A computational model of human decision making and learning for assessment of co-adaptation in neuro-adaptive human-robot interaction

Stefan K. Ehrlich[1] and Gordon Cheng[1]

***Abstract*— Studies have demonstrated the potential of using error-related potentials (ErrPs), online decoded from the electroencephalogram (EEG) of a human observer, for robot skill learning and mediation of co-adaptation in collaborative human-robot interaction (HRI). While these studies provided proof-of-concept of this approach as a highly promising avenue in the field of HRI, a systematic understanding of the dyadic interacting system (human and machine) remained unexplored. This research aims to address this gap by proposing a computational model of the human counterpart and simulating the integrated dyadic system. The model can be employed for the systematic study of both human behavioral and technical factors influencing co-adaptation as exemplarily demonstrated in this paper for hypothetical variations of ErrP-decoder performance. The obtained findings have practical implications for future steps along this line of research, for instance to what extent and how improvements of ErrP-decoder performance can benefit co-adaptation in ErrP-based HRI. The proposed computational model enables the prediction of human behavior in the context of ErrP-based HRI. As such it allows the simulation of future empirical studies prior to their conductance and thereby providing a means for accelerating progress along this line of research in a resource-saving manner.**

I. INTRODUCTION

Error-related potentials (ErrPs) are a specific type of event-related potential (ERP) occurring in response to a human subject observing an external instance performing an erroneous or unexpected action and can be robustly decoded from electroencephalography (EEG) signals with relatively high accuracies [1], [2]. Previous studies have demonstrated the usability of ErrPs, online decoded from a human subject's ongoing EEG signals, as a feedback signal for intuitive reinforcement learning of robot skills [3], [4], [5]. Our recent study [6] extended this line of research and provided empirical proof-of-concept for the usability of ErrPs to mediate co-adaptation in human-agent interaction. Here, subjects performed a game-like interactive task with a humanoid robot in which they learned to infer goals from the robot's gaze behavior, while the robot learned to convey these goals by adapting its gaze behavior. Importantly, feedback to the robot for adapting gaze behavior was only provided via online decoded ErrPs (see Fig. 1). Results demonstrated successful co-adaptation in the majority of subjects.

While the above mentioned works [3], [4], [5], [6] provided strong support for the potential of deploying online decoded ErrPs in the domain of HRI and human-machine interaction in general, they were largely focused on technical

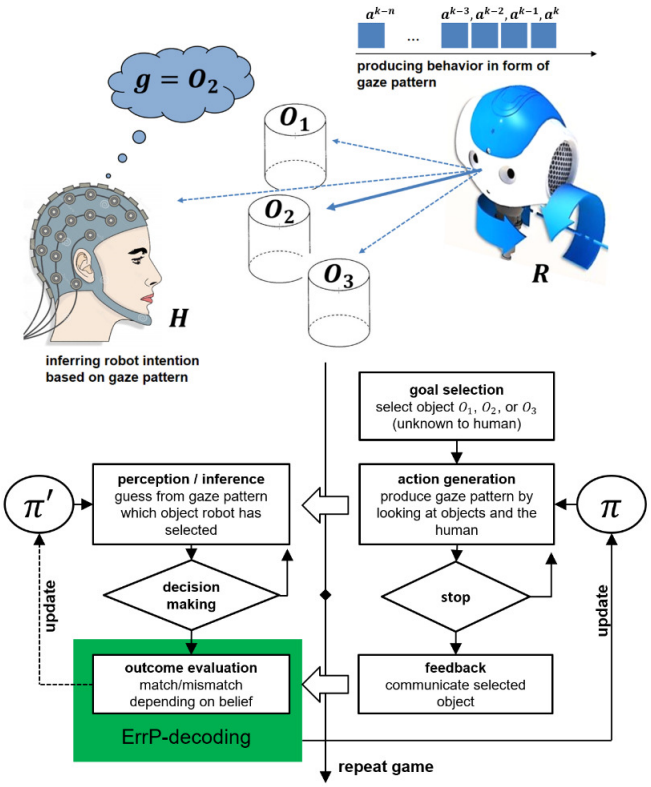

Fig. 1. **Experimental paradigm:** Human subject and robot play a guessing game in which the robot covertly selects one out of three objects. Subsequently the robot generates a gaze pattern based on which the subject has to guess the secret object. The subject's brain responses are measured (marked in green) and used as a feedback signal to adapt the robot's gaze policy π, while the subject may likewise adapt expectations π' about the robot's gaze behavior. Adapted from Ehrlich & Cheng, 2018 [6].

aspects, and limited in providing a systematic understanding of the dyadic interacting system (human and machine). After all, human participants are involved, which vary in their way of perceiving, interpreting, and coping with the task and robot stimulus. This may be a reason why in our study co-adaptation was successfully achieved in some participants and less so in others [6]. For instance, it remained unclear to what extent variations of human factors (e.g. learning/adaptation style) contributed to the joint performance of the human-agent system. Also, the influence of technical factors such as online ErrP decoding performance or the type of agent learning paradigm remained largely unexplained. Instead of conducting additional empirical studies, these questions were here addressed by means of a computational model simulating the integrated dyadic co-adaptive system

