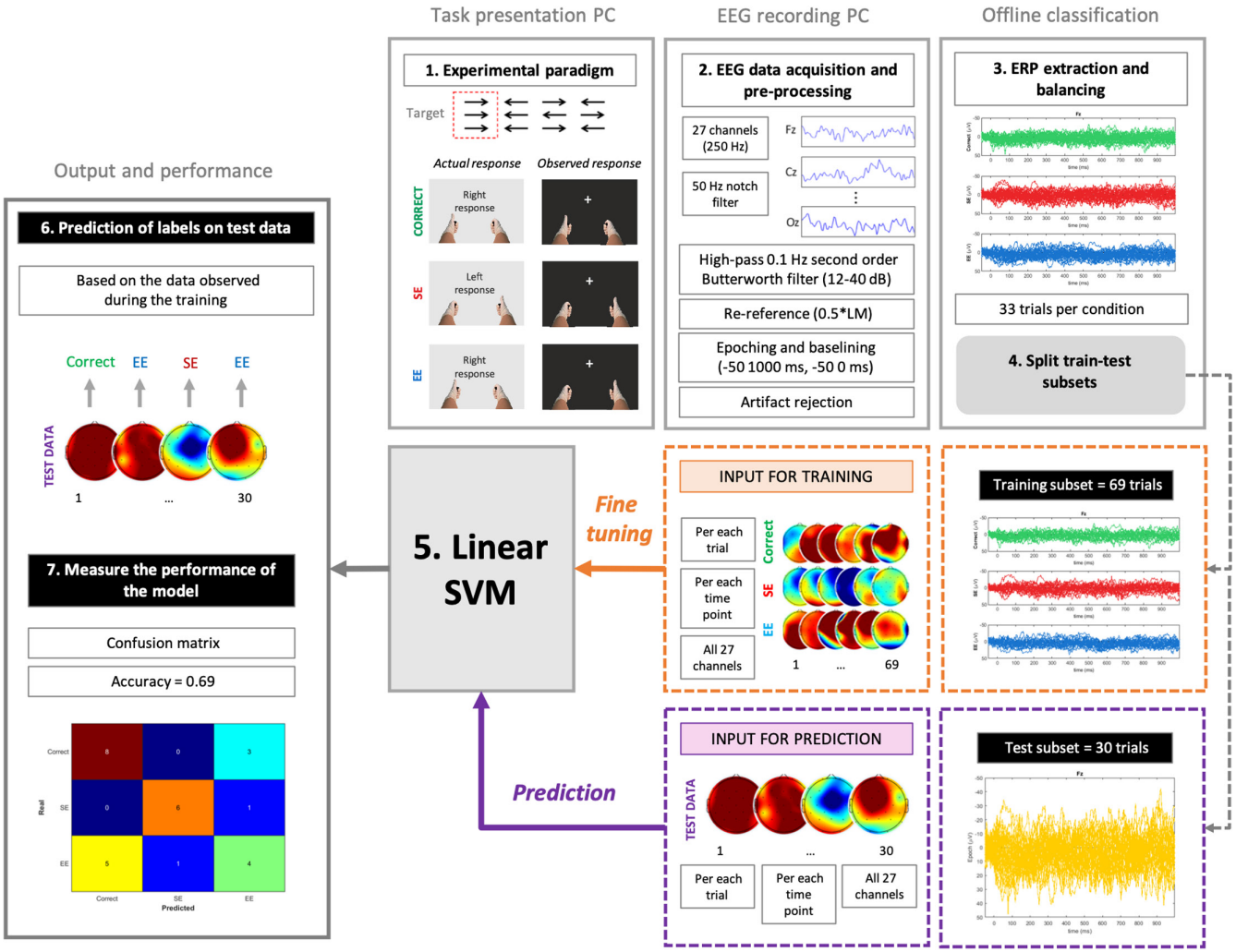

FIGURE 2 Classification pipeline description. EEG data previously acquired (Step 1) was pre-processed (Step 2). We extracted the ERPs and balanced the number trials across conditions (33 trials per condition) (Step 3) and split the data into the training and test subsets (Step 4). The training subset was used to fine tune the linear SVM classifier (Step 5). We then applied this fine-tuned SVM classifier to predict the labels of the test data (Step 6). Finally, we examined the output and performance of the classifiers, in terms of accuracy and confusion matrices.

2.3 | EEG recording

The electroencephalographic (EEG) signal was recorded from the scalp using sintered Ag-AgCl ring electrodes mounted in an elastic cap (Easycap, International 10–20 System locations) and located at 27 standard positions (Fp1/2, Fz, F3/4, F7/8, Fc1/2, Fc5/6, Cz, C3/4, Cp1/2, Cp5/6, T7/8, Pz, P3/4, P7/8, O1/2). Data acquisition was done using a BrainAmp Standard amplifier and Brain-Vision Recorder v1.20.05 software for EEG recording. Biosignals were referenced online to the right mastoid electrode and posteriorly re-referenced offline to half of the signal acquired on the left mastoid electrode. Electrode impedances were kept below 5 kΩ. For all experiments, vertical eye movements were monitored with an electrode at the infraorbital ridge of the right eye. The electrophysiological signals were filtered online with a notch-filter (50 Hz) and a high-pass filter (0.016 Hz) and digitized at a rate of 250 Hz. Participants were given instructions about how to reduce muscle artifacts by minimizing movement and to wait for a visual signal, an array of five asterisks appearing every 10 trials, to free blink for 5 s.

2.4 | Data and statistical analysis

2.4.1 | Behavioral measures

To inspect the behavioral effects during the task execution, we computed the RT and accuracy rates for all trial types (Correct, SE and EE), exploring the compatibility effect (Danielmeier & Ullsperger, 2011; Eriksen & Schultz, 1979), in terms of RT and accuracy rates.

TABLE 1 Item description of the subjective experience questionnaire examining agency attribution.

Subjective feeling	Questionnaire item	Description
Internal attribution	My movements	Q1. Most of the time, the movements of the digital hands seemed to be my own movements
	Feeling of control	Q2. I felt I could control the movements of the digital hands most part of the time
External attribution	Not my movements	Q3. Sometimes, the digital hands seemed to be moving by themselves
	External errors	Q4. It sometimes seemed as if the errors were not caused by me
Control	Influence	Q5. Sometimes I felt as if the movements of the digital hands were influencing my own movements
	More than 2 hands	Q6. It seemed as if I had more than two hands
	My hands (SoO)	Q7. I felt as if the digital hands were my hands

Abbreviation: SoO, sense of ownership.

Also, we examined the subjective experience of agency after the standard and error induction blocks using a 7-item questionnaire (in Spanish) (see Table 1). After each experimental condition, participants were asked to rate their subjective experience focusing on the experience of Agency attribution, which partially depended on the sensorimotor input resulting from the action performed (Haggard, 2017). The questionnaire was designed to address internal (Q1 and Q2, i.e., *'Most of the time, the movements of the digital hands seemed to be my own movements'* and Q2) and external attribution of actions (Q3 and Q4, i.e., *'It sometimes seemed as if the errors were not caused by me'*). Additionally, two control questions (Q5 and Q6, i.e., *'It seemed as if I had more than two hands'*) and one item addressing the Sense of Ownership (SoO, Q7: *'I felt as if the digital hands were my hands'*) were also included. Participants were asked to rate their level of agreement with these 7 statements using a 7-level Likert-type response, ranging from "strongly disagree" (1) to "strongly agree" (7). Wilcoxon test (pairwise comparisons) were employed for testing the possible differences regarding the participants' scores on the questionnaire, with the significance alpha level adjusted to multiple comparisons.

