International Journal of Social Robotics
https://doi.org/10.1007/s12369-018-0501-8

A Feasibility Study for Validating Robot Actions Using EEG-Based Error-Related Potentials

Stefan K. Ehrlich[1] · **Gordon Cheng**[1]

Accepted: 19 October 2018

Abstract
Validating human–robot interaction can be a challenging task, especially in cases in which the robot designer is interested in the assessment of individual robot actions within an ongoing interaction that should not be interrupted by intermittent surveys. In this paper, we propose a neuro-based method for real-time quantitative assessment of robot actions. The method encompasses the decoding of error-related potentials (ErrPs) from the electroencephalogram (EEG) of a human during interaction with a robot, which could be a useful and intuitive complement to existing methods for validating human–robot interaction in the future. To demonstrate usability, we conducted a study in which we examined EEG-based ErrPs in response to a humanoid robot displaying semantically incorrect actions in a simplistic HRI task. Furthermore, we conducted a procedurally identical control experiment with computer screen-based symbolic cursor action. The results of our study confirmed decodeability of ErrPs in response to incorrect robot actions with an average accuracy of $69.0 \pm 7.9\%$ across 11 subjects. Cross-comparisons of ErrPs between experimental tasks revealed high temporal and topographical similarity, but more distinct signals in response to the cursor action and, as a result, better decodeability with a mean accuracy of $90.6 \pm 3.9\%$. This demonstrated that ErrPs can be sensitive to the stimulus eliciting them despite procedurally identical protocols. Re-using ErrP-decoders across experimental tasks without re-calibration is accompanied by significant performance losses and therefore not recommended. Overall, the outcomes of our study confirm feasibility of ErrP-decoding for human–robot validation, but also highlight challenges to overcome in order to enhance usability of the proposed method.

Keywords Electroencephalography (EEG) · Passive brain–computer interface (BCI) · Error-related potentials (ErrP) · Event-related potentials (ERP) · Error monitoring · Human–robot interaction (HRI)

1 Introduction

1.1 Validating Human–Robot Interaction

More than a decade of research on human–robot interaction [22] has been dedicated to the question of how to make interaction with robots more intuitive and "natural" for the human user [10]. In this regard, the assessment and validation of robot behavior during interaction with humans is crucial for successfully directing technical improvements towards more widespread and effective integration of robots in society. This task can be particularly challenging in the domain of humanoid and social robotics. Typical scenarios include collaboration tasks in shared environments [21,29] and game-based or dialogue social interaction tasks [28,53]. For robot validation, the human's subjective experience of the robot's behavior is usually assessed with survey-based methods, such as questionnaires [3,4,34] or interviews [41]. The quality of interaction is often additionally assessed with objective performance measures, such as "time to complete the task", as in [21,29]. To avoid interruption of the interaction flow, these measures are often taken at the end of a task and as such constitute an average assessment of the performed sequence of actions or events. However, the assessment of individual robot actions may be beneficial or,

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s12369-018-0501-8) contains supplementary material, which is available to authorized users.

✉ Stefan K. Ehrlich
stefan.ehrlich@tum.de

Gordon Cheng
gordon@tum.de

[1] Chair for Cognitive Systems, Department of Electrical and Computer Engineering, Technical University of Munich (TUM), Arcisstrasse 21, 80333 Munich, Germany

Published online: 02 November 2018

in specific cases, even required to effectively pinpoint factors that have influenced the interaction.

1.2 EEG-Based Validation of Human–Robot Interaction Using Error-Related Potentials

In this paper, we propose a neuro-based method for real-time quantitative assessment of the human perception of individual robot actions immediate to their occurrence, see also [12]. We understand the proposed method as a potentially useful complement to existing methods for robot validation, particularly for cases in which the robot designer is interested in the assessment of robot actions embedded in ongoing interaction that should not or cannot be interrupted by intermittent surveys. The method proposed here encompasses the measurement of brain activity during interaction with a robot. Brain responses, time-locked to the occurrence of robot actions, are captured and analyzed. In case the human observes incorrect robot actions, we expect the occurrence of deviating brain responses compared to the observation of correct robot actions. Classifying these responses from the ongoing EEG signals allows for implicit and real-time labeling of single robot actions immediate to their occurrence in a binary fashion (see Fig. 1). In a collaborative assembly task, an incorrect robot action may be the robot providing the human a wrong object for the next step in the assembly. In a game-based or dialogue interaction task, an incorrect robot action may be the robot performing a social cue, e.g. gaze contact with the human, in an unexpected, contextually inappropriate moment.

1.3 Error-Related Brain Responses

Prior research has found that human brain activity is modulated by both performed [6,16,25,40] and observed [48] erroneous actions. This neural process is understood to be related to error-/performance monitoring, crucial for goal-directed behavior, decision making, planning and execution of tasks, error handling as well as learning [1,45]. Related neural modulations, believed to originate from the prefrontal cortex, mainly the anterior cingulate cortex [1], have been shown to be observable as signal deflections in the human electroencephalogram (EEG), termed error-related potentials (ErrP) [16,25], a specific type of event-related potentials (ERP) [5]. Detecting human-observed errors from the ongoing EEG has already been employed in the field of brain–computer interfaces (BCI), which were originally developed to provide a communication channel between human and machine using brain activity only [56]. In the context of BCIs for communication and control, ErrPs have been used to automatically detect erroneous feedback from the BCI [8,17], and to use this information for correction or adaptation of the BCI [18,38,43,50,52]. These works showed that ErrPs can be used to improve the precision of decoding the mental commands from EEG signals and as such enhance the quality of interaction between human and device. The same principle has been applied to the domain of robotics for automatic improvement of the robotic device [30,37,55]. Recent works have shown that ErrPs are useful as feedback for reinforcement learning of robot behavior, e.g. robot trajectories [32], association of objects in a sorting task [46], as well as recognition and imitation of human gestures [36].

1.4 Aim of the Present Study

Most previous studies on ErrPs used computer screen-based interaction with perceptually simple symbolic stimuli [17,18,31,51,52]. Others have studied ErrPs in the context of human–robot interaction [30,32,36,37,46]. An important fundamental question is whether and to what extend observability and decodeability of EEG-based ErrPs are transferrable between different types of stimuli. In the present study, we examined the observability and decodeability of EEG-based ErrPs in response to a humanoid robot displaying incorrect actions in a simplistic HRI task, where robot actions either conform (congruent, correct action) or disconform (incongruent, wrong action) to a selection the human partner made. Furthermore, we conducted a procedurally identical control experiment with computer screen-based symbolic (cursor) action. To the best of our knowledge, no such comparative studies are currently available. With this comparative experimental design, we address the following items with practical implications on the usability of deploying EEG-based ErrP-decoding for validating human–robot interaction:

- We verify the observability and decodeability of EEG-based ErrPs in response to the human observation of semantically incorrect robot actions.
- We demonstrate that the type of stimulus executing the action (cursor, robot) cause variations in the EEG signals and consequently affect the observability and decodeability of ErrPs, despite procedurally identical experimental protocols and the use of the same decoding method.
- We investigate the possibility of deploying ErrP-decoders across experimental tasks, e.g. calibrate the decoder based on one task and apply it for decoding ErrPs based on the other task. Despite identical experimental protocols, we show that this task transfer is accompanied by a significant reduction in decoding performance and discuss possibilities to overcome this challenge.