[1]Chair for Cognitive Systems, Department of Electrical and Computer Engineering, Technische Universität München (TUM), Munich, Germany stefan.ehrlich@tum.de

(see Fig. 2). The model presented focuses on our recent empirical HRI study [6]. While research on modeling human (and animal) decision making has a longstanding history in neuroscience, cognitive psychology, and game theory, it appears rather scarce in the domain of brain-computer interfaces (BCI), with a few works over the last decades, such as [7]. Also in the domain of HRI, modelling human behavior appears to be not yet common sense, although some recent research works demonstrated the potential of improving HRI by including models emulating human behavior (see [8] and [9] for an overview). At the intersection between HRI and BCI, e.g. in the relatively new field of ErrP-mediated collaborative HRI, computational modeling of human behavior appears to be yet entirely unexplored. This is why this work mainly integrates concepts from previous works in the domain of cognitive neuroscience. Here, a series of computational models have been proposed over the last decades describing human (and animal) decision making and corresponding neural activities in the context of effortful and conflicting tasks, including social interaction (see [10] for an overview). The underlying process is usually referred to as *'performance monitoring'*; the computational center in the brain is understood to be the anterior cingulate cortex (ACC) and projecting areas. The majority of the proposed models employed reinforcement learning based computational frameworks, in particular temporal-difference learning in actor-critic architectures (see for instance [11], [12], [13]). A direct application of the so far proposed models was challenged by the fact that they were largely instantiated to account for data based on experimental tasks and stimuli typical in the domain of neuroscience (e.g. choice-reaction-time tasks, Stroop task, or gambling tasks). In contrast, our problem encompasses a relatively complex stimulus (robot) in a much less constrained task setting. This led us to decide to develop a model tailored to our problem by taking inspiration mainly from the *Predicted-Response-Outcome (PRO)* model [13], but also others which are acknowledged in respective sections. An important prerequisite was to allow flexible extensions of our model to different HRI scenarios in future work.

This research is divided into four steps: (i) proposition of a computational model of human decision making and learning in the context of ErrP-based human-agent co-adaptation, (ii) fitting the model based on empirical data from our previous study [6], (iii) embedding the model into an interaction environment with an agent, and (iv) performing model-based simulations of the integrated model to elucidate potential factors influencing co-adaptation. With this, the present work contributes to research on ErrP-based HRI in two ways:

- Generally, the proposed model provides a platform for the systematic study of the influence of both human behavioral and technical factors on ErrP-based co-adaptation in HRI. Furthermore, it allows the derivation of testable hypotheses and informed experiment design of future studies.
- Specifically, this paper reports results from simulations of varying ErrP-decoder performance. The obtained findings allow an estimate of the impact of hypothetical improvements of ErrP-decoding accuracy on human-agent co-adaptation performance which has practical implications on future steps along this line of research.

II. COMPUTATIONAL MODEL

A. Experimental paradigm and empirical data

This section briefly recapitulates the experimental setup and technical implementation of our previous study; for more detailed information, the reader is referred to [6]. In the experiment, subjects had to perform a collaborative task together with a real humanoid robot[1] (see Fig. 1), which required co-adaptation from both sides (human and robot). The task was designed as a repeated guessing game in which subjects were asked to infer from the robot's gaze behavior which one out of three available objects it has selected (underlying goal/intention). A single guessing game (further referred to as *trial*) started with the robot secretly deciding for one of the three objects and then proceed with alternatingly gazing at either of three objects or the subject in a fixed pace (one gaze transition in 400 ms). Meanwhile the subject would attempt to infer the correct object from observing the robot's ongoing gaze behavior and eventually take a decision for either of the three target objects. This means that subjects could continue observing the behaving robot until they were certain about their decision. Subjects were asked to indicate their decision in a self-passed fashion via keypress responses (these were only used as ground truth for validation). Afterwards, the robot would communicate the true choice of object to the subject (feedback); time-locked to the moment of feedback presentation, the subject's EEG-based ERP would be classified online into non-error (match) or error (mismatch) and then administered to the robot for adapting its current gaze policy. This way, the robot's gaze behavior would gradually adapt to subjects' expectations; likewise subjects may adapt their expectations, e.g. learn a gaze pattern 'suggested' by the robot. A measure for successful co-adaptation was *guessing performance*, i.e. the subject's accuracy in correctly inferring the robot's chosen object from its gaze behavior. Sixteen healthy subjects took part in the experiment (age: 29.2$\pm$5.0, 7 females, 9 males). Each one performed first a calibration session (CALIB) followed by four co-adaptation sessions (CORL-I-IV). The calibration session consisted of 150 trials (guessing games) in which the robot gaze behavior followed a fixed (e.g. non-adaptive) policy. Data collected during this session was used to derive subject-specific ErrP decoders. During the following four closed-loop co-adaptation sessions, each consisting of 50 trials, the previously built ErrP-decoder was employed for online adaptation of the robot's gaze policy. In each co-adaptation session, the gaze policy was (re-)initialized such that the robot would generate random gaze behavior in the first trial (starting without prior knowledge).