2.4.2 | EEG data

EEG analyses (Step 2 in Figure 2) were conducted using routines taken from the ERPLAB toolbox V6.1.4 (Lopez-Calderon & Luck, 2014) and custom routines from MATLAB (The MathWorks, Inc. Natick, MA). A high-pass filter of 0.1 Hz (non-causal Butterworth impulse response function, half-amplitude cutoff of 0.1 Hz, roll-off of −12 dB/oct) was applied to the raw EEG data. Epochs of −50 to 1000 ms were defined, time-locked to the onset of the participants response, and baseline-corrected to its preceding 50 ms. To perform artifact rejection, we excluded epochs with step-like artifacts when the amplitude jumps in the electro-oculograms exceeded 25 μV (moving window = 400 ms, moving step = 10 ms) or in which activity was ±100 μV in any channel (average of 84.95% of retained trials after artifact rejection). No additional filtering was applied for the subsequent analyses. Since no bad channels were detected, data interpolation was not necessary.

To study the multichannel evoked potentials, we visualized the superimposed activity of all the electrodes at once using a Butterfly plot, and the Global Mean Field Power (GMFP) was calculated (Hamburger & Van der Burgt, 1991; Lehmann & Skrandies, 1980; Skrandies, 1990). The GMFP has been shown as a reference-independent descriptor of the potential field (Skrandies, 1990), and it corresponds to the standard deviation of the activity across scalp electrodes with respect to the mean channel potential at that specific time point (Esser et al., 2006). Subsequently, a baseline correction (from −100 to 0 ms) was applied (Ort et al., 2023).

2.4.3 | Decoding

The aim of this work was to decode, at a single-trial level, the timeseries of the ERPs to be able to separate the different defined conditions (Correct vs. SE vs. EE), and to study at which time points these three conditions were separable by a supervised learning model such as a Support Vector Machine (SVM) classifier. This decoding procedure was conducted using *scikit-learn* routines from Python (Pedregosa et al., 2011), and visualization of the results was done using custom routines from MATLAB. See Figure 2 for a pipeline on the classification algorithm.

To ensure the predictive performance of the decoding procedure, the number of trials per condition was balanced (Step 3 in Figure 2). Since the number of EE trials was set beforehand, the limiting condition was the SE. In that sense, the participant who had less SE trials had 33, therefore, we considered 33 trials per condition and subject. For the participants who had more than 33 trials for any condition, we randomly selected 33, resulting in a total of 99 ERPs for each subject. The 70% of these trials ($N_{train} = 69$) were used to train the classifier, and the 30% of the trials ($N_{test} = 30$) were used in the test phase (Step 4 in Figure 2). The proportion of trials of each class/condition (33%) was maintained in both subsets.

A SVM classifier with a linear kernel was considered for the decoding procedure (Step 5 in Figure 2). For each subject and time point, a single multiclass SVM was fitted using the training trials, and the accuracy of the model was calculated using the testing trials. For each classifier, the hyperparameter C was fine-tuned using a cross-validation method of the training subset with $K = 5$. The input of the classifier $SVM_{s,t}$ was the amplitude (in μV) from all the electrodes of the subject s at the time point t. To discriminate the classes, the classifier considered the distribution of the EEG signal across the scalp (i.e., the topography) at the time point t.

Once the $SVM_{s,t}$ classifier was trained, the labels of the testing trials were predicted (Step 6 in Figure 2), and the performance of the model was calculated (Step 7 in Figure 2). To calculate this performance, we used the confusion matrix, which considers both the predicted and the real labels, the F1-score, and the accuracy. To consider a prediction as correct, the predicted and the real labels had to match, providing a very strict assessment of decoding. To compare the time-point accuracy against the chance performance (i.e., theoretical chance level equals to 0.33, because we defined three conditions; actual chance level equals to 0.34 ± 0.00, mean actual chance $\pm SD$), we used a nonparametric cluster-based Monte Carlo simulation analysis (Bae & Luck, 2018; Groppe et al., 2011; Maris & Oostenveld, 2007) with 10,000 permutations. This analysis allowed the authors to detect the clusters of contiguous time-points for which the performance of the classifier is above the actual chance ($p < .05$), and to obtain a cluster-level t mass. If the obtained cluster-level t mass was larger than the maximum of the Monte Carlo cluster-level t mass ($max\ t\ mass = 42.90$), we reported $p < 10^{-4}$. We rejected the null hypothesis (H0: the classifier performance was not different to chance performance), for any cluster-based t mass above the top 95% of the null distribution ($critical\ t\ mass = 14.75$, one-tailed, $a < .05$; see Bae & Luck, 2018 for further details on this analysis).

Moreover, to analyze the confused classes/conditions at a specific time point, we used the confusion matrices at this time point to calculate a *confusion metric* for each pair of classes (Correct vs. SE, Correct vs. EE, and SE vs. EE). This metric, for a given pair of classes (c1, c2), is defined as $(TP_{c1} + TP_{c2})/(FP_{c1/c2} + FP_{c2/c1})$, where $FP_{c1/c2}$ are the c2 trials predicted as c1 and the $FP_{c2/c1}$ the c1 trials predicted as c2. This metric is between 0 and 1, and the higher the metric, the lower confusion between the pair of classes. Since this metric was calculated at each time-point, it allowed us to capture how the different pairs of classes were confused along the time.

Interestingly, since model prediction relied on electrodes' amplitudes (topographies), we analyzed the relation between the difference waveform and the confusion metric by means of Pearson's correlation (significance level was set at $p < .05$). We averaged (across electrodes) the absolute value of the amplitude difference between conditions. Therefore, we obtained an averaged absolute difference waveform at each time point for every pair of conditions. A positive correlation between these two measures would indicate that the larger the waveform difference between conditions, the higher the confusion metric between them (i.e., the lower the model confusion).

Finally, since the GMFP captures the variability of the data at a certain time point by quantifying the amount of activity at each time point in the field considering the data from all recording electrodes, we inspected the association between the GMFP and the accuracy of the decoding procedure by means of Pearson correlation. Since the times of GMFP maxima are used to determine the latencies of the relevant ERP components (Skrandies, 1990), we expected to find a positive correlation between the GMFP and the decoding accuracy driven the topographical and latency-dependent characteristics of the ErrPs.

2.4.4 | Statistical analysis

Categorical variables were reported as absolute values (gender), while continuous variables were reported as the mean ± either the standard deviation or the standard error of means. The normality distribution of the data was

checked using the Shapiro–Francia test (function *sf.test* from *nortest* package) and visual inspection.

All statistical analyses were conducted in R (Version 3.6.0, R Core Team, 2019. https://www.R-project.org). The relationship between continuous variables was assessed using Pearson's correlations. To compare groups of two factors, for the continuous normal-distributed data, two-sided, unpaired t-tests of equal variances checked by the Levene's test (R *car* package) were used and both the *t*- and *p*-values were reported (R *stats* package). For groups of more than two factors or several levels of interaction, an ANOVA for balanced designs was used, reporting the *F*- and *p*-values. Moreover, when necessary, *p*-values were corrected for multiple comparisons (p_{adj}) using the False Discovery Rate (FDR). Finally, for the group's comparisons, the effect sizes were reported, i.e., Cohen's *d* for the *t*-tests and η_p^2 for the ANOVA.

Raw data were generated at University of Barcelona, and derived data supporting the findings of this study are available from the corresponding author [ARF] on request.

3 | RESULTS

3.1 | Behavioral results

The participant's performance was as expected for this paradigm, with a mean percentage of self-generated errors (SE) approx. of 9% ± 1 (mean ± SEM). A compatibility effect (Danielmeier & Ullsperger, 2011; Eriksen & Schultz, 1979) was encountered, with participants responding more accurately [percentage of SE during compatible vs. incompatible trials, 13.4% ± 1.1 vs. 86.6% ± 1.1, respectively; $t(22) = 32.78$, $p < .0001$, $d = 13.8$] and faster [mean RT for correct: 282 ± 4 ms vs. 299 ± 5 ms; $t(22) = 9.73$, $p < .001$, $d = 0.8$] during compatible versus incompatible trials. Moreover, participants also showed significantly faster RT for SE compared to correct responses [mean RT correct: 290 ± 4 ms vs. mean RT SE: 236 ± 5 ms; $t(22) = 21.7$, $p < .0001$, $d = 2.4$]. Altogether, these results indicated a correct implementation of the Eriksen flanker task.