By addressing the first item, we aim to confirm feasibility of detecting ErrPs in response to human observation of incorrect robot actions. The second item concerns the understanding whether and to what extent the type of stimulus can

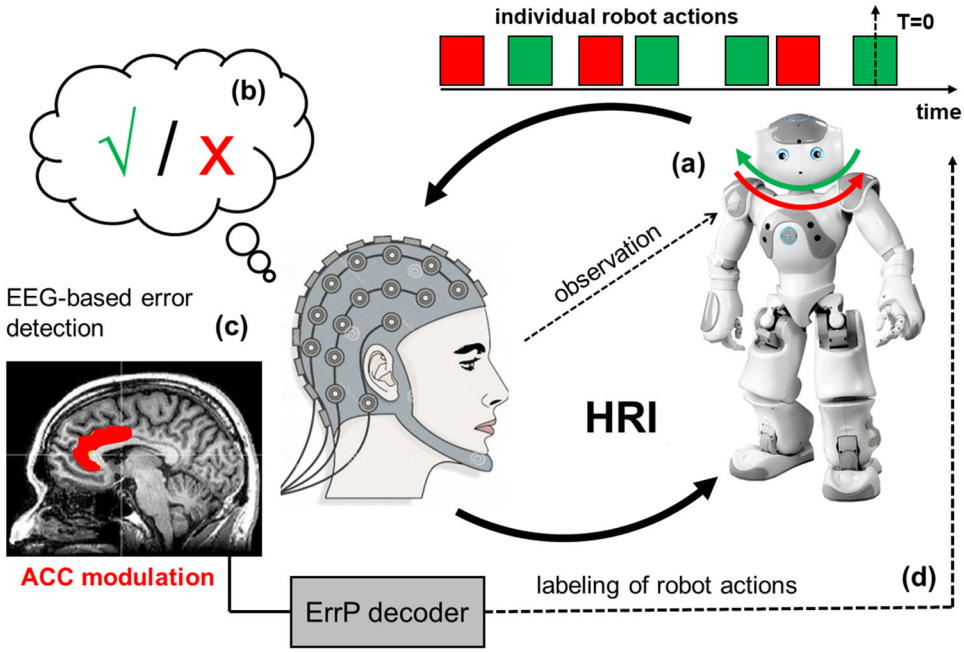

Fig. 1 Conceptual illustration of using ErrPs for robot validation: **a** during interaction with a robot, a human observes the behaving robot. Meanwhile, the brain activity is recorded via electroencephalography and responses to single robot actions are captured and analyzed. **b** Brain responses associated to the observation of incorrect robot action are expected to deviate from those associated to the observation of correct actions. These deviating responses are believed to originate from differential activation of the anterior cingulate cortex (ACC). **c** Classifying these EEG responses allows implicit real-time labeling of single robot actions immediate to their occurrence, usable for **d** post-hoc validation of the robot behavior. (Color figure online)

affect the observability of ErrPs and if this results in altered efficiency for decoding ErrPs in response to robot actions. By addressing the third item, we aim at validating the feasibility of re-using ErrP-decoders across experimental tasks (choice-reaction time tasks with identical protocols) without the need for re-calibration. After having introduced the motivation and objectives of the study here presented, the remainder of this paper is structured as follows: In Sect. 2, the design (Sect. 2.2) and implementation (Sect. 2.3) of the study as well as data analysis (Sect. 2.4) are presented. The corresponding results are reported in Sect. 3 and interpreted and discussed in Sect. 4. Section 5 concludes the paper.

2 Experimental Study and Data Analysis

2.1 Participants

Thirteen healthy participants took part in the experiment. The data of two participants were excluded from further analysis: subject s01 due to data corruption during the experiment, and subject s12 due to having been on medication during the experiment. The remaining 11 participants were 6 males and 5 females with average age: 29.4 $\pm$ 7.4 years. Prior experience and familiarity with humanoid robots scored 2.8 $\pm$ 1.8 on a scale of 1 "unfamiliar" to 7 "familiar". The participants were equally instructed about the experiment protocol and provided informed consent regarding participation in the experiment. Each participant was paid an honorarium of 8 EUR/h. The study was approved by the institutional ethics review board of the Technical University of Munich.

2.2 Experimental Tasks and Protocol

Traditionally, the study of ErrPs was largely performed using choice-reaction time tasks (CRT) and variants [16,25]. To allow for systematic comparisons with the current body of literature on ErrPs and draw from a well-established paradigm, the experimental tasks were designed based on the principle of a CRT task with identical protocols in both tasks (Fig. 2): One out of three possible target stimuli appeared on a computer screen and participants were requested to respond, with a corresponding key-press, as quickly and precisely as possible. Feedback to the participant key-press was one of two possible outcomes: congruent or incongruent response to the target stimulus. The two experimental tasks differed only in the type of feedback presented to the participant: In experimental task 1 (*cursor scenario*), a cursor, centrally placed on the computer screen, would either move towards (i.e. congruent) or away (i.e. incongruent) from the target stimulus (Fig. 2, left); in experimental task 2 (*robot scenario*), the head of a humanoid robot would either turn towards (i.e. congruent) or away (i.e. incongruent) from the target stimulus (Fig. 2, right). Robot head turns as the corresponding counterpart to the cursor actions were chosen for having the robot performing actions which can be understood in the sense of gaze cues [20] and as such being useful for real-world HRI tasks [42]. This way we realized an experimental setting with two tasks of identical procedure, but different connotation: while the cursor scenario follows traditional protocols used in the study of ErrPs, the robot scenario deploys an embodied humanoid robotic presence performing actions that can be useful in realistic human robot social interaction. In both

experimental tasks, we expected to observe different event-related potentials (ERPs) in the participants EEG evoked by congruent (semantically correct) versus incongruent (semantically incorrect) feedback. By varying the type/form of the feedback only, our design allowed us to test whether and how the observability and decodeability of ErrPs evoked by simplistic symbolic feedback can be transferred to a scenario involving the execution of actions by a humanoid robot in a procedurally identical protocol.

Experimental Setting The experiment took place in a quiet room which was partitioned into two sections by means of a sight-proof wall (Fig. 3). On the right side of the room, a participant was comfortably seated in front of the computer screen/the humanoid robot. The computer keyboard for capturing the participant responses was located in near distance to the participant to allow for comfortable access. All but one participant (s07) performed the key presses with their right hand. The left side of the room was reserved for the experimenter monitoring the experiment protocol and a live visualization of the recorded EEG data.

Experimental Protocol The experimental protocol was divided into two recording sessions (one for each scenario) which took place one after another. About half of the participants (6 out of 11, cf. Supplementary Table 1) started with the cursor scenario and the others with the robot scenario. Each scenario was further divided into 10 blocks of 50 trials each; the duration of one block was approximately 2.5 min. Thus, the total duration of the experiment was around 60 min. After each block, the participant would take a rest and decide when to continue with the next block in a self-paced fashion. The participants were first instructed (verbally and by written instruction) about the experimental setup including the recording modalities and handed a questionnaire about personal details. Participants were informed about the approximate duration of the experiment, but not about the specific number of trials per block. Participants were instructed to react as quickly and precisely as possible to the appearing target stimuli.

Trial Structure Each trial (Fig. 4) started with a pause of random duration between 500 and 2000 ms in order to avoid habituation to timing of appearance of the target stimuli. After the initial pause, one out of three possible target stimuli appeared on the screen, followed by the participants self-paced key-press. In response, the feedback was presented in form of cursor movement (cursor scenario) or robot head turn (robot scenario) towards or away from the target stimulus. The feedback ended after 130 ms: cursor reached target stimulus/robot head movement reached end location. A second feedback was presented 200 ms afterwards in form of the appearance of a colored frame around the target stimulus (green frame = correct, red frame = incorrect). The framed target stimulus disappeared 300 ms later, which initiated the robot head turning back to the initial location and the cursor re-appearing in the center of the screen 600 ms later.

2.3 Stimuli and Apparatus

Stimuli were presented on a 24-in. flat screen LCD computer monitor with 60 Hz refresh rate placed at a distance of approximately 150 cm from an observer. Participant responses were registered with the arrow keys of an ordinary computer keyboard. The experiment was programmed with Python using the Psychopy library [44] and executed on an Intel®Core™ i5 CPU 750@2.67 GHz. The target stimuli were realized as white squares of size 3x3 cm appearing in three possible locations on the computer screen (left, right, or up). Per trial, one out of the three possible target stimuli appeared. Participants were requested to respond with a corresponding arrow key (left target = left arrow key, right target = right arrow key, upper target = upper arrow key). Upon participant key-press, feedback was initiated in form of a cursor (cursor scenario) or a robot head movement (robot scenario). In case of correct participant response (response key-press congruent to the target stimulus location), false feedback events were introduced in a uniform random fashion with a pre-defined probability p_{Err} (these events are termed "machine-errors" for the rest of this paper). Machine-errors were manifested as cursor movement or robot head turns towards the wrong direction/incongruent to the target stimulus location. The wrong direction was selected in a uniform random fashion among the two remaining non-target directions. To avoid habituation to the machine error probability, half of the blocks were executed with a machine-error probability of $p_{Err} = 20\%$ and the other half with $p_{Err} = 50\%$. The order was pseudo-randomized such that no more than two subsequent blocks would belong to the same error probability category. The first block of each scenario was always executed with $p_{Err} = 20\%$ to avoid any confusion in the beginning of each scenario. In total, 500 trials were collected per participant and scenario among which were on average 35% machine-error trials and a negligible number of human-committed error trials. The cursor feedback was realized as a white square of size 2×2 cm, initially located in the center of the screen. Upon participant key-press in response to the appearance of the target stimulus, the cursor would start moving towards or away from the target with uniform speed until reaching the target position after 130 ms. In the robot scenario, the cursor was substituted with a NAO humanoid robot located in a crouched posture in front of the computer screen such that the head position matched the center of the screen. NAO is a 58 cm tall humanoid robot with 21–25 degrees of freedom [23] which was controlled by the experiment program via local area network (LAN) using the Python-based NAOqi library. In the initial head position (yaw = 0°,