[1]The robotic platform chosen for the experiment was the humanoid robot NAO, which is a commercially available (SoftBank Robotics) 58 cm tall robot with 21-25 degrees of freedom.

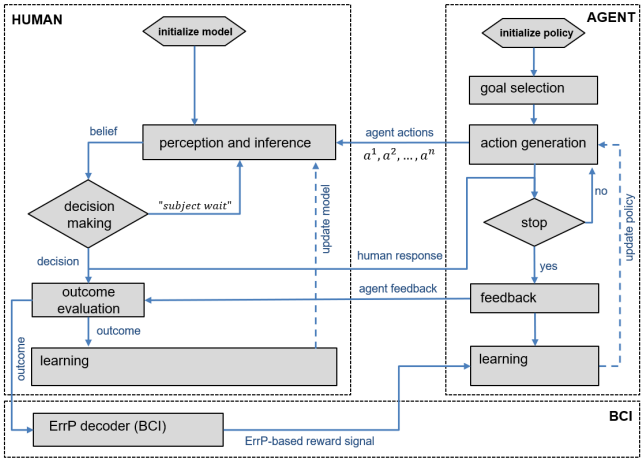

Fig. 2. **Model architecture:** Overview of proposed architecture of the human computational model embedded in the interaction process with the agent. Agent adaptation is performed based on feedback from the evaluation module of the human computational model. Decoding and transmission of that feedback signal derived from online decoded EEG-based error-related potentials is simulated in the BCI block.

B. Proposed model of human decision making and learning

Overview: The proposed model consists of four modules (see Fig. 2): (1) The *perception and inference* module transfers observations of the agent's actions into a belief about the agent's goal. (2) The *decision making* module turns the belief into an explicit action stating the predicted agent's goal. As long as no decision was made, the perception and inference module continues observing the agent and further updates the belief. (3) The *outcome evaluation* module receives the agent feedback and compares it with the predicted agent's goal, based on which it derives a trial outcome and a measure of expectation violation (further referred as 'prediction error'). (4) The *learning* module updates the current knowledge in the model based on the prediction error and the history of observed agent actions.

Perception and inference: Pisauro et al. (2017) recently provided empirical evidence from a combined EEG-fMRI study that humans accumulate evidence in favour of the different alternatives before committing to a decision [14]. To model this evidence accumulation process, we used Bayesian inference based on the formulation of Behrens et al. in 2007 for describing human behavioral data in a perceptual decision making task [15]. Bayesian inference prescribes a standard computation resulting in a posterior belief $p(g_m|a^k_{i,j})$ that alternative g_m is true given observation $a^k_{i,j}$. Here, g_m is the alternative among the set of agent goals $g = \{g_{O1}, g_{O2}, g_{O3}\}$, with M being the number of possible goals (or target objects: $M = 3$). $a^k_{i,j}$ is the observed agent action from gaze state s_i to s_j in time step k. The agent can be in four possible gaze states; either gazing at the human or any of the three target objects $S = \{s_{O1}, s_{O2}, s_{O3}, s_H\}$. Transitions are possible from one gaze state to another or staying in the current state. As observations arrive sequentially over time, the posterior belief $p(g_m|a^{1:k}_{i,j})$ can be updated recursively using all observations $a^{1:k}_{i,j} = \{a^1_{i,j}, ..., a^k_{i,j}\}$ up to time step k. Equation (1) describes the computation of the posterior belief after observation of the first agent action in the current trial t. Before observing any agent action, the belief about the agent's goal is uniform among alternatives, hence the prior is initialized with $p(g_m) = \frac{1}{3}$. The posterior belief is used as the new prior upon observation of the next robot action according to equation (2). The likelihoods $p(a_{i,j}|g_m)$ are computed based on internal weights $w^{t-1}_{i,j,m}$ of the previous trial $t-1$ using the Softmax function according to equation (3). These weights associate the set of observable actions $a_{i,j}$ with the goal alternatives g_m and as such describe the current knowledge of the model. Upon start of the experiment, the weights are initialized uniformly among all possible actions and goals with $w^0_{i,j,m} = 0$ for all i, j, m.

$$p(g_m|a^1_{i,j}) = \frac{p(a^1_{i,j}|g_m)p(g_m)}{\sum_{m=1}^{M} p(a^1_{i,j}|g_m)p(g_m)} \tag{1}$$

$$p(g_m|a^{1:k}_{i,j}) = \frac{p(a^k_{i,j}|g_m)p(g_m|a^{1:k-1}_{i,j})}{\sum_{m=1}^{M} p(a^k_{i,j}|g_m)p(g_m|a^{1:k-1}_{i,j})} \tag{2}$$

$$p(a_{i,j}|g_m) = \frac{e^{w^{t-1}_{i,j,m}}}{\sum_{m=1}^{M} e^{w^{t-1}_{i,j,m}}} \tag{3}$$