In relation to the SoA experience, the insertion of an EE in the error induction blocks lead to an external attribution of the actions, revealing significant differences between the standard and error induction blocks for both external attribution questions: Q3 (*"Sometimes, the digital hands seemed to be moving by themselves"*) ($Z = -3.64$, $p < .001$) and Q4 ("*It sometimes seemed as if the errors were not caused by myself*") ($Z = -3.64$, $p < .001$), revealing a significant external attribution judgment when introducing EE but without affecting the internal attribution ratings (Q1: $Z = -1.75$, $p = .08$; Q2: $Z = -2.26$, $p = .024$). High levels of SoO were observed for both conditions, and no other significant differences were found for any of the other control measures.

3.2 | Error-related brain potentials (ERPs)

Figure 3a shows the average waveforms for Correct (green line), SE (red line), and EE (blue line) trials time-locked to the participants' response. During SE, the well-known error-related negativity (ERN) component was elicited, peaking at around 80 ms after error commission. The ERN's topography depicted a fronto-central distribution, with its maximum activity at Fz electrode locations (see Figure 3b). On the contrary, the insertion of EE elicited a more latter and posterior P600 component, peaking at 580 ms after the EE induction evidencing a centro-parietal topographical distribution (Figure 3b, EE) (reported in Gomez-Andres et al., 2023). Although grand average waveforms across participants are depicted, single-subject data were entered to the classification algorithm.

The multichannel evoked potential analysis (Figure 3c) indicates two time periods where the variability of the electrodes' activity is increased (the GMFP is higher). The first component is sharp and peaks around 95 ms, and the second one peaks around 585 ms. In both cases, these correspond clearly to the ERN and P600 periods previously identified.

3.3 | Decoding

3.3.1 | Cluster permutation analysis

The linear SVMs were fitted for each participant and time-point, leading to 6049 (23 subjects × 263 time-points) classifiers. For each $SVM_{s,t}$ classifier, a confusion matrix was obtained, and the subject's mean accuracy along the time was calculated (Figure 4). The classifier performed better than the actual chance when it was able to distinguish at least one class/condition over the others. This fact occurred from −6 ms to the end of the ERP (*t mass* = 1804.6; $p < 10^{-4}$).

3.3.2 | Confusion metric

Figure 5a shows the confusion metric along the time between all the pairs of classes. The confusion metric suggests that two processes can be differentiated in the significant interval. During the first one (which starts

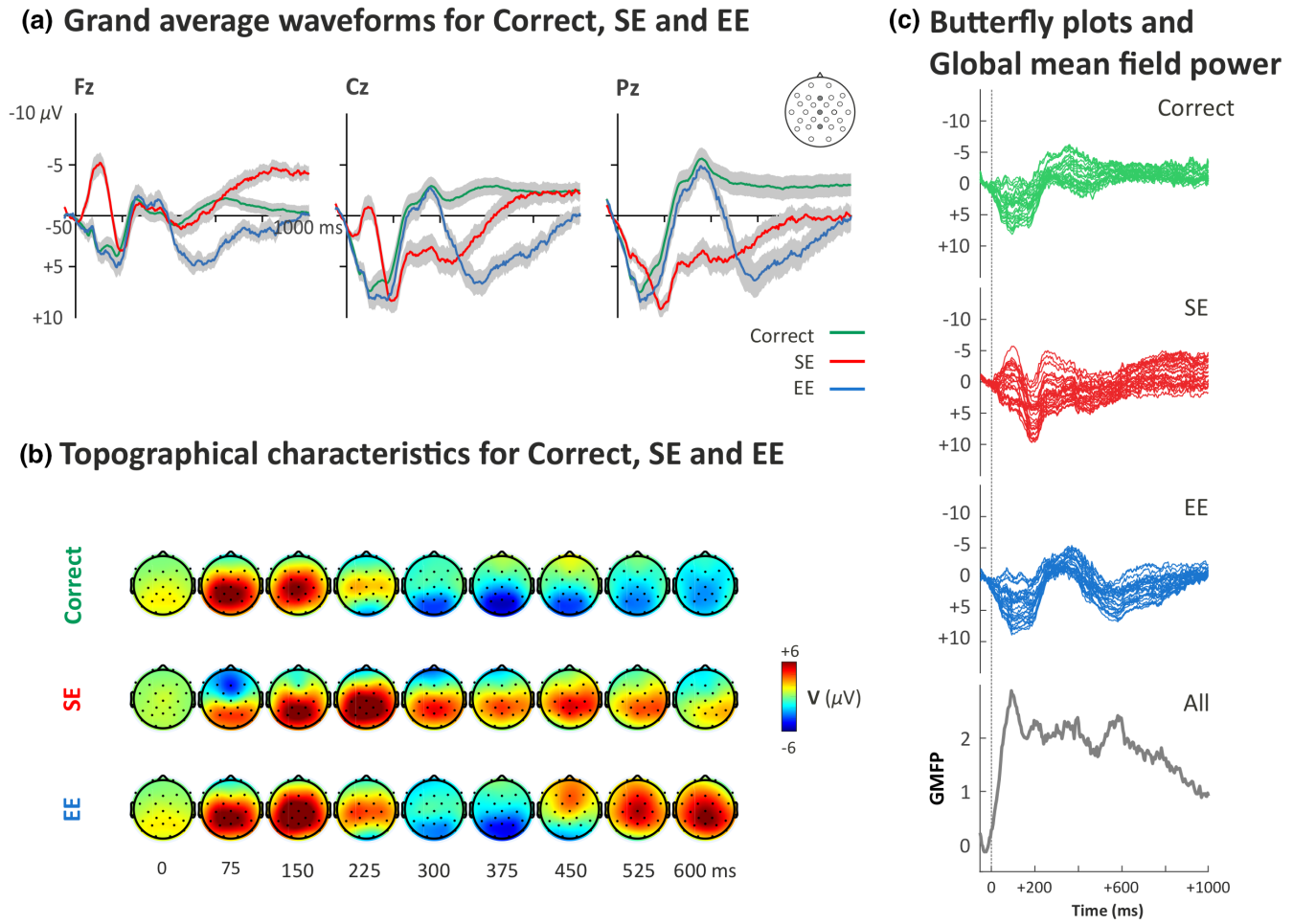

FIGURE 3 Description of EEG data. (a) Response-locked grand-average event-related potentials (ERPs) for Correct, self-errors (SE) and external errors (EE) at Fz, Cz, and Pz electrodes. In gray we show the standard deviation (*SD*). (b) Topographical characteristics for Correct, SE, and EE trials from 0 to 600 ms. (c) Butterfly plots for Correct, SE, and EE, and the global mean field power (GMFP) for all electrodes at each time point.

at −6 ms approx.), the SE class pops-up over the rest and roughly corresponds to the ERN component time-window). During the second one, starting at 400 ms, Correct and EE classes become more distinguishable. This observation is strengthened by both the topographies and confusion matrices shown in Figure 5b. At 0 ms, the topographies between the three classes are very similar, which leads to classifiers that do not perform better than chance (high number of mismatches). In contrast, as the time is increased (bottom panels of Figure 5b), the condition topographies become more distinguishable, with a maximum differentiation between Correct and EE classes at 585 ms (corresponding to the P600 period).

As previously stated, since topographies were the input of the classifiers, we tested the correlation between the average of the absolute difference waveform and the confusion metric. Figure 6a depicts the difference waveform between conditions, showing that EE and Correct conditions started to be distinguishable from 160 ms onwards. Furthermore, a significant high positive correlation was observed between the averaged absolute value of the difference waveform and the confusion metric (Figure 6b; $r = .6988$, $p = 2.20 \times 10^{-16}$). This positive relationship indicated that the larger the difference waveform between conditions, the higher the confusion metric (the lower the confusion between conditions).