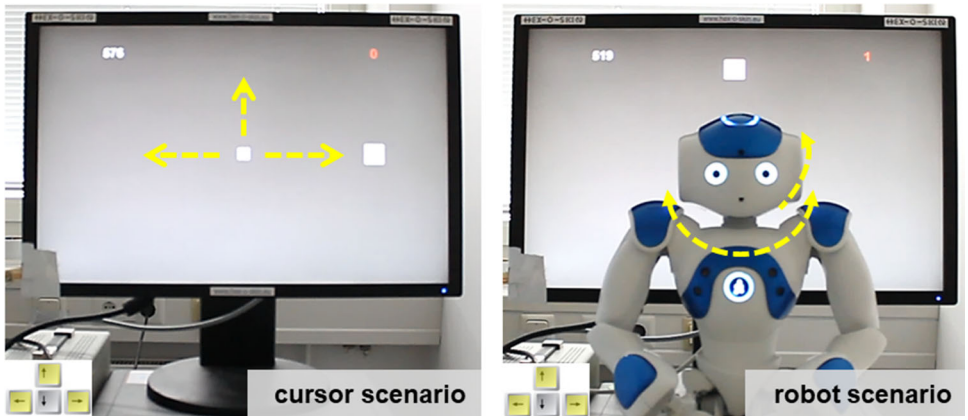

Fig. 2 Illustration of the two experimental tasks. Participants were requested to respond to one out of three possible target stimuli (left, right, up) appearing on the screen with corresponding arrow key-presses. In response, either a cursor (left) or the robot head (right) would move towards or away from the given target. The yellow dashed arrows show the directions of possible cursor movements and robot head turns. (Color figure online)

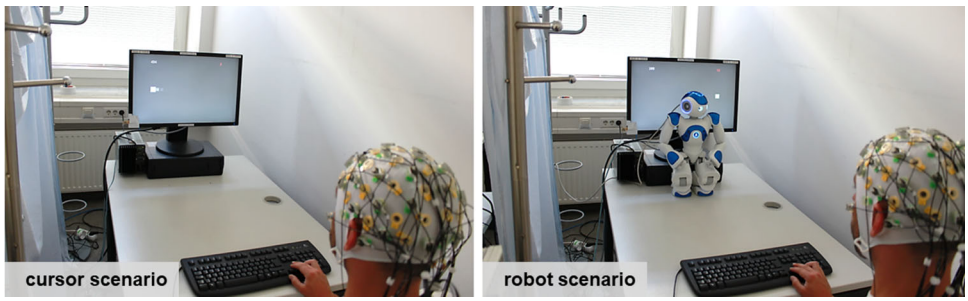

Fig. 3 The experimental setup showing a participant performing the cursor scenario (left) and the robot scenario (right). (Color figure online)

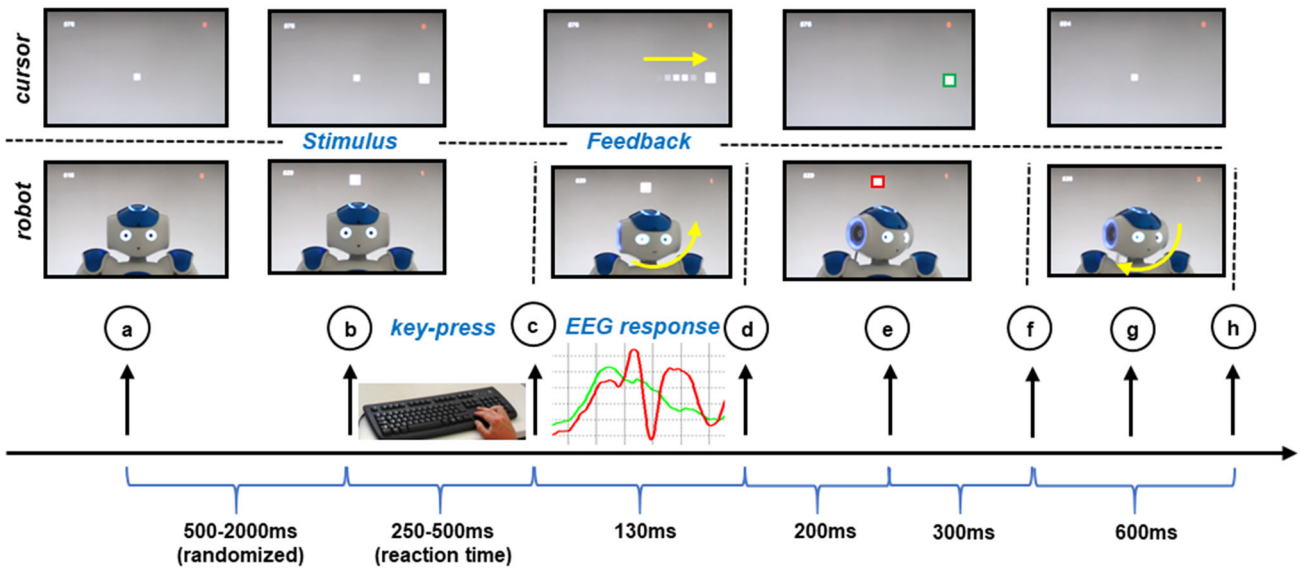

Fig. 4 Trial structure with exemplary illustration of the cursor scenario (top panel, example for a congruent/correct trial) and the robot scenario (middle panel, example for an incongruent/incorrect trial). (a) Trial start and pause of random duration between 500 and 2000 ms, (b) appearance of target stimulus, (c) participant response in form of arrow key press, start cursor/robot head movement (d) end cursor/robot head movement (e) target border feedback presentation (correct: green; incorrect: red), (f) disappearance target, disappearance cursor/start robot head turning back, (g) re-appearance cursor/ongoing robot head turning back (h) re-appearance cursor, end of robot head turning back, updating average reaction time and error count. (Color figure online)

pitch $= 0°$), the robot was gazing directly towards the participant. Upon participant key-press in response to the target stimulus, the robot head would turn towards or away from the target (left: pitch $= 0°$, yaw $= -40°$; right: pitch $= 0°$, yaw $= +40°$, up: pitch $= -20°$, yaw $= 0°$) and reach the end position after 130 ms, keep the position for 500 ms and move back to the initial head position. The left upper corner of the screen informed the participants about the average reac-

tion time per block in ms. The right upper corner informed about the number of errors per block with no distinction of errors committed by the participant or the machine. This additional feedback was shown to the subjects to keep up their engagement in the task by self-monitoring their own performance throughout the experiment.