Decision making: The decision making module turns the current belief $p(g_m|a^k_{i,j})$ into an explicit action stating the predicted agent's goal $\hat{g}$. Here, we draw from the formulation of the PRO-model by Alexander & Brown in 2011 in which decisions are initiated when the belief exceeds a certain threshold (*decision bound*) [13]. However, as mentioned initially, the PRO-model was largely instantiated based on choice reaction time tasks, in which subjects were to respond as quickly as possible. In our case, subjects could freely decide when to respond; decision time was neither rewarded nor penalized. This means that on the one hand subjects could continue observing the agent although they were already certain about its goal in the hope of collecting more evidence supporting the current belief. On the other hand, subjects could respond earlier even though they were not certain about the agent's goal, possibly because they lost confidence that future observations would help consolidate the current belief. Based on this, it is assumed that subject's decisions did not only depend on their belief, but also on the elapsed time spend on observing the agent. Therefore, the decision variable $\beta(g_m)$ was realized as a function of the posterior belief and k. Both variables were linearly combined according to equation (4) and the influence of k was scaled with a factor ϵ (further referred to as *timeout* parameter). A decision $\hat{g} = \text{argmax}_i\, \beta(g_m)$ is initiated whenever the maximum of $\beta(g_m)$ is equal or greater than the decision bound Γ otherwise, the next observation is awaited according to equation (5). The decision bound Γ is recomputed in every trial t based on a Gaussian random process: $\Gamma \sim \mathcal{N}(\mu_\beta, \sigma^2_\beta)$.

$$\beta(g_m) = p(g_m|a^{1:k}_{i,j}) + \epsilon k \tag{4}$$

$$\hat{g} = \begin{cases} \text{argmax}_m\, \beta(g_m) & \text{if } \max_m \beta(g_m) \geq \Gamma \\ 0 \text{ "wait"} & \text{otherwise} \end{cases} \quad (5)$$

Outcome evaluation: The evaluation module validates the outcome of the trial with regard to the decision about the predicted agent's goal $\hat{g}$ and the feedback about the true agent's goal g. First, a binary outcome O is computed according to equation (6). Second, the prediction error δ is computed based on the expected outcome and the true outcome. The expected outcome is the posterior belief at the moment of which the decision was taken.

$$O = \begin{cases} 1 & \text{if } \hat{g} = g \text{ (non-error)} \\ 0 & \text{if } \hat{g} \neq g \text{ (error)} \end{cases} \quad (6)$$

$$\delta = O - p(\hat{g}|a_{i,j}^{1:k}) = \begin{cases} 0 \leq \delta \leq +1 & \text{if } \hat{g} = g \\ -1 \leq \delta \leq 0 & \text{if } \hat{g} \neq g \end{cases} \quad (7)$$

Learning: The learning module is responsible for updating the internal weights after the outcome of the trial has been evaluated. The learning function used in this model was based on previous works describing behavioral and neurophysiological responses in choice reaction time and gambling tasks as a reinforcement learning process [13], [10]; particularly based on the formulation of Cohen et al. in 2007 [16]. In equation (8) the weights of the previous trial are updated for the next trial by adding the probabilities of observed actions, weighted by the prediction error δ and scaled by the learning rate λ. As subjects may react differently to erroneous versus correct guesses, we introduced separate learning rates for success and failure outcomes ($\lambda^{(+)}$ and $\lambda^{(-)}$, respectively). The gating parameter $\eta(\hat{g})$ is 1 for the chosen target and 0 for all non-chosen targets according to equation (9). This way, the weights associated with the chosen target are either positively or negatively reinforced. After updating, the weights are finally turned into new likelihoods according to equation (3) and used by the perception and inference module in the next trial.

$$w_{i,j,m}^{t} = \begin{cases} w_{i,j,m}^{t-1} + \eta(\hat{g})\delta\lambda^{(+)}\frac{1}{n}\sum_{k=1}^{n} a_{i,j}^{1:k} & \text{if } O = 1 \\ w_{i,j,m}^{t-1} + \eta(\hat{g})\delta\lambda^{(-)}\frac{1}{n}\sum_{k=1}^{n} a_{i,j}^{1:k} & \text{if } O = 0 \end{cases} \quad (8)$$

$$\eta(g_m) = \begin{cases} 1 & \text{for } g_m = \hat{g} \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

C. Parameter fitting

The human computational model consists of 7 parameters ($\lambda^{(+)}$, $\lambda^{(-)}$, μ_β, σ_β^2, ϵ, TNR, TPR). ErrP-decoder true-negative and true-positive rates (TNR, TPR) were determined based on the online decoding performance during the experimental study. Non-error (match) events were defined as the negative class, whereas error (mismatch) events were defined as the positive class. The remaining four model parameters ($\lambda^{(+)}$, $\lambda^{(-)}$, μ_β, σ_β^2) were fit to each subject individually. The original experiment data consisted of 16 subjects out of which the data of one subject (s06) had to be removed due to incomplete marker information. Model fitting was performed in a sequential two step procedure: (1) Success and failure learning rates were first fitted using an optimization procedure, and (2) the decision bound parameters were retrieved by fitting a Gaussian process to the distribution of $\beta(\hat{g})$ across trials. Finally, the timeout parameter ϵ turned out to only affect how well the model describes the experimental data in terms of decision times (number of observed agent actions until subject decision). Since decision times were not in the focus of our investigation, the timeout parameter was empirically set to $\epsilon = 0.02$ during model validation (see Section II.D). This resulted in a good match of simulated and real decision times for all subjects.