3.3.3 | Individual performance accuracy

The individual differences on the decoding performance are reported in Table 2, where the accuracy of the decoding procedure strongly depends on the subject (the subject accuracy, averaged during the significant interval, ranges from 0.35 for S22 to 0.50 for S23). Although individual differences are noticeable, Figure 7 shows that, as expected, the subject's mean accuracy along time and the GMFP have a very strong correlation ($r = .7025$, $p = 2.20 \times 10^{-16}$), indicating that the higher the GMFP, the higher the accuracy of the model.

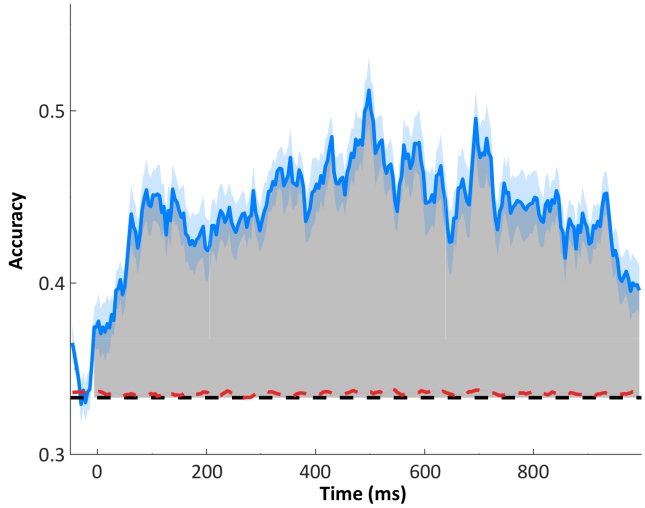

FIGURE 4 Classification accuracy over time. Graphical representation of the accuracy (ranging from 0 to 1) for the time period −50 to 1000 ms. In blue we show the mean accuracy for all participants and for every time point, and the SEM (bluish shadow). In gray we highlight the presence of significant clusters after performing the nonparametric cluster-based Monte Carlo simulation analysis (Bae & Luck, 2018; Groppe et al., 2011; Maris & Oostenveld, 2007) with 10,000 permutations. Black dashed line indicates the theoretical chance level (0.33), and the red line indicates the actual chance level.

3.3.4 | F1-scores

In Figure 8, we analyzed the mean F1-score, which indicates how well a class/condition is classified, at these two components time-windows (ERN: 95 ± 25 ms, and P600: 580 ± 25 ms). An analysis of variance model was fitted, considering the component time-windows (ERN vs. P600) and the condition (Correct vs. SE vs. EE) factors on the F1-score. A subsequent ANOVA test indicated a main effect of Condition ($F(2) = 3.29$, $p = .0403$, $\eta_p^2 = 0.05$), together with an interaction between Condition and Component ($F(2) = 14.20$, $p = 2.59 \times 10^{-6}$, $\eta_p^2 = 0.18$). Post hoc analyses reported significant differences between SE and Correct and EE during the elicitation of the ERN (p_{adj}(*SE* vs. *Corr.*) = .0005; p_{adj}(*SE* vs. *EE*) = .0005).

At more later stages of processing (∿580 ms—P600 period), the classification of the Correct and EE conditions improved, being their F1-score higher at P600 than at ERN (F1-score(Correct) = 0.50 ± 0.01, and F1-score(EE) = 0.46 ± 0.02; p_{adj}(Correct) = .0005, p_{adj}(EE) = 0.0148). On the other hand, the classification of the SE worsened in P600 compared to ERN (F1-score(SE) = 0.43 ± 0.02; p_{adj}(SE) = 0.0148). Interestingly, the Correct condition was the best classified at the P600 latency, with significant differences between Correct and SE ($p_{adj} = .0061$), but not between Correct and EE ($p_{adj} = .1497$), and SE and EE ($p_{adj} = .2917$).

4 | DISCUSSION

In the present study, our aim was to investigate the feasibility of applying a machine learning decoder to decipher ErrPs from an EEG experimental paradigm. As previously mentioned in the introductory section, the ability to distinguish self versus externally generated actions is a key factor influencing learning and adaptive behavior and is crucial for agency inference. Here, we classified three types of trials, namely, Correct, SE and EE on a single-trial basis using a linear SVM model based on the ERP latency and its topographical distribution. Significantly, our results showed that the classifier performance was better than chance from −6 ms onwards, showing two distinguishable peaks of accuracy overlapping with the latencies of the ERN during self-made errors and the P600 for externally induced errors.

As previously stated, ErrPs can be observed at the single-trial level allowing us to distinguish correct versus erroneous actions (Chavarriaga et al., 2014; Iturrate et al., 2015; Kim et al., 2017; Usama et al., 2021; Zander et al., 2016). In the present study, the existence of two dissociable components for SE and EE was confirmed by the GMFP measures, suggesting the existence of two main ERP components, the first one peaking at 95 ms, and the second one at 585 ms, approximately (Gomez-Andres et al., 2023). Considering the latencies of these two components, it is feasible to say that they correspond to the ERN (for SE) and P600 (for EE) components (see Figure 3).

Our results from the linear SVMs (fitted for each participant and time-point) evidenced a performance above chance from −6 ms onwards. When looking at the confusion metric and matrices at the different time points between all the pairs of classes, showed a good classification of the SE class (i.e., SE vs. Correct and SE vs. EE) (see Figure 5a), evidencing the distinguishable topographical characteristics of the ERN, while the confusion metric was maintained for the Correct versus EE. This result confirms previous evidence highlighting the utility of the ERN for decoding self-made errors, confirming its dissociable neural characteristics from correct responses (Schmidt et al., 2012; Spüler et al., 2012). Moreover, SE versus EE could also be correctly classified at this point in time, favoring the possibility of performing rapid error corrections if the error is coming from the self. More posteriorly, at approximately +400 ms, the Correct and EE classes became more distinguishable,