EEG Recording and Data Pre-processing EEG data were acquired with a Brain Products actiChamp amplifier equipped with 32 active EEG electrodes arranged according to an extended international 10–20 system [27] (FP1, FP2, F3, F4, F7, F8, FC1, FC2, FC5, FC6, C3, C4, T7, T8, CP5, CP6, P3, P4, P7, P8, TP9, TP10, O1, O2, Fz, Cz, Pz, EOG1, EOG2, EOG3). All leads were referenced to the average of TP9 and TP10 (average mastoids referencing) and the sampling rate was set to 1000 Hz. The impedance levels of all leads were kept below 10 $k\Omega$. Three channels were used for capturing electrooculogram (EOG1–3) signals in three locations of the participant's face (forehead, left and right outer canthi) according to a method suggested by Schlögl et al. [49]. The EEG amplifier was battery-driven and located on a tray nearby the participant. The data was transferred via USB to a separate recording PC (Intel®Core™ i5 CPU 750@2.67 GHz). The amplifier was connected to the PC executing the experiment protocol via parallel port over which event triggers were sent to be stored synchronously with the EEG signals. All EEG data preprocessing was carried out in MATLAB®, in part using functions provided by the EEGLAB toolbox [11]. The subsequent processing steps were applied to each dataset (11 subjects × 2 sessions) separately in the following order: In order to remove high frequency and power-line noise, we first filtered the signals of the EEG and EOG channels using a zero phase Hamming windowed sinc FIR band-pass filter with cutoff frequencies of 1 Hz and 20 Hz. Next, we identified and interpolated contaminated EEG channels using kurtosis with a threshold of 5%. EOG activity in the EEG signals (eye-blink and lateral eye movements) was corrected using Schlögl et al.'s method [49]. Afterwards, we re-referenced the EEG signals to common average (CAR) to further reduce signal contamination due to external noise sources.

Stimuli Timing In order to ensure precise information about the moments of presentation of the feedback, the robot head was equipped with a light emitting diode (LED) and a photodiode to record the onset of head movements synchronously with the recording of the EEG signals. The computer screen was also equipped with a separate photodiode to record the timing of the cursor movement. Both photodiode setups were not directly visible to the participants and thus not distracting. Ground truth timing of onset of both cursor and robot head movement was obtained by analyzing the signals captured by the two photodiodes and introducing additional event markers into the EEG datasets.

2.4 Data Analysis

2.4.1 Analysis of Behavioral Data

Our experimental design featured identical protocols in both scenarios with only the appearance of feedback differing. Therefore, we did not expect systematic behavioral differences regarding reaction times (RT) and number of errors committed by the participants ($nErr$) across scenarios. The purpose of the analysis of behavioral data was to verify the absence of such systematic behavioral differences. For that, we performed several statistical tests: we tested whether the distributions of mean reaction times $\overline{RT}$ and $nErr$ differ across the two scenarios. Furthermore, we tested whether the distributions of $\overline{RT}$ and $nErr$ differed across first and second performed scenario, irrespective of type cursor or robot.

2.4.2 Electrophysiological Analysis of Error-Related Potentials (ErrP) and Their Stimulus-Dependent Variations

The data was segmented into epochs by extracting time intervals of −500 to 1500 ms relative to the presentation of the feedback (onset of cursor movement/robot head turning, t = 0 ms). These segments were further separated into three categories: (1) correct trials (non-error), (2) false feedback trials (machine-error), (3) human error trials. Per participant and recording session, we extracted on average approximately 325 correct trials, 159 machine error trials, and 16 human error trials. Since human errors were not in the focus of our investigation, the corresponding epochs were discarded from further analyses. Error-related potentials encompass the appearance of three main components on the difference (error minus correct) average time courses: an N2, a P3, and an N4 component [8,18,51]. These are a negative deflection around 200 ms, a positive deflection around 300 ms, and another negative deflection around 400 ms, time-locked to the appearance and observation of an event of type error or correct, mainly observable over fronto-central and fronto-parietal sites. An additional late positive component (P600, 600 ms latency) was frequently reported in the context of error processing in choice-reaction time tasks [15] and the experience of syntactic and semantic anomalies in language comprehension [35,54]. Our investigation encompassed the comparative analysis of the shape and timing of the potentials of each scenario. This was carried out through the computation of the time-locked average potentials for the machine-error and non-error potentials in channel Cz, through the difference average (machine-error minus non-error averages) and by the computation of the coefficient of determination r^2 [56]. Before averaging, we performed a per channel baseline correction by subtracting the average amplitude of the period 200 ms prior to the onset of the feedback

from the entire signal epoch. Spatial ERP activity patterns were compared by computing the topographic interpolation of the potentials at the time of the main peaks of the difference average. Finally, we assessed the similarity of ERP time courses within the period 0 to 800 ms per subject by computing the 2D correlation coefficients between the difference average of the cursor and the robot scenario. Furthermore, we computed the 2D correlation coefficient between the difference grand average within the period 0 to 800 ms of the cursor and the robot scenario according to Eq. 1; with C, R being the difference average ERP of cursor C and robot R (machine-error minus non-error); with $\overline{C}$, $\overline{R}$ being the means of all elements in C and R; c being the spatial dimension (channel), and t being the temporal dimension (sample time point).

$$r = \frac{\sum_c \sum_t (C_{ct} - \overline{C})(R_{ct} - \overline{R})}{\sqrt{\sum_c \sum_t (C_{ct} - \overline{C})^2 \sum_c \sum_t (R_{ct} - \overline{R})^2}} \quad (1)$$

2.4.3 Single-Trial Classification and Analysis of Impact of Stimulus-Dependent Variations

This section describes the methodology of obtaining (calibrating) subject-specific ErrP-decoders in order to classify EEG signals into responses due to the observation of non-error or machine-error events. The calibration step included the extraction of relevant features from the EEG signals and the training of a classification model. The testing/validation step included feature extraction and application of that classification model on unseen data within session and across session.

Feature Extraction In the context of single-trial classification of error-related potentials, different types of features have been used and reported in previous works. Temporal features extracted from the time series [18,30] have been used in most cases, being reported as stable and reliable even across recording sessions [7]. Furthermore, we performed a feature cross-comparison in one of our earlier works [12] where we found temporal features being superior over spectral features in the context of decoding ErrPs. Therefore, temporal features were used in this work. For each trial and each channel, the signal amplitude was averaged within 9 overlapping 100 ms-long windows, relative to the occurrence of machine-error/non-error events for each channel (windows: 100–200 ms, 150–250 ms, 200–300 ms, 250–350 ms, 300–400 ms, 350–450 ms, 400–500 ms, 450–550 ms, 500–600 ms) and concatenated into the single feature vector of length 243 (27 channels $\times$ 9 windows).

Classification The classifier used in the analysis was a regularized version of the linear discriminant analysis (rLDA) [19]. The rLDA classifier has been established as a robust method to discriminate mental states based on EEG signals in the field of brain–computer interfaces [5]. The LDA discriminant function is the hyperplane discriminating the feature space corresponding to two classes: $y(x) = sign(w^T x + b)$, with x being the feature vector, w being the normal vector to the hyperplane (or weight vector), b the corresponding bias, and $y(x) \in \{-1, 1\}$ the classifier decision. The weight vector and bias were computed by $w = (\hat{\mu_2} - \hat{\mu_1})(\tilde{\Sigma_1} + \tilde{\Sigma_2})^{-1}$ and $b = -w^T(\hat{\mu_1} + \hat{\mu_2})$, with $\hat{\mu_j}$ being the class-wise sample means, and $\tilde{\Sigma_j}$ being the class-wise regularized covariance matrices. Regularization aims at minimizing the covariance estimation error by penalizing very small and large eigenvalues. This leads to robust covariance estimates even for high dimensional feature spaces [5] as in our case. The regularized covariance matrices were computed by $\tilde{\Sigma_j} = (1 - \lambda)\Sigma_j + \lambda \nu I$, with $\lambda \in [0, 1] \subset \mathbb{R}$ being the shrinkage parameter, ν the trace (sum of diagonal elements) of Σ_j divided by the number of features, and I the identity matrix. The optimal shrinkage parameter λ was determined automatically based on the given training data using the analytic method proposed by Schäfer and Strimmer [47].

Within-Session Validation We validated the above described modeling approach within session (cursor, robot) using a 10-times-10-fold cross-validation scheme. Per session and subject, the trials were randomly split in 10 folds, 9 folds were used for model calibration and the remaining fold was used for testing. This procedure was repeated until all folds were once used for testing. The entire procedure was furthermore repeated for 10 times. Each time and fold, the number of trials per class of the calibration data was balanced by random pick and replace (please note that the number of trials per class was initially unbalanced with ˜65% non-error and ˜35% machine-error trials). This analysis allowed us to obtain an estimate of how well subject-specific ErrP-decoders would perform in classifying unseen data when being calibrated with different data of that same session. Individual classification results per time and fold were averaged and reported per session and subject as percentage of correctly classified trials (overall accuracy, $ACC = (TP + TN)/(TP + FP + FN + TN)$); percentage of correctly classified machine-error trials (true positive rate, $TPR = TP/(TP + FP)$); percentage of correctly classified non-error trials (true negative rate, $TNR = TN/(TN + FN)$); and in addition as the area (AUC) under the receiver operator curve (ROC) since this measure can be more informative when reporting classification results based on data with class imbalance.