Learning rates: The computational model was presented with the actions performed by the robot in the actual experimental sessions, starting with the calibration session and continuing with the co-adaptation runs CORL-I, CORL-II, and CORL-IV[2]. The model started with zero knowledge and continued learning across the entire experiment. The model's simulated decisions were compared to the subject's real decisions on a single-trial level. The cost function to be minimized is depicted in equation (10) and computes the root-mean-squared error between the simulated $O_{SIM}^{1:t}$ and the actual subject's trial outcomes $O_{DATA}^{1:t}$ across all trials T (see equation (6)). Outcomes of individual trials were first smoothed with a 10-trial kernel before being administered to equation (10), a procedure often used to examine correspondence between model predictions and behavioral selections, such as in [16]. The cost function was computed separately for the calibration (150 trials) and the three co-adaptation sessions (each 50 trials) and subsequently averaged to obtain a single measure of goodness of fit F. Optimization of success and failure learning rates were performed by minimizing F in a sequential two-step procedure: (1) by localization of the global minimum using a 2D grid search across the values $\lambda^{(+)} = \{0, 0.2, ..., 3\}$ and $\lambda^{(-)} = \{0, 0.2, ..., 1\}$, and (2) by fine-grained optimization using the nonlinear, unconstrained Nelder-Mead simplex method starting with initial values resulting from step 1.

$$F = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(\overline{O}_{SIM}^{1:t} - \overline{O}_{DATA}^{1:t})^2} \quad (10)$$

Decision bound: Visual inspection of the discrete probability distributions of the decision bound across trials revealed normal distributions in all subjects. Therefore, the variations of the decision bound were assumed to result from a Gaussian random process reflecting the subject's uncertainty about the underlying belief. Based on this, the decision bound of each subject was modelled individually via a Gaussian random process (described by μ_β and σ_β^2) based on trials

[2]In CORL-III a different protocol was employed which did not capture ground truth information about subject decisions; therefore it was excluded from this research.

in which the model's decision matched the decision of the subject ($\hat{g}_{SIM} = \hat{g}_{DATA}$).

Fitting results: Exemplary results of the fitting procedure for two subjects can be visually inspected in Fig. 3; numerical results for all subjects are provided in Table I. Overall, simulations match well the experimental data with an average across subjects goodness of fit of $F = 0.19 \pm 0.04$ (AVG±SD). This translates to an average percentage of correct predictions of subject decisions of $P = 70.4 \pm 11.1\%$ (AVG±SD), with 12 out of 15 subjects resulting in $P \geq 65\%$. These results suggest that the model captured relevant behavioral effects observed in the human data.

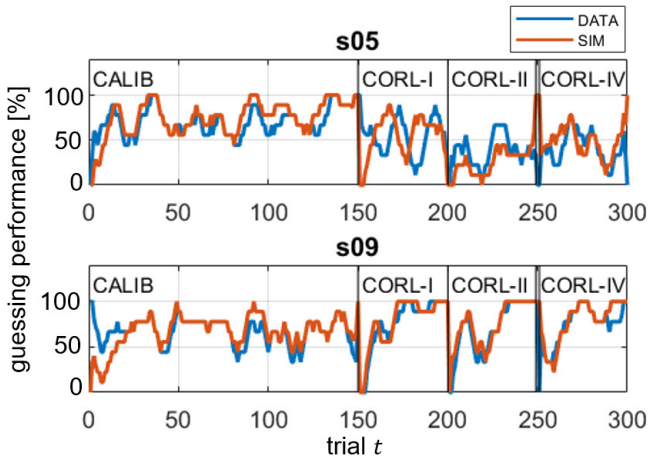

Fig. 3. **Qualitative results of goodness of fit:** Exemplary results of two subjects depicting simulated (orange) and real (blue) learning curves (guessing performance smoothed with a 10-trials kernel for CALIB, CORL-I, CORL-II, and CORL-IV separately).