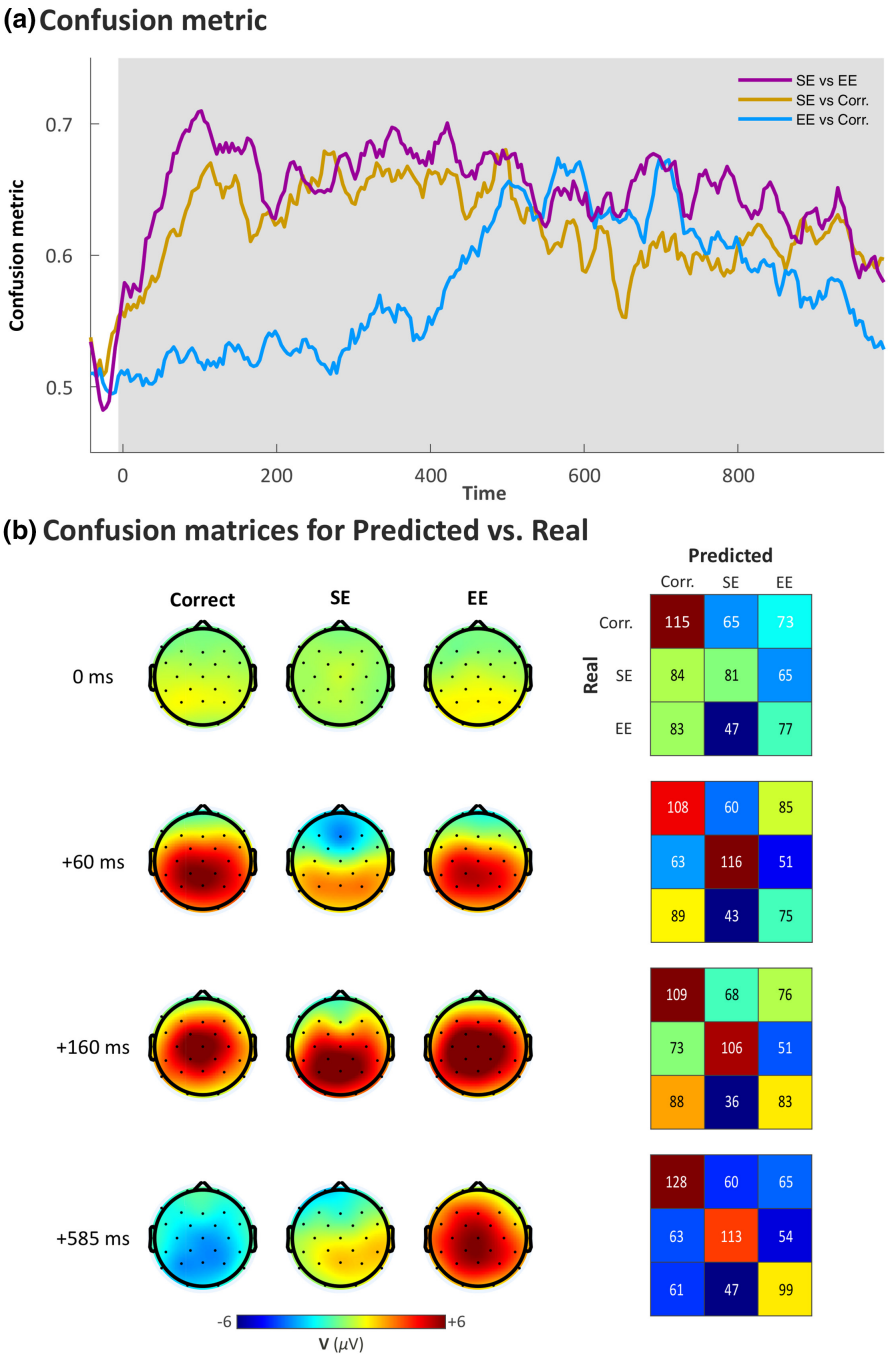

FIGURE 5 Confusion metric and matrices. (a) Confusion metric (ranging from 0 to 1) along time for the pairs Correct versus SE (orange line), Correct versus EE (blue line), and SE versus EE (purple line). (b) Topographical representations of Correct, SE, and EE at several time points of interest (left) and confusion matrices for Correct, SE, and EE for predicted versus real trial classifications (right).

reflecting the P600 component appearing during the EE condition, with a maximum differentiation at 585 ms. The ability to correctly classify EE seems to be more related to later stages of processing probably indicating the need to recruit more reflective aspects of agentic processing related to higher cognitive functions (Moore & Haggard, 2008; Synofzik et al., 2008).

Worth noting, in the second interval, the accuracy of the model decreased at latencies around 200 ms (Figure 4), indicating that, at this time point, conditions were less distinguishable. Considering the grand-averaged ERPs (Figure 3a), this decrease of accuracy coincided with a positivity in all conditions (including the SE). This positivity on incorrect choices, corresponding to the Pe component and peaking at around 200 ms at centroparietal electrodes, was also reported in previous studies (Di Gregorio et al., 2018; Falkenstein et al., 2000; see review in Ullsperger et al., 2014). Although the origin of this positivity remains unclear, it seems that this error positivity increases the confusion in the condition's classification.

Individual differences were observed in both the Agency attribution questionnaire and the model performance. One could argue that the model performance, at least on the EE condition, could correlate with the attribution of agency, but there was no significant correlation between the F1-score of the EE class during the significant interval, and the Agency

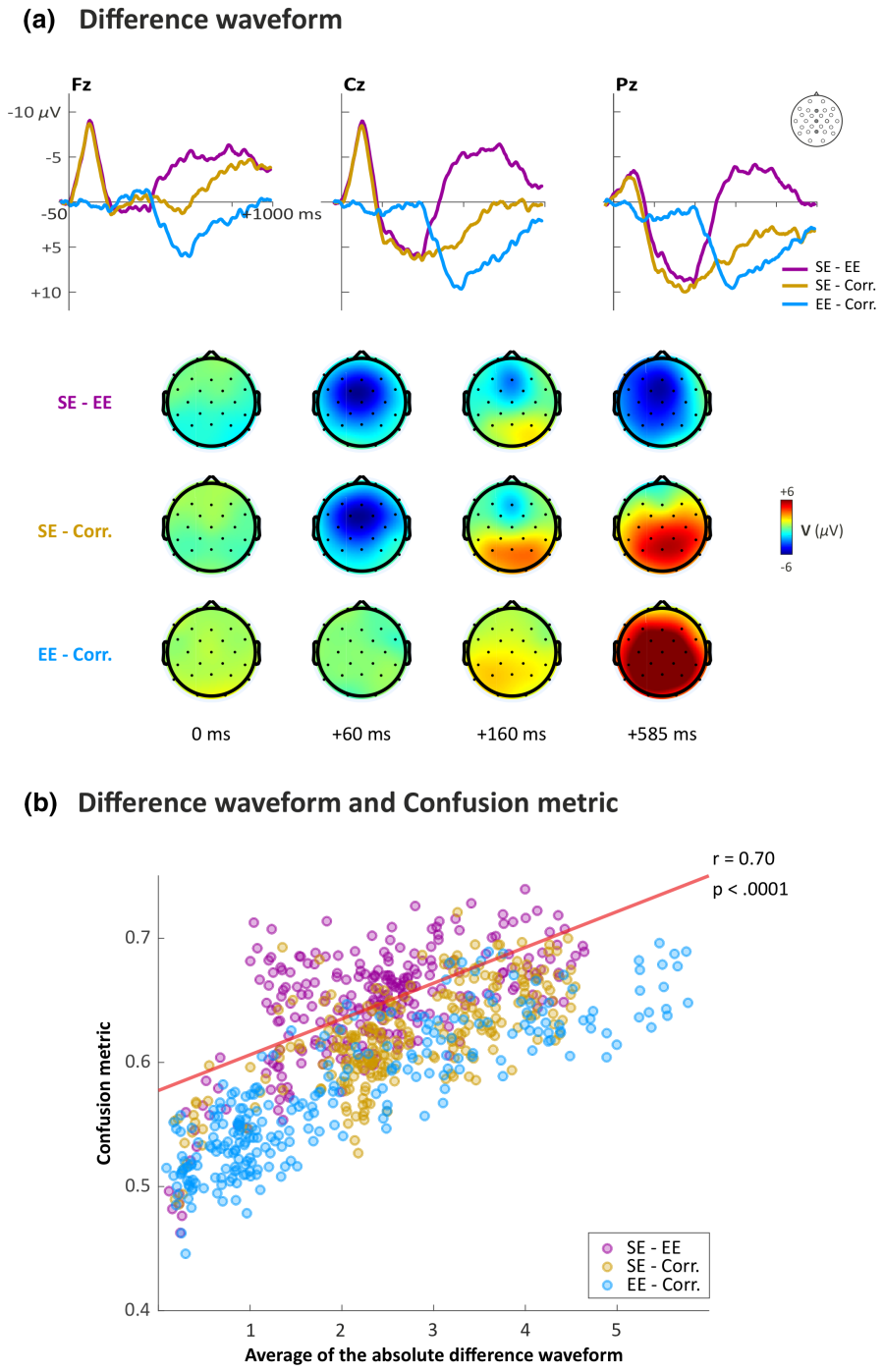

FIGURE 6 Description of differences between conditions. (a) Both time (top) and topographic plots (bottom) considering the difference waveform between conditions. (b) Pearson's correlation for averaged absolute value of the difference waveforms and confusion metric. Pairs of conditions are indicated by colors. The positive correlation indicates that the larger the topographical difference, the lower the confusion by the model.

attribution score (r=.31, p=.1647). Probably, this indicates that the model performance depends on how clear and robust the ERP signal is, and not on how much agency attribution the participant has. Moreover, although individual differences existed in terms of accuracy rates, the mean accuracy along time and the GMFP had a very strong correlation (see Figure 7), indicating that the higher the GMFP, the higher the accuracy of the model.