Cross-session Validation Furthermore, we performed a per subject cross-session validation in two steps: calibration with the cursor data and testing on the robot data, calibration with the robot data and testing on the cursor data. This analysis allowed us to obtain an estimate of how well a subject-specific ErrP-decoder would perform in classifying data of one session when being calibrated with data of the

other session. The number of trials per class was balanced in the calibration data by random pick and replace. To increase the likelihood that most of the trials were used for calibration at least once, this procedure was repeated 100 times. The weight vectors w and biases b of the resulting 100 individual rLDAs were averaged to obtain a final rLDA. Testing was performed on the unbalanced data of the test session. Results are reported identical to the within-session validation as *ACC*, *TPR*, *TNR*, and *AUC*.

3 Results

3.1 Behavioral Data

Subject individual data are summarized in Supplementary Tables 1 and 2. Mean reaction times did not significantly vary across type of scenarios ($\overline{RT}_{cursor} = 421 \pm 75$ ms, $\overline{RT}_{robot} = 403 \pm 43$ ms, Wilcoxon's signed-rank test, $p = 0.123$). Number of human-committed errors did not significantly vary across type of scenario ($\overline{nErr}_{cursor} = 17 \pm 14$, $\overline{nErr}_{robot} = 16 \pm 8$, Wilcoxon's signed-rank test, $p = 0.916$). Mean reaction times did not significantly vary across first and second performed scenario irrespective of type ($\overline{RT}_{first} = 415 \pm 52$ ms, $\overline{RT}_{second} = 410 \pm 71$ ms, Wilcoxon's signed-rank test, $p = 0.240$). Number of human-committed errors did not significantly vary across first and second performed scenario irrespective of type ($\overline{nErr}_{first} = 13 \pm 10$, $\overline{nErr}_{second} = 19 \pm 13$ ms, Wilcoxon's signed-rank test, $p = 0.095$). We conclude no effect of type of scenario or scenario order on reaction times and number of human-committed errors.

3.2 Error-Related Potentials (ErrP)

Results of the electrophysiological analysis of error-related potentials (ErrP) and stimulus-dependent signal variations are depicted in Fig. 5. The left and middle panel show the grand average ERPs for each category (blue: non-error, red: machine-error) and the difference average (dashed black: machine-error minus non-error) for each scenario (left panel: cursor, middle panel: robot). In both scenarios, the shape of the difference grand averages were similar to those previously reported [8,18,51] with regard to the timing and topographical distribution of the components N2 and P3. The expected N4 component is not observable in our data. Instead, we observed another late positive component with a latency of approximately 500 ms. This effect might be related to the P600 component which has been reported in the context of error processing as well as in response to syntactic and semantic anomalies [35,54]. These findings indicate that the observed effects originated from error-/performance monitoring processes. The characteristics we observed in the difference ERPs of the cursor scenario are significantly less pronounced in the robot scenario. The grand average time courses of the robot scenario are attenuated in peak amplitudes and topologically less clearly distinguished than in the cursor scenario. This is also reflected in the results of the analysis of the coefficient of determination r^2 based on channel Cz. In both scenarios, highest grand average r^2-values were observed at similar time points (cursor: $r^2_{\max} = 0.14$ at $t = 261$, ms, robot: $r^2_{\max} = 0.04$ at $t = 263$ ms). The similarity analysis revealed a correlation coefficient of $r = 0.70$ between spatio-temporal difference grand averages of both scenarios. Subject individual computation of the 2D correlation coefficient between spatio-temporal difference averages resulted in median $r = 0.48$ and exclusively all subjects with a positive 2D correlation coefficient (subject-individual results are reported in Supplementary Table 3). We conclude that the ErrPs observable in both scenarios were qualitatively similar in terms of shape, timing, and topographical distribution. This indicates that the observed effects originated from the same underlying neural process.

3.3 Single-Trial Classification

The within-session single-trial classification results are depicted in Fig. 6 (left panel) and detailed in Supplementary Tables 4 and 5. Across-subject average classification performance for the cursor scenario resulted in $90.6 \pm 3.9\%$ accuracy with TPR: $87.3 \pm 4.3\%$, TNR: $92.2 \pm 3.8\%$ and AUC: 0.95 ± 0.03. For the data of the robot scenario, we observed significantly reduced and higher variant within-session classification performance of $69.0 \pm 7.9\%$ accuracy with TPR : $66.1 \pm 6.5\%$, TNR: $70.6 \pm 9.1\%$, and AUC: 0.73 ± 0.1. The reduced classification performance based on the data of the robot scenario was systematic across all subjects (ACC: $-21.6 \pm 7.8\%$, TPR: $-21.2 \pm 8.1\%$, TNR: $-21.7 \pm 8.2\%$, AUC: -0.21 ± 0.1), however, with non-significant correlation between classification results of each session ($r_{ACC} = 0.27$, $p = 0.41$; $r_{AUC} = 0.17$, $p = 0.61$, Pearson's correlation, n = 11). All subject-individual classification accuracies were above the sample size adapted chance-level of 53.6% ($p < 0.05$) for binary classification according to [9].

The cross-session single-trial classification results are depicted in Fig. 6 (right upper and lower panel) and detailed in Supplementary Tables 6 and 7. For calibration with the cursor data and testing with the robot data, across-subject average classification performance resulted in $68.3 \pm 6.1\%$ accuracy with TPR: $34.9 \pm 13.0\%$, TNR: $86.4 \pm 10.8\%$, and AUC: 0.68 ± 0.11 (Fig. 6, right upper panel). Cross-session classification performance showed high correlation with subject-individual spatio-temporal ERP similarity measures ($r_{ACC} = 0.80$, $p = 0.003$; $r_{AUC} = 0.79$, $p = 0.003$, Pearson's correlation, n = 11). Except for subject s10, clas-