D. Integration of the model into the co-learning system

After fitting the free parameters of the computational model based on experimental data, the model is ready to be integrated into the co-learning system including the adaptive robotic agent, and the brain-computer interface (ErrP-decoder) providing implicit human feedback to the agent (see Fig. 2). Algorithm 1 describes the perception and inference, decision making, outcome evaluation, and learning procedure of the human subject. Algorithm 2 describes the action generation and learning procedure of the robotic agent. The simulation of the ErrP-decoder interfacing human and agent is described in equation (11). The function computes a reward estimate $\tilde{R}$ for agent policy updating (see algorithm 2) based on the true outcome of the trial and a comparison of a uniform random number $\mathcal{U}(0, 100)$ with the subject's individual ErrP-decoder rates. As such, the function emulates the derivation of a reward from an imprecise ErrP-decoder decision scaled with the experimental parameters TNR and TPR.

$$\tilde{R} = \begin{cases} +1 & \text{if } (O = 1)\&(\mathcal{U}(0,100) \leq TNR); \text{else } -1 \\ -1 & \text{if } (O = 0)\&(\mathcal{U}(0,100) \leq TPR); \text{else } +1 \end{cases} \tag{11}$$

TABLE I

MODEL PARAMETERS, SUCCESS / FAILURE LEARNING RATES ($\lambda^{(+)}$, $\lambda^{(-)}$), DECISION BOUND (μ_β, σ_β^2), AND EXPERIMENTALLY DETERMINED ONLINE ERRP-DECODER PERFORMANCE (TNR, TPR [%])[3] AND RESULTING GOODNESS OF FIT F, AND PERCENTAGE OF CORRECT PREDICTIONS OF SUBJECT DECISIONS P [%].

ID	$\lambda^{(+)}$	$\lambda^{(-)}$	μ_β	σ_β^2	TNR	TPR	F	P
s03	1.5	0.3	1.3	0.2	92.1	69.0	0.11	87.0
s04	0.0	0.2	0.5	0.1	96.8	86.4	0.28	44.0
s05	2.4	0.8	1.0	0.2	62.8	42.2	0.24	66.0
s07	2.2	1.0	0.9	0.2	83.2	81.9	0.19	56.3
s08	2.6	0.8	1.1	0.2	95.5	71.1	0.21	77.3
s09	2.6	0.4	1.2	0.2	90.1	77.0	0.13	84.7
s10	0.2	0.2	0.7	0.1	100.0	1.0	0.16	72.0
s11	2.4	0.4	1.1	0.2	76.5	56.6	0.21	60.7
s12	2.8	0.3	1.3	0.3	95.0	71.6	0.22	74.7
s13	2.6	0.4	1.2	0.2	71.7	63.9	0.21	69.3
s14	0.7	0.3	1.0	0.2	90.6	66.0	0.19	73.7
s15	1.0	0.4	1.0	0.2	90.5	68.2	0.14	82.0
s16	2.0	0.2	1.2	0.3	58.4	81.5	0.15	73.0
s17	2.6	0.4	1.1	0.2	84.3	89.4	0.18	68.7
s18	1.4	0.8	1.0	0.2	89.0	65.2	0.19	67.0
AVG	**1.80**	**0.46**	**1.04**	**0.20**	**85.1**	**66.1**	**0.19**	**70.4**
±SD	0.94	0.26	0.22	0.05	12.5	21.6	0.04	11.1

[3]In the experimental study [6], ErrP-decoder calibration turned out below chance level in three subjects (s05, s10, s13). This resulted in TNR and TPR of subject s10 being highly biased during online ErrP decoding.

Algorithm 1: HUMAN perception and inference, decision making, outcome evaluation, and learning algorithm

Initialize model weights $w^0_{i,j,m} = 0$ for all i, j, m
for $t \leftarrow 1$ **to** T **do**
  Initialize prior belief: $p(g_m) = \frac{1}{M}$
  Initialize action count: $k = 0$
  Initialize action history: $a_{i,j} = 0$ for all i, j
  Compute likelihoods: $w^{t-1}_{i,j,m}$ [equation (3)]
  Compute decision bound: $\Gamma \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$
  while $\beta(g_m) < \Gamma$ *("subject wait")* **do**
    $k \leftarrow 1$
    Observe agent action a^k
    Update posterior belief based on observed action a^k [equation (2)]
    Update decision variable $\beta(g_m)$ [equation (4)]
    Update prior belief: $p(g_m|a^{1:k}_{i,j}) \leftarrow p(g_m|a^{1:k-1}_{i,j})$ [equations (1) and (2)]
    Update observed action history: $a^k_{i,j} \leftarrow 1$
  end
  Take decision $\hat{g}$ [equation (5)]
  Observe agent feedback g
  Evaluate outcome O [equation (6)]
  Compute prediction error δ [equation (7)]
  Update weights $w^t_{i,j,m} \leftarrow w^{t-1}_{i,j,m}$ [equation (8)]
end

Algorithm 2: AGENT action generation and learning algorithm

Initialize policy: $\pi = \frac{1}{4}$ for all s, a
for $t \leftarrow 1$ **to** T **do**
  Initialize action count: $k = 0$
  Choose initial gaze state s_i with $s = \mathcal{U}\{s_{O1}, s_{O2}, s_{O3}, s_H\}$
  Choose goal g with $g = \mathcal{U}\{g_{O1}, g_{O2}, g_{O3}\}$
  Initialize gaze transition history: $a_{i,j} = 0$ for all i, j
  while $\hat{g} = 0$ *("subject wait")* **do**
    $k \leftarrow 1$
    Choose action a_j from state s_i using $\pi(g)$
    Execute action a_j (perform gaze shift), observe next gaze state s_j
    Update gaze transition history: $a^k_{i,j} \leftarrow 1$
    $s_i \leftarrow s_j$
  end
  Present feedback g to subject
  Observe reward $\tilde{R}$ from ErrP-decoder
  Update policy: $\pi \leftarrow \pi + \tilde{R}\alpha A_{i,j}$
  Truncate policy: $\pi \leftarrow clamp(\pi)^1_0$ for all s, a
  Normalize policy: $\pi \leftarrow \frac{\pi}{\sum_s \pi}$
end