Not surprisingly, when examining the F1-scores, the SE condition was the best classified at the ERN time window. This observation goes in line with previous electrophysiological studies that reported the ERN as a key component to detect the self-generated errors (Falkenstein et al., 2000; Gehring et al., 1993; Taylor et al., 2007). Probably, both the polarity (negativity) and the localization (frontal electrodes) of the activity make the ERN easy to detect by a classifier focused on the proper latencies (∼95 ms). In the same way, this too early component does not permit the model to classify the Correct condition better than the EE condition. In fact, the classification of the Correct and EE conditions improves along the time, that is, their F1-score at P600 is higher than that of at ERN.

TABLE 2 Individual results for F1-scores and classifier accuracy.

	F1 score			
Subject	**Correct**	**SE**	**EE**	**Accuracy**
S01	0.47	0.58	0.43	0.50
S02	0.45	0.50	0.34	0.44
S03	0.44	0.45	0.47	0.45
S04	0.48	0.43	0.44	0.45
S05	0.47	0.45	0.52	0.49
S06	0.48	0.47	0.46	0.47
S07	0.46	0.39	0.42	0.43
S08	0.42	0.54	0.42	0.46
S09	0.50	0.50	0.40	0.48
S10	0.42	0.50	0.33	0.42
S11	0.45	0.41	0.40	0.43
S12	0.49	0.53	0.47	0.49
S13	0.42	0.40	0.37	0.40
S14	0.43	0.51	0.36	0.43
S15	0.45	0.50	0.37	0.44
S16	0.42	0.43	0.41	0.42
S17	0.47	0.47	0.39	0.45
S18	0.42	0.40	0.43	0.42
S19	0.42	0.35	0.41	0.40
S20	0.43	0.44	0.36	0.42
S21	0.47	0.45	0.42	0.45
S22	0.36	0.36	0.35	0.35
S23	0.49	0.59	0.40	0.50
Mean	0.45	0.46	0.41	0.44

Note: Performance of the models for each subject, averaged along the significant interval.

Abbreviations: EE, externally generated errors; SE, self-generated errors.

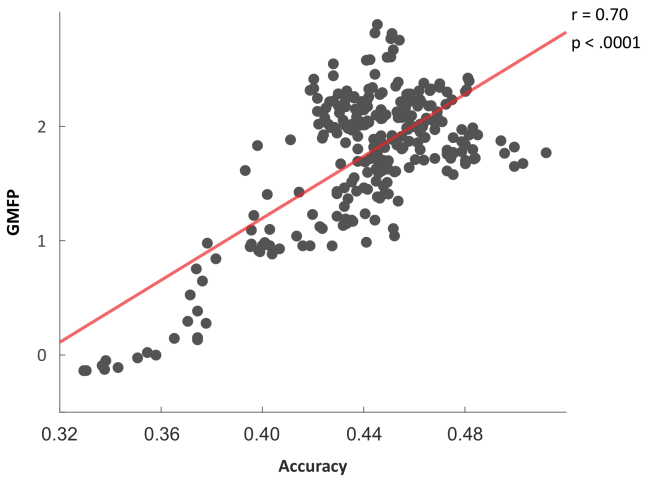

FIGURE 7 Correlation between global mean field power (GMFP) and accuracy of the SVM decoder. Pearson's correlation for GMFP and accuracy measures depicting a significant positive association between an increase GMFP and increase in model accuracy.

Interestingly, the Correct condition was the best classified at the P600 latency, with significant differences between Correct versus SE but, not between Correct versus EE and SE versus EE. These results indicate that the EE condition is more distinguishable by later components such as P600, which goes in-line with previous studies that reported a centro-parietal large positive waves (P600) produced by erroneous or incongruent conditions (Pijnacker et al., 2010).

Although it was not the aim of this manuscript, we tested whether a model which can consider full information (i.e., all amplitudes during the whole ERP waveform, including the ERN and the P600 components) improved the accuracy compared to a classifier which only considered the individual time points (as described in this manuscript). We constructed a model which input was the down-sampled (50 Hz) ERPs, and the procedure was the same one as described in the Methods section, with the only difference of the model's input. Therefore, a total of 23 different "full models" were trained (one per participant). The averaged accuracy across subjects (full model accuracy = 0.65 ± 0.11; [0.50–0.83]) was significantly higher than that of the "time point model" (time point model accuracy = 0.44 ± 0.03; [0.35–0.49]; $t(26) = -8.95$, $p = 1.86 \times 10^{-9}$, $d = -2.64$), that is, the full model (which could consider more information about the waveform) was able to improve the classification of errors' attribution.

To sum up, in the present study we were able to correctly classify, at the single-trial level, Correct, SE and EE ERPs by means of a linear SVM classifier based on the latency and topographical characteristics of several ErrPs. Importantly, the model indicates that SE condition is the best classified condition during the frontal ERN time window, while the Correct and EE conditions were more accurately classified at more later time window (from 400 ms onwards), corresponding to the parietal P600 component. We believe the present results provide crucial new evidence on the importance of agency attribution of our actions and the actions supposedly governed by our mind (on external agents or surrogated bodies). We encountered that single-trial EEG activity contains decodable information about our capacity to accurately monitor our actions, both the ones we initiated and the ones that are imposed to us. This research has also potential value for further research applications in the near future, not only for understanding the origin of our intentions to act, actions and the sense of agency but also for the potential implications it might have regarding moral responsibility over our supposed actions.

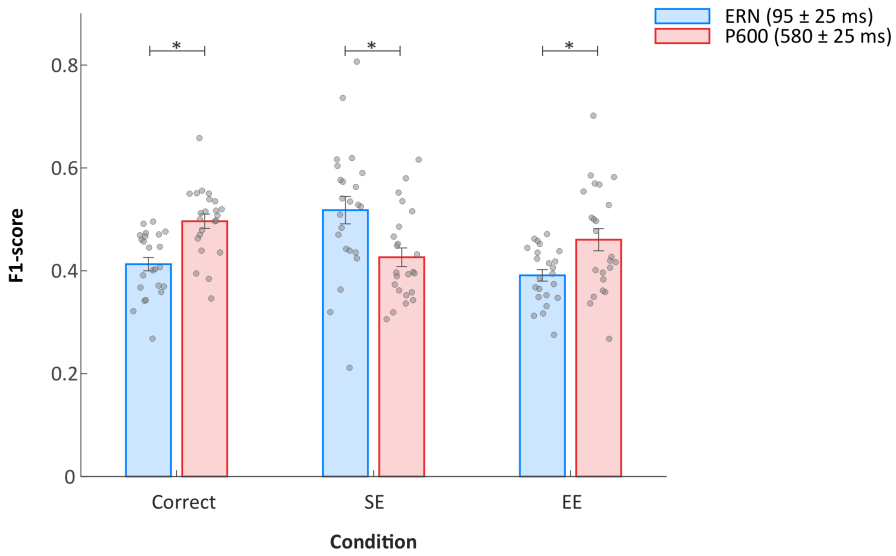

FIGURE 8 Mean F1-scores at the ERN and P600 time windows. Graphical representation of the F1-scores for the ERN and P600 time windows depicting the significant differences between conditions (Correct/SE/EE). During the ERN time window, significant differences were encountered between SE and EE. Significant differences between time windows were observed for Correct and EE classes. SE: self-generated errors, EE: externally generated errors.