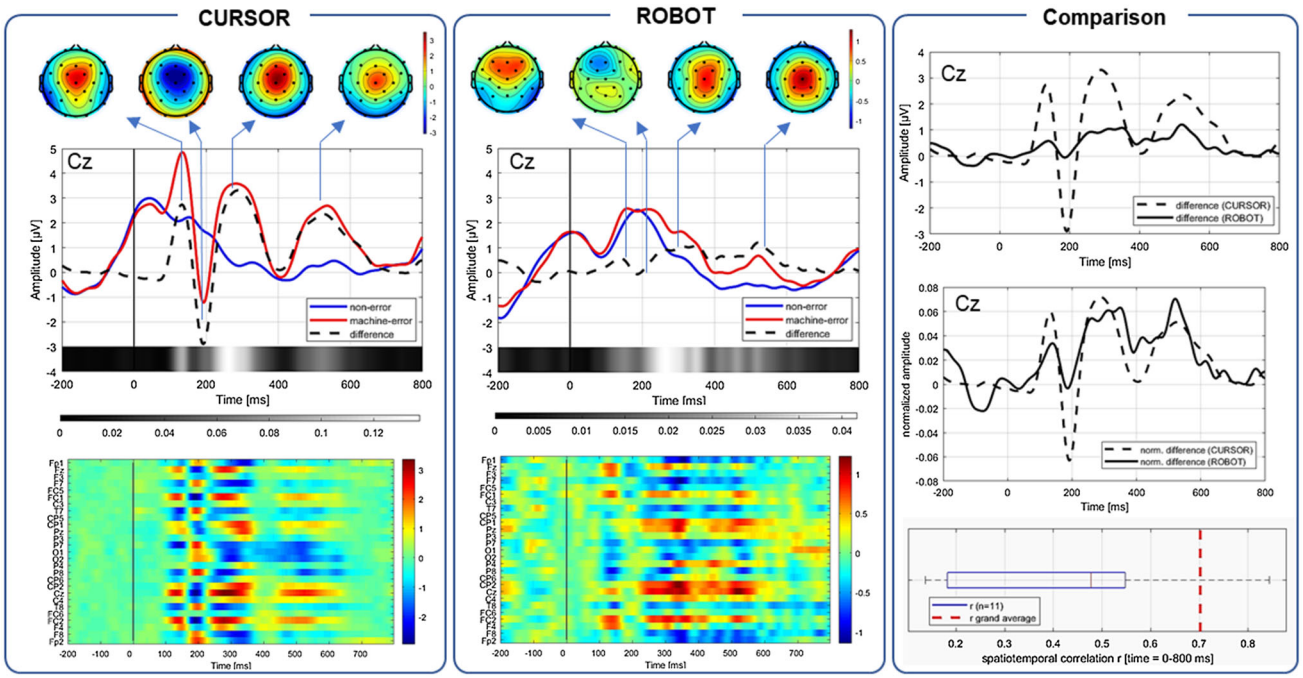

Fig. 5 Grand average signals over Cz time-locked to the onset of feedback (cursor movement, robot head turn) for each category (blue: non-error, red: machine-error) and the difference grand average (dashed black: machine-error minus non-error), for both scenarios (left panel: cursor, middle panel: robot). The r^2-values for between non-error and machine-error are depicted below each plot, where bright colors indicate high values. The difference grand average is furthermore depicted as topographic plots for the main peaks above each plot and in form of a spatio-temporal activity matrix across all channels and time points below each plot. The right panel shows the resemblance of difference (and normalized difference) grand averages of channel Cz of both scenario as well as the across-subject distribution of 2D correlation coefficients between difference average of cursor and robot scenario with median $r = 0.42$. The dashed red line depicts the difference grand average 2D correlation coefficient $r = 0.70$. Please note that different axes scaling was used to facilitate qualitative comparisons. (Color figure online)

sification accuracies were above chance-level, however with a systematic bias between TPR and TNR: across all subjects, we observed consistent low classification rates for class machine-error and high classification rates for class non-error. This is most likely related to shifts in the distributions of features favored by the rLDA classifier for separating the cursor data, causing the decision boundary to favor one class over the other in the robot data. For calibration with the robot data and testing with the cursor data, average classification performance resulted in $73.1 \pm 13.0\%$ accuracy with TPR: $70.3 \pm 24.8\%$, TNR: $74.7 \pm 11.6\%$, and AUC: 0.78 ± 0.18 (Fig. 6, right lower panel). No classification bias was observed for the robot-to-cursor transfer, but a decrease of accuracy compared to the within-cursor validation accuracies. This is most likely related to the rLDA favoring features in the robot data which have less discriminative power in the cursor data. Cross-session classification performance showed high correlation with subject-individual spatio-temporal ERP similarity measures ($r_{ACC} = 0.57$, $p = 0.06$; $r_{AUC} = 0.61$, $p = 0.05$, Pearson's correlation, $n = 11$). In all but two subjects (s08, s10), classification accuracies were close to or above 70% with an $AUC > 0.73$ and no systematic bias between TPR and TNR. The fact that s08 and s10 revealed also very low accuracies in the within-session validation of the robot data may explain why their cross-session classification results turned out to be low as well.

4 Discussion

ErrPs are decodeable in response to incorrect robot actions, but decoding performance is sensitive to the stimulus

With the results of our study, we confirm feasibility of decoding ErrPs in response to the human observation of semantically incorrect robot action. We obtained a classification accuracy of average $ACC_{robot} = 69.0 \pm 7.9\%$ across 11 subjects, which is comparable to results in response to robot actions obtained by others [32,46,55]. For the cursor task, we obtained an average classification accuracy of $ACC_{cursor} = 90.6 \pm 3.9\%$ which is comparable or higher than previously reported single-trial classification results based on ErrPs in response to screen-based stimuli [8,18,51]. From these results, we draw the conclusion that the observability and

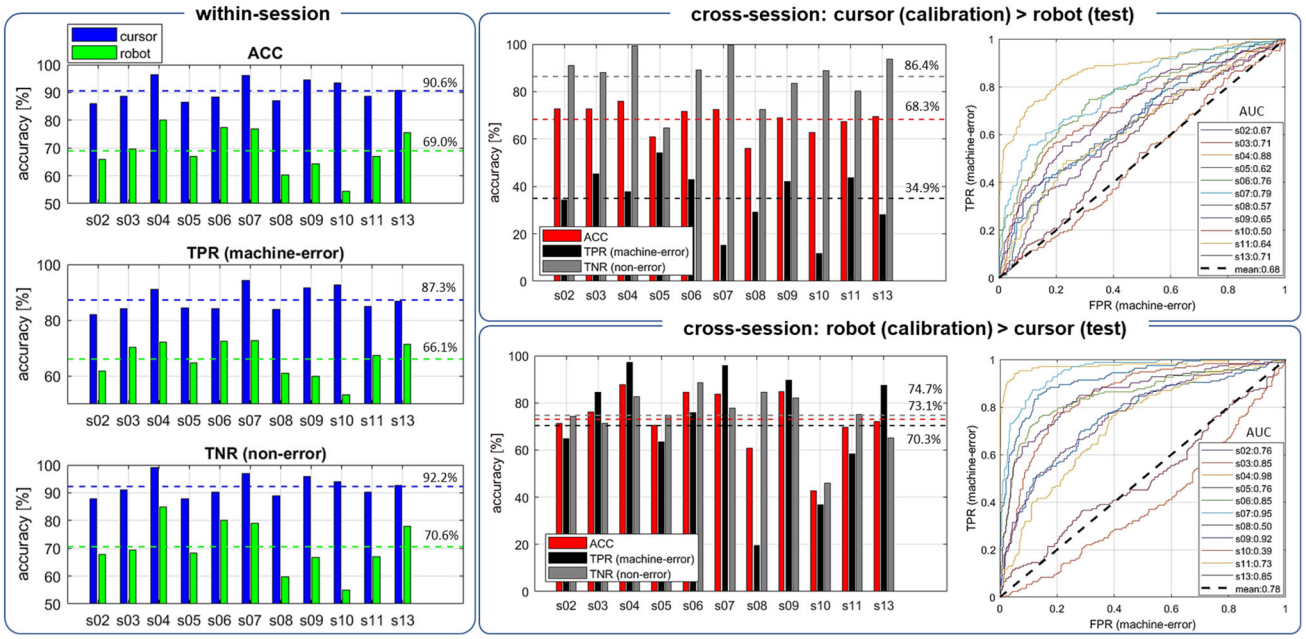

Fig. 6 Single-trial classification results: Per subject and average within-session classification results (left panel). Per subject and average cross-session classification results with calibration based on cursor data and validation on robot data (right upper panel) and calibration based on robot data and validation on cursor data (right lower panel). Classification results are reported as percentage of correctly classified machine-error events (TPR), percentage of correctly classified non-error events (TNR), total percentage of correctly classified events (ACC), and area under receiver operator curve (AUC). (Color figure online)

decodeability of ErrPs is sensitive to the appearance of the stimulus eliciting them, even in procedurally and semantically simplistic tasks. Besides the lower accuracy in the robot task, we observed higher variations of decoding accuracy indicating that subjects seemed to have responded more differently to the robot than to the cursor, e.g. s04, s06, s07, s13 with $ACC_{robot} > 75\%$ and s08, s09, s10 with $ACC_{robot} < 65\%$, and s10 close to chance level (c.f. Supplementary Table 5). We hypothesize that the variations of observability and decodeability of ErrPs across experimental tasks are most likely related to the differential perceptual complexity of the feedback stimuli eliciting them. A follow-up study examining ErrPs in response to a wider range of stimuli with systematic variation of complexity may help to explain the observed variations. The ErrPs observable in both scenarios were qualitatively similar in terms of shape, timing, and topographical distribution. This indicated that the observed effects originated from the same underlying neural process. Based on this finding, we tested whether ErrP-decoders calibrated with the data of one task can be re-used to classify responses based on the other task. The cross-session single-trial classification analysis resulted in an across-subject consistent classification bias for the cursor to robot transfer (high classification rate for non-error and low classification rate for machine-error events) and a significant drop of classification accuracy for the robot-to-cursor transfer compared to the respective within-session results. Based on these results, we do not recommend a straightforward re-use of the ErrP decoders across tasks without re-calibration. Despite the observed variation between experimental tasks, our results demonstrated and confirmed feasibility of decoding ErrPs in response to the observation of semantically incorrect robot actions given that ErrP-decoders are calibrated based on data of the same task. In reference to the huge potential of ErrP-decoding shown by others [32,36,46] and combined with the benefit of real-time assessment of individual robot actions, we are confident that the proposed method can be of substantial help in validating human–robot interaction in the future.