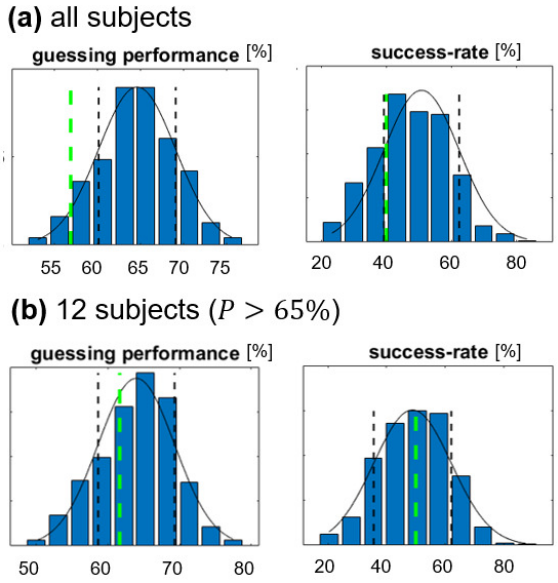

Fig. 4. **Validation of the integrated human-agent model:** (a) Distributions of performance measures (guessing performance and success-rate) across 500 simulated studies including all subjects and (b) a subset of subjects whose decisions the model could predict with $P > 65\%$. Results show that empirical measures of the original study (green dashed lines) are well within the single standard deviation of the simulated data (black dashed lines) for the two performance measures when considering the subset of subjects compared to considering all subjects.

E. Validation of the integrated human-agent model

The fitting results (see Section II.C) demonstrated that the model is capable of sufficiently capturing behavioural effects observed in the human data. The following section assesses the validity of the computational model in the context of large-scale simulations in comparison with the empirical data observed in the experimental study [6].

To establish equal conditions for the comparison, the original experimental protocol was simulated with a large number of individual experiments. The simulation was focused on the calibration-session (CALIB) and subsequent first co-adaptation run (CORL-I) only, as these made up the most critical parts of the original experiment. The comparison between simulated and real experimental data was performed on the *study-level*, e.g. a single simulated study comprised all 15 subjects and reported measures of co-adaptation performance averaged across subjects. Four measures of co-adaptation performance were compared:

- *guessing performance*: mean guessing performance in 50 trials of CORL averaged across subjects.
- *success-rate*: ratio of successful[4] co-adaptation runs across subjects CORLs across subjects.

In total 500 studies were simulated; distributions of the above measures were then compared to the empirical measures from the original study. The underlying proposition of this comparison is that the model reflects well the experimental data (and can be trusted at the level of large-scale simulations) if the empirical measures lie within the single standard deviation of the simulated measures.

Results are depicted in Fig. 4(a) and show the empirical performance measures (dashed green line) superimposed on the distributions of simulated measures. Sufficient consistency between simulated and real data can be observed for success-rate, but average guessing performance turned out to be consistently higher than the empirical average guessing performance (see Fig. 4(a)). This discrepancy could be due to the fact that the model could not be fitted equally well to all subjects. Fitting results turned out rather low in a few subjects (see Table I). A comparison was therefore performed based only on subjects for which the model could be well fitted. As a threshold we chose $P < 65\%$ which excluded three out of fifteen subjects (s04, s07, s11). Results, depicted in Fig. 4(b) show that now empirical measures (green dashed lines) are well within the single standard deviation of the simulated data (black dashed lines) for all four measures. This consistency between the empirical and simulated data corroborates the above hypothesis that the observed discrepancy was essentially due to the cases for which no good fit could be found. However, they also show the limitations of the model, since the latter can only account for a subset (approx. 80%) of the participants. Nonetheless, the consistency observed validates using the model for large scale simulations, as long as these are based and constrained to this subset of subjects. All further investigations and simulations therefore excluded subjects s04, s07, and s11 from the analyses.

[4]A co-adaptation run consisting of 50 trials was denoted successful if guessing performance was $\geq 70\%$ in three subsequent segments of 10 trials, $p < 7.6 * 10^{-6}$, one-sided binomial test with $p_{chance} = 1/3$.