AUTHOR CONTRIBUTIONS

Alba Gomez-Andres: Conceptualization; data curation; formal analysis; writing – original draft. **Xim Cerda-Company:** Conceptualization; data curation; formal analysis; writing – original draft. **David Cucurell:** Data curation. **Toni Cunillera:** Conceptualization; funding acquisition; project administration; supervision; writing – review and editing. **Antoni Rodríguez-Fornells:** Conceptualization; funding acquisition; project administration; resources; supervision; writing – review and editing.

ACKNOWLEDGMENTS

This study was supported by the Spanish Government with a MINECO Grant awarded to A.R-F. (PSI2015-69178-P), a MINECO Grant awarded to T.C. (PSI2016-79678-P), a predoctoral FPI grant awarded to A.G-A. (BES-2016-078889), and X.C-C has been awarded with a Margarita Salas grant funded by the Ministerio de Universidades and the European Union—NextGenerationEU.

CONFLICT OF INTEREST STATEMENT

The authors declare no competing financial interests.

DATA AVAILABILITY STATEMENT

Raw data were generated at University of Barcelona, and derived data supporting the findings of this study are available from the corresponding author [ARF] on request.

ORCID

Xim Cerda-Company https://orcid.org/0000-0001-7359-5453
Antoni Rodríguez-Fornells https://orcid.org/0000-0002-3249-6931

REFERENCES

Artusi, X., Niazi, I. K., Lucas, M. F., & Farina, D. (2011). Accuracy of a BCI based on movement-related and error potentials. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, *2011*, 3688–3691. https://doi.org/10.1109/IEMBS.2011.6090624

Bae, G. Y., & Luck, S. J. (2018). Dissociable decoding of spatial attention and working memory from EEG oscillations and sustained potentials. *The Journal of Neuroscience*, *38*(2), 409–422. https://doi.org/10.1523/JNEUROSCI.2860-17.2017

Bhattacharyya, S., Konar, A., & Tibarewala, D. N. (2014). Motor imagery, P300 and error-related EEG-based robot arm movement control for rehabilitation purpose. *Medical & Biological Engineering & Computing*, *52*(12), 1007–1017. https://doi.org/10.1007/s11517-013-1123-9

Chavarriaga, R., Sobolewski, A., & Millán, J. D. R. (2014). Errare machinale est: The use of error-related potentials in brain-machine interfaces. *Frontiers in Neuroscience*, *8*, 208. https://doi.org/10.3389/fnins.2014.00208

Dal Seno, B., Matteucci, M., & Mainardi, L. (2010). Online detection of P300 and error potentials in a BCI speller. *Computational Intelligence and Neuroscience*, *2010*, 307254. https://doi.org/10.1155/2010/307254

Danielmeier, C., & Ullsperger, M. (2011). Post-error adjustments. *Frontiers in Psychology*, *2*, 233. https://doi.org/10.3389/fpsyg.2011.00233

David, N., Cohen, M. X., Newen, A., Bewernick, B. H., Shah, N. J., Fink, G. R., & Vogeley, K. (2007). The extrastriate cortex distinguishes between the consequences of one's own and others' behavior. *NeuroImage*, *36*(3), 1004–1014. https://doi.org/10.1016/j.neuroimage.2007.03.030

David, N., Newen, A., & Vogeley, K. (2008). The "sense of agency" and its underlying cognitive and neural mechanisms. *Consciousness and Cognition*, *17*(2), 523–534. https://doi.org/10.1016/j.concog.2008.03.004

Di Gregorio, F., Maier, M. E., & Steinhauser, M. (2018). Errors can elicit an error positivity in the absence of an error negativity: Evidence for independent systems of human error monitoring.

NeuroImage, *172*, 427–436. https://doi.org/10.1016/j.neuroimage.2018.01.081

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Perception & Psychophysics*, *25*(4), 249–263. https://doi.org/10.3758/bf03198804

Esser, S. K., Huber, R., Massimini, M., Peterson, M. J., Ferrarelli, F., & Tononi, G. (2006). A direct demonstration of cortical LTP in humans: A combined TMS/EEG study. *Brain Research Bulletin*, *69*(1), 86–94. https://doi.org/10.1016/j.brainresbull.2005.11.003

Falkenstein, M., Hoormann, J., Christ, S., & Hohnsbein, J. (2000). ERP components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, *51*(2–3), 87–107. https://doi.org/10.1016/s0301-0511(99)00031-9

Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: A positron emission tomography study. *NeuroImage*, *18*(2), 324–333. https://doi.org/10.1016/s1053-8119(02)00041-1

Farrer, C., & Frith, C. D. (2002). Experiencing oneself vs another person as being the cause of an action: The neural correlates of the experience of agency. *NeuroImage*, *15*(3), 596–603. https://doi.org/10.1006/nimg.2001.1009

Ferrez, P. W., & Millán, J. D. R. (2008). Error-related EEG potentials generated during simulated brain–computer interaction. *IEEE Transactions on Biomedical Engineering*, *55*(3), 923–929. https://doi.org/10.1109/tbme.2007.908083

Frith, C. D., Blakemore, S. J., & Wolpert, D. M. (2000). Abnormalities in the awareness and control of action. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *355*(1404), 1771–1788. https://doi.org/10.1098/rstb.2000.0734

Gehring, W. J., Goss, B., Coles, M. G., Meyer, D. E., & Donchin, E. (1993). A neural system for error detection and compensation. *Psychological Science*, *4*(6), 385–390. https://doi.org/10.1111/j.1467-9280.1993.tb00586.x

Gomez-Andres, A., Cunillera, T., Cucurell, D., & Rodríguez-Fornells, A. (2023). The complex nature of agency attribution: Neurophysiological signatures associated to monitoring self vs. external erroneous actions. Under review.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711–1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Haggard, P. (2017). Sense of agency in the human brain. *Nature Reviews. Neuroscience*, *18*(4), 196–207. https://doi.org/10.1038/nrn.2017.14

Haggard, P., & Tsakiris, M. (2009). The experience of agency: Feelings, judgments, and responsibility. *Current Directions in Psychological Science*, *18*(4), 242–246. https://doi.org/10.1111/j.1467-8721.2009.01644.x

Hamburger, H. L., & Van der Burgt, M. A. (1991). Global field power measurement versus classical method in the determination of the latency of evoked potential components. *Brain Topography*, *3*(3), 391–396. https://doi.org/10.1007/BF01129642

Iturrate, I., Chavarriaga, R., Montesano, L., Minguez, J., & Millán, J. D. R. (2015). Teaching brain-machine interfaces as an alternative paradigm to neuroprosthetics control. *Scientific Reports*, *5*(1), 1–10. https://doi.org/10.1038/srep13893

Kalckert, A., & Ehrsson, H. H. (2012). Moving a rubber hand that feels like your own: A dissociation of ownership and agency. *Frontiers in Human Neuroscience*, *6*, 40. https://doi.org/10.3389/fnhum.2012.00040

Kalckert, A., & Ehrsson, H. H. (2014). The moving rubber hand illusion revisited: Comparing movements and visuotactile stimulation to induce illusory ownership. *Consciousness and Cognition*, *26*, 117–132. https://doi.org/10.1016/j.concog.2014.02.003

Kim, S. K., Kirchner, E. A., Stefes, A., & Kirchner, F. (2017). Intrinsic interactive reinforcement learning–using error-related potentials for real world human-robot interaction. *Scientific Reports*, *7*(1), 1–16. https://doi.org/10.1038/s41598-017-17682-7

Kumar, A., Gao, L., Pirogova, E., & Fang, Q. (2019). A review of error-related potential-based brain–computer interfaces for motor impaired people. *IEEE Access*, *7*, 142451–142466. https://doi.org/10.1109/ACCESS.2019.2944067

Lehmann, D., & Skrandies, W. (1980). Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and Clinical Neurophysiology*, *48*(6), 609–621. https://doi.org/10.1016/0013-4694(80)90419-8