ErrPs can potentially be used in more complex HRI scenarios

We purposely designed a relatively simple HRI scenario to demonstrate principal feasibility of decoding the human perception of semantically incorrect robot actions. Furthermore, the chosen experimental paradigm allowed us to compare the ErrP responses in the robot task to those obtained in response to a simplistic stimulus in a procedurally identical task. The robot scenario in our study resembles most that of a real-world HRI task, in which the robot designer is interested in validating the congruency of robot gaze cues with the human partner's expectations. In a recent follow-up study we investigated as to which extent the results of the present

study are transferable to such a more realistic HRI task. The results of this work showed that ErrPs, online decoded from the human interaction partner, can be used to both validate the congruency of robot gaze cues with the human partner's expectations and likewise to successfully adapt the robot's gaze behavior during interaction [13]. Whether and to what extent the outcomes of our present study scale to even more complex HRI scenarios with different types of robot actions remains open for future work. In light of the findings we can, however, state that the brain responses we observed in both experimental tasks were related to error/performance monitoring due to their temporal and topographical similarity with previously reported ErrPs in different experimental paradigms [18,51]. This supports the notion that we indeed observed and decoded the effects of a high-level "generic" neural process, that is understood to be largely unrelated to the situational context or the associated stimulus [26,40]. A relevant research question for follow-up investigations is whether and how these observed effects extrapolate to responses due to robot actions that are not per se categorized into semantically correct or wrong. Possible examples are situations in which robots perform correct actions, but in unexpected or inappropriate moments, or situations in which the judgment of correctness of actions is dependent on the human's subjective interpretation. The study of ErrPs in the context of human–robot interaction is therefore always a co-investigation of both the technical system and the human with potential contributions to a better understanding of both sides.

Practicality in ErrP-based validation of HRI denotes challenges to overcome

A few aspects render the method of ErrP-decoding for robot validation laborious and impractical to be readily deployable: Firstly, the cumbersome, expensive, and sensitive EEG setup, and secondly, the necessity for subject-specific (re-)calibration of the ErrP-decoder. The need for inexpensive and easy-to-use EEG systems with sufficient signal quality has already been recognized by the BCI community [24]. Along this line of research, we made a contribution in the form of the development of a simple, mobile, and comparably inexpensive (˜800USD) EEG system [14]. Our device was deployed in a study investigating sensorimotor rhythms in patients suffering from cerebral palsy while performing an adapted serial reaction time task [2]; the usability of our device for measuring and decoding ErrPs remains to be tested. Subject-specific (re-)calibration is a generally recognized issue thwarting practicality of brain–computer interfaces [39]. Non-stationarities and dissimilarities of EEG signals across recording sessions and subjects generally render (re-)calibration a necessity for sufficient functionality of BCIs. ErrPs have, however, been shown to be stable across recording sessions within the same subject and experimental task up to 600 days [7,18]. This indicates that re-using ErrP-decoders without re-calibration is principally possible if subject and task remain the same across experimental sessions. Furthermore, cross-task transfer learning has been shown to significantly reduce calibration based on just a few observations of the new task [31,33].

5 Conclusions

In this paper, we presented a neuro-based method for real-time quantitative assessment of robot actions during HRI using EEG-based error-related potentials. To demonstrate usability, we conducted a study examining the observability and decodeability of ErrPs in response to incorrect robot actions in comparison to responses to simplistic computer screen-based cursor actions. The results of our study demonstrated decodeability of ErrPs in response to incorrect robot actions with an average accuracy of $69.0 \pm 7.9\%$. This supports feasibility of our proposed method of using ErrPs for validating robot behavior. Comparisons across experimental tasks revealed more distinct signals in response to the cursor action and, as a result, better decodeability with a mean accuracy of $90.6 \pm 3.9\%$. This demonstrated that ErrPs can be sensitive to the stimulus eliciting them despite procedurally identical protocols. The observed ErrPs were qualitatively similar across experimental tasks, indicating that they originated from the same underlying neural process. However, a straightforward re-use of ErrP decoders across experimental tasks without re-calibration is accompanied by performance losses and therefore not recommended. Overall, the outcomes of our study confirm feasibility of ErrP-decoding for human–robot validation, but also highlight challenges to overcome in order to enhance usability of the proposed method.

Acknowledgements We thank Ana Alves-Pinto and Sae Franklin for helpful comments on revision and editing of the manuscript. We thank the anonymous reviewers for their detailed comments and references, which have led to significant clarification of the work in this paper. This research was partially supported by Deutsche Forschungsgemeinschaft (DFG) through the International Graduate School of Science and Engineering (IGSSE) at the Technical University of Munich (TUM).

Compliance with Ethical Standards

Conflict of interest The authors declare no conflict of interest nor competing financial interests.

Ethics Approval This work was approved by the ethics commission of the Faculty of Medicine, Technische Universität München (TUM) under the Reference Number 236/15s.

Informed Consent Consent to participate and publish was obtained from the participants in verbal and written form.

Availability of Data and Material Data and material was not made publicly available but can be obtained from the corresponding author.

References

1. Alexander WH, Brown JW (2011) Medial prefrontal cortex as an action-outcome predictor. Nat Neurosci 14(10):1338
2. Alves-Pinto A, Ehrlich S, Cheng G, Turova V, Blumenstein T, Lampe R (2017) Effects of short-term piano training on measures of finger tapping, somatosensory perception and motor-related brain activity in patients with cerebral palsy. Neuropsychiatr Dis Treat 13:2705
3. Bartneck C, Croft E, Kulic D (2008) Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. In: Metrics for HRI workshop, technical report, Citeseer, vol 471, pp 37–44
4. Bartneck C, Kulić D, Croft E, Zoghbi S (2009) Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. Int J Soc Robot 1(1):71–81
5. Blankertz B, Lemm S, Treder M, Haufe S, Müller KR (2011) Single-trial analysis and classification of ERP components—a tutorial. NeuroImage 56(2):814–825
6. Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD (2001) Conflict monitoring and cognitive control. Psychol Rev 108(3):624
7. Chavarriaga R, Millán JR (2010) Learning from eeg error-related potentials in noninvasive brain–computer interfaces. IEEE Trans Neural Syst Rehabil Eng 18(4):381–388
8. Chavarriaga R, Sobolewski A, Millán JR (2014) Errare machinale est: the use of error-related potentials in brain–machine interfaces. Front Neurosci 8:208
9. Combrisson E, Jerbi K (2015) Exceeding chance level by chance: the caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. J Neurosci Methods 250:126–136
10. Dautenhahn K, Woods S, Kaouri C, Walters ML, Koay KL, Werry I (2005) What is a robot companion-friend, assistant or butler? In: 2005 IEEE/RSJ international conference on intelligent robots and systems, 2005 (IROS 2005). IEEE, pp 1192–1197
11. Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial eeg dynamics including independent component analysis. J Neurosci Methods 134(1):9–21
12. Ehrlich S, Cheng G (2016) A neuro-based method for detecting context-dependent erroneous robot action. In: 2016 IEEE-RAS 16th international conference on humanoid robots (humanoids). IEEE, pp 477–482
13. Ehrlich S, Cheng G (2018) Human-agent co-adaptation using error-related potentials. J Neural Eng 15:066014
14. Ehrlich S, Alves-Pinto A, Lampe R, Cheng G (2017) A simple and practical sensorimotor EEG device for recording in patients with special needs. In: Neurotechnix2017, CogNeuroEng 2017
15. Falkenstein M, Hohnsbein J, Hoormann J, Blanke L (1991) Effects of crossmodal divided attention on late erp components. II. Error processing in choice reaction tasks. Electroencephalogr Clin Neurophysiol 78(6):447–455
16. Falkenstein M, Hoormann J, Christ S, Hohnsbein J (2000) ERP components on reaction errors and their functional significance: a tutorial. Biol Psychol 51(2–3):87–107
17. Ferrez PW, Millán JdR (2005) You are wrong!—automatic detection of interaction errors from brain waves. In: Proceedings of the 19th international joint conference on artificial intelligence, EPFL-CONF-83269
18. Ferrez PW, Millán JR (2008) Error-related eeg potentials generated during simulated brain–computer interaction. IEEE Trans Biomed Eng 55(3):923–929
19. Friedman JH (1989) Regularized discriminant analysis. J Am Stat Assoc 84(405):165–175
20. Frischen A, Bayliss AP, Tipper SP (2007) Gaze cueing of attention: visual attention, social cognition, and individual differences. Psychol Bull 133(4):694
21. Gombolay MC, Gutierrez RA, Clarke SG, Sturla GF, Shah JA (2015) Decision-making authority, team efficiency and human worker satisfaction in mixed human–robot teams. Auton Robots 39(3):293–312
22. Goodrich MA, Schultz AC (2007) Human–robot interaction: a survey. Found Trends Hum Comput Interact 1(3):203–275
23. Gouaillier D, Hugel V, Blazevic P, Kilner C, Monceaux J, Lafourcade P, Marnier B, Serre J, Maisonnier B (2008) The nao humanoid: a combination of performance and affordability. CoRR arXiv:abs/08073223
24. Hairston WD, Whitaker KW, Ries AJ, Vettel JM, Bradford JC, Kerick SE, McDowell K (2014) Usability of four commercially-oriented EEG systems. J Neural Eng 11(4):046,018
25. Holroyd CB, Coles MG (2002) The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. Psychol Rev 109(4):679
26. Holroyd CB, Dien J, Coles MG (1998) Error-related scalp potentials elicited by hand and foot movements: evidence for an output-independent error-processing system in humans. Neurosci Lett 242(2):65–68
27. Homan RW, Herman J, Purdy P (1987) Cerebral location of international 10–20 system electrode placement. Electroencephalogr Clin Neurophysiol 66(4):376–382
28. Huang CM, Mutlu B (2012) Robot behavior toolkit: generating effective social behaviors for robots. In: Proceedings of the seventh annual ACM/IEEE international conference on human–robot interaction. ACM, pp 25–32
29. Huang CM, Cakmak M, Mutlu B (2015) Adaptive coordination strategies for human–robot handovers. In: Robotics: science and systems
30. Iturrate I, Montesano L, Minguez J (2010) Single trial recognition of error-related potentials during observation of robot operation. In: 2010 annual international conference of the IEEE engineering in medicine and biology society (EMBC). IEEE, pp 4181–4184
31. Iturrate I, Montesano L, Minguez J (2013) Task-dependent signal variations in eeg error-related potentials for brain–computer interfaces. J Neural Eng 10(2):026,024
32. Iturrate I, Chavarriaga R, Montesano L, Minguez J, Millán JR (2015a) Teaching brain–machine interfaces as an alternative paradigm to neuroprosthetics control. Sci Rep 5(13):893
33. Iturrate I, Grizou J, Omedes J, Oudeyer PY, Lopes M, Montesano L (2015b) Exploiting task constraints for self-calibrated brain–machine interface control using error-related potentials. PLoS ONE 10(7):e0131,491
34. Joosse M, Sardar A, Lohse M, Evers V (2013) Behave-II: the revised set of measures to assess users? attitudinal and behavioral responses to a social robot. Int J Soc Robot 5(3):379–388
35. Kaan E, Harris A, Gibson E, Holcomb P (2000) The P600 as an index of syntactic integration difficulty. Lang Cogn Process 15(2):159–201
36. Kim SK, Kirchner EA, Stefes A, Kirchner F (2017) Intrinsic interactive reinforcement learning-using error-related potentials for real world human–robot interaction. Sci Rep 7(1):17,562
37. Kreilinger A, Neuper C, Müller-Putz GR (2012) Error potential detection during continuous movement of an artificial arm controlled by brain–computer interface. Med Biol Eng Comput 50(3):223–230

38. Llera A, van Gerven MA, Gómez V, Jensen O, Kappen HJ (2011) On the use of interaction error potentials for adaptive brain computer interfaces. Neural Netw 24(10):1120–1127
39. Lotte F, Guan C (2010) Learning from other subjects helps reducing brain–computer interface calibration time. In: 2010 IEEE international conference on acoustics speech and signal processing (ICASSP). IEEE, pp 614–617
40. Miltner WH, Braun CH, Coles MG (1997) Event-related brain potentials following incorrect feedback in a time-estimation task: evidence for a generic neural system for error detection. J Cogn Neurosci 9(6):788–798
41. Mutlu B, Forlizzi J (2008) Robots in organizations: the role of workflow, social, and environmental factors in human–robot interaction. In: Proceedings of the 3rd ACM/IEEE international conference on human robot interaction. ACM, pp 287–294
42. Mutlu B, Shiwa T, Kanda T, Ishiguro H, Hagita N (2009) Footing in human–robot conversations: how robots might shape participant roles using gaze cues. In: Proceedings of the 4th ACM/IEEE international conference on human robot interaction. ACM, pp 61–68
43. Parra LC, Spence CD, Gerson AD, Sajda P (2003) Response error correction—a demonstration of improved human–machine performance using real-time EEG monitoring. IEEE Trans Neural Syst Rehabil Eng 11(2):173–177
44. Peirce JW (2007) Psychopy—psychophysics software in python. J Neurosci Methods 162(1–2):8–13
45. Ridderinkhof KR, Ullsperger M, Crone EA, Nieuwenhuis S (2004) The role of the medial frontal cortex in cognitive control. Science 306(5695):443–447
46. Salazar-Gomez AF, DelPreto J, Gil S, Guenther FH, Rus D (2017) Correcting robot mistakes in real time using EEG signals. In: 2017 IEEE international conference on robotics and automation (ICRA). IEEE, pp 6570–6577
47. Schäfer J, Strimmer K (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. Stat Appl Genet Mol Biol 4(1):1175–1189
48. van Schie HT, Mars RB, Coles MG, Bekkering H (2004) Modulation of activity in medial frontal and motor cortices during error observation. Nat Neurosci 7(5):549
49. Schlögl A, Keinrath C, Zimmermann D, Scherer R, Leeb R, Pfurtscheller G (2007) A fully automated correction method of EOG artifacts in EEG recordings. Clin Neurophysiol 118(1):98–104
50. Schmidt NM, Blankertz B, Treder MS (2012) Online detection of error-related potentials boosts the performance of mental typewriters. BMC Neurosci 13(1):19
51. Spüler M, Niethammer C (2015) Error-related potentials during continuous feedback: using EEG to detect errors of different type and severity. Front Hum Neurosci 9:155
52. Spüler M, Rosenstiel W, Bogdan M (2012) Online adaptation of a c-VEP brain–computer interface (BCI) based on error-related potentials and unsupervised learning. PLoS ONE 7(12):e51,077
53. Szafir D, Mutlu B (2012) Pay attention!: designing adaptive agents that monitor and improve user engagement. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, pp 11–20
54. Van Herten M, Kolk HH, Chwilla DJ (2005) An ERP study of P600 effects elicited by semantic anomalies. Cogn Brain Res 22(2):241–255
55. Welke D, Behncke J, Hader M, Schirrmeister RT, Schönau A, Eßmann B, Müller O, Burgard W, Ball T (2017) Brain responses during robot-error observation. ArXiv preprint arXiv:170801465
56. Wolpaw JR, Birbaumer N, McFarland DJ, Pfurtscheller G, Vaughan TM (2002) Brain–computer interfaces for communication and control. Clin Neurophysiol 113(6):767–791

Stefan K. Ehrlich is research assistant and doctoral candidate at the Chair for Cognitive Systems, Technical University of Munich (TUM). He received his Master degree in 2012 in electrical engineering and information technologies at TUM and his Bachelor degree in 2006 from Baden-Württemberg Cooperative State University. Stefan Ehrlich is currently working towards his Ph.D. which is centered around EEG-based non-invasive Brain–Computer Interfaces (BCI), with a focus on modeling and decoding human error- and performance-monitoring in human–robot interaction (HRI).

Gordon Cheng holds the Chair of Cognitive Systems at Technical University of Munich (TUM). He is Founder and Director of the Institute for Cognitive Systems in the Department of Electrical and Computer Engineering at TUM. He is also the coordinator of the CoC for Neuro-Engineering - Center of Competence Neuro-Engineering within the department. He is also involved in a number of major European Union Projects. Over the past years Gordon Cheng has been the co-inventor of approximately 20 patents and is the author of approximately 300 technical publications, proceedings, editorials and book chapters.