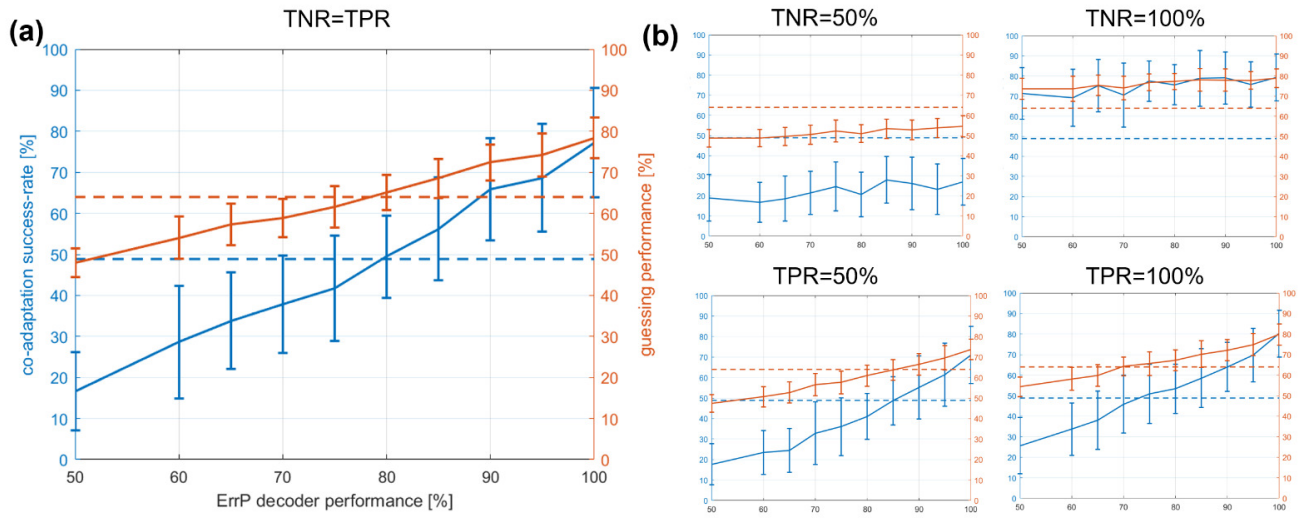

Fig. 5. **Simulated impact of ErrP-decoder performance:** average guessing performance (orange lines) and success-rate (blue lines) for fixed ErrP-decoder performance across subjects with (a) balanced non-error and error decoding performance, (b) fixed TNR and varying TPR (upper panels), and vice-versa (lower panels). Error bars represent the single standard deviation of the mean across 50 studies. Results indicate an approx. linear relationship between ErrP-decoder performance and co-adaptation performance measures. Variations of co-adaptation performance seem to be mainly driven by TNR; variations of TPR have a negligible effect on co-adaptation performance. Average baseline measures (computed based on individual subject's TNR and TPR) are depicted as dashed lines.

III. EFFECT OF VARYING ERRP-DECODER PERFORMANCE ON CO-ADAPTATION

Higher ErrP-decoder performances are naturally assumed to yield better co-adaptation performance as the amount of false feedback provided to the robotic agent is reduced. What remains unclear is how much any hypothetical improvement of ErrP-decoder performance would impact the performance of the entire system (human and agent). The simulations presented in this subsection were conducted to assess the extent to which it is worth to try to improve ErrP-decoder performance.

This investigation was performed by running large sets of simulated experiments on single subject level, e.g. in each experiment the model parameters ($\lambda^{(+)}$, $\lambda^{(-)}$, μ_β, σ^2_β, TNR, TPR) were fixed to one specific subject (in accordance to the approach used to validate the integrated human-agent model, see Section II.E). To achieve robust statistics, all simulations comprised 50 experiments per subject and condition, e.g. 50 simulated studies. This resulted in 600 experiments (12 subjects x 50) per condition, each starting with 150 trials of simulated CALIB and a single subsequent CORL of 50 trials. Conditions were defined by a hypothetical set of TNR and TPR which were fixed across subjects (while no other model parameters were modified). Condition-wise results were compared to the *baseline condition* in which the empirical subject-specific measures of TNR and TPR were used (see Table I). Measures for quantifying co-adaptation performance were *guessing performance* and *success-rate* in accordance with the definitions in Section II.E.

Results are depicted in Fig. 5(a) and indicate an approx. linear relationship between ErrP-decoder performance and co-adaptation performance measures. The results show, that even a non-functioning ErrP-decoder ($TNR = TPR = 50\%$) results in an avg. guessing performance of $\sim 48\%$ and an average success-rate of $\sim 16\%$. Optimal ErrP-decoder performance ($TNR = TPR = 100\%$) on the other hand results in an avg. guessing performance of $\sim 78\%$ and a success-rate of $\sim 77\%$. According to the simulation, an overall improvement of i.e. 10% in ErrP-decoding performance results in $\sim 5\%$ improvement of guessing performance and $\sim 10\%$ improvement of success-rate. As hypothesized, the results reflect that ErrP-decoder performance plays a critical role, but also suggest that it is not the only factor affecting co-adaptation. Some level of co-adaptation can even be achieved with a non-functioning ErrP-decoder (possibly compensated by the human counterpart) and, more importantly, optimal ErrP-decoding does not straightforwardly result in optimal co-adaptation performance. Further results (see Fig. 5(b)) indicate that variations of co-adaptation performance are mainly driven by variations of TNR (non-error decoding rate). Variations of TPR (error decoding rate) on the other hand have a negligible