Liu, D. X., Wu, X., Du, W., Wang, C., Chen, C., & Xu, T. (2017). Deep spatial-temporal model for rehabilitation gait: Optimal trajectory generation for knee joint of lower-limb exoskeleton. *Assembly Automation*, *37*, 369–378. https://doi.org/10.1108/AA-11-2016-155

Llera, A., van Gerven, M. A., Gómez, V., Jensen, O., & Kappen, H. J. (2011). On the use of interaction error potentials for adaptive brain computer interfaces. *Neural Networks*, *24*(10), 1120–1127. https://doi.org/10.1016/j.neunet.2011.05.006

Lopez-Calderon, J., & Luck, S. J. (2014). ERPLAB: An open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, *8*(213), 1–14. https://doi.org/10.3389/fnhum.2014.00213

Manyakov, N. V., Combaz, A., Chumerin, N., Robben, A., Vliet, M. V., & Hulle, M. M. V. (2012). Feasibility of error-related potential detection as novelty detection problem in P300 mind spelling. In *International conference on artificial intelligence and soft computing* (pp. 293–301). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-29350-4-35

Margaux, P., Emmanuel, M., Sébastien, D., Olivier, B., & Jérémie, M. (2012). Objective and subjective evaluation of online error correction during P300-based spelling. *Advances in Human-Computer Interaction*, *2012*, 1–13. https://doi.org/10.1155/2012/578295

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Miall, R. C., & Wolpert, D. M. (1996). Forward models for physiological motor control. *Neural Networks*, *9*(8), 1265–1279. https://doi.org/10.1016/s0893-6080(96)00035-4

Moore, J., & Haggard, P. (2008). Awareness of action: Inference and prediction. *Consciousness and Cognition*, *17*(1), 136–144. https://doi.org/10.1016/j.concog.2006.12.004

Nielsen, T. I. (1963). Volition: A new experimental approach. *Scandinavian Journal of Psychology*, *4*(1), 225–230. https://doi.org/10.1111/j.1467-9450.1963.tb01326.x

Ort, A., Smallridge, J. W., Sarasso, S., Casarotto, S., von Rotz, R., Casanova, A., Seifritz, E., Preller, K. H., Tononi, G., & Vollenweider, F. X. (2023). TMS-EEG and resting-state EEG applied to altered states of consciousness: Oscillations, complexity, and phenomenology. *iScience*, *26*(5), 1–16. https://doi.org/10.1016/j.isci.2023.106589

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

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Pijnacker, J., Geurts, B., Van Lambalgen, M., Buitelaar, J., & Hagoort, P. (2010). Exceptions and anomalies: An ERP study on context sensitivity in autism. *Neuropsychologia*, *48*(10), 2940–2951. https://doi.org/10.1016/j.neuropsychologia.2010.06.003

Rabbitt, P. M. A. (1966). Error correction time without external error signals. *Nature*, *212*, 438. https://doi.org/10.1038/212438a0

Rakshit, A., Konar, A., & Nagar, A. K. (2020). A hybrid brain-computer interface for closed-loop position control of a robot arm. *IEEE/CAA Journal of Automatica Sinica*, *7*(5), 1344–1360. https://doi.org/10.1109/JAS.2020.1003336

Rodriguez-Fornells, A., Kurzbuch, A. R., & Münte, T. F. (2002). Time course of error detection and correction in humans: Neurophysiological evidence. *The Journal of Neuroscience*, *22*(22), 9990–9996. https://doi.org/10.1523/JNEUROSCI.22-22-09990.2002

Rotermund, D., Ernst, U. A., & Pawelzik, K. R. (2006). Towards online adaptation of neuro-prostheses with neuronal evaluation signals. *Biological Cybernetics*, *95*(3), 243–257. https://doi.org/10.1007/s00422-006-0083-7

Sato, A., & Yasuda, A. (2005). Illusion of sense of self-agency: Discrepancy between the predicted and actual sensory consequences of actions modulates the sense of self-agency, but not the sense of self-ownership. *Cognition*, *94*(3), 241–255. https://doi.org/10.1016/j.cognition.2004.04.003

Schalk, G., Wolpaw, J. R., McFarland, D. J., & Pfurtscheller, G. (2000). EEG-based communication: Presence of an error potential. *Clinical Neurophysiology*, *111*(12), 2138–2144. https://doi.org/10.1016/s1388-2457(00)00457-0

Schmidt, N. M., Blankertz, B., & Treder, M. S. (2012). Online detection of error-related potentials boosts the performance of mental typewriters. *BMC Neuroscience*, *13*(1), 1–13. https://doi.org/10.1186/1471-2202-13-19

Skrandies, W. (1990). Global field power and topographic similarity. *Brain Topography*, *3*(1), 137–141. https://doi.org/10.1007/BF01128870

Spüler, M., Bensch, M., Kleih, S., Rosenstiel, W., Bogdan, M., & Kübler, A. (2012). Online use of error-related potentials in healthy users and people with severe motor impairment increases performance of a P300-BCI. *Clinical Neurophysiology*, *123*(7), 1328–1337. https://doi.org/10.1016/j.clinph.2011.11.082

Synofzik, M., Vosgerau, G., & Newen, A. (2008). I move, therefore I am: A new theoretical framework to investigate agency and ownership. *Consciousness and Cognition*, *17*(2), 411–424. https://doi.org/10.1016/j.concog.2008.03.008

Taylor, S. F., Stern, E. R., & Gehring, W. J. (2007). Neural systems for error monitoring: Recent findings and theoretical perspectives. *The Neuroscientist*, *13*(2), 160–172. https://doi.org/10.1177/1073858406298184

Tonin, L., & Millán, J. D. R. (2021). Noninvasive brain–machine interfaces for robotic devices. *Annual Review of Control, Robotics, and Autonomous Systems*, *4*, 191–214. https://doi.org/10.1146/annurev-control-012720-093904

Tsakiris, M., Prabhu, G., & Haggard, P. (2006). Having a body versus moving your body: How agency structures body-ownership. *Consciousness and Cognition*, *15*(2), 423–432. https://doi.org/10.1016/j.concog.2005.09.004

Ullsperger, M., Danielmeier, C., & Jocham, G. (2014). Neurophysiology of performance monitoring and adaptive behavior. *Physiological Reviews*, *94*(1), 35–79. https://doi.org/10.1152/physrev.00041.2012

Usama, N., Niazi, I. K., Dremstrup, K., & Jochumsen, M. (2021). Detection of error-related potentials in stroke patients from EEG using an artificial neural network. *Sensors*, *21*(18), 6274. https://doi.org/10.3390/s21186274

van Schie, H. T., Mars, R. B., Coles, M. G., & Bekkering, H. (2004). Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience*, *7*(5), 549–554. https://doi.org/10.1038/nn1239

Zander, T. O., Krol, L. R., Birbaumer, N. P., & Gramann, K. (2016). Neuroadaptive technology enables implicit cursor control based on medial prefrontal cortex activity. *Proceedings of the National Academy of Sciences*, *113*(52), 14898–14903. https://doi.org/10.1073/pnas.1605155114

Zhang, T., & Huang, H. (2018). A lower-back robotic exoskeleton: Industrial handling augmentation used to provide spinal support. *IEEE Robotics & Automation Magazine*, *25*(2), 95–106. https://doi.org/10.1109/MRA.2018.2815083

Zito, G. A., Wiest, R., & Aybek, S. (2020). Neural correlates of sense of agency in motor control: A neuroimaging meta-analysis. *PLoS One*, *15*(6), e0234321. https://doi.org/10.1371/journal.pone.0234321

How to cite this article: Gomez-Andres, A., Cerda-Company, X., Cucurell, D., Cunillera, T., & Rodríguez-Fornells, A. (2024). Decoding agency attribution using single trial error-related brain potentials. *Psychophysiology*, *61*, e14434. https://doi.org/10.1111/psyp.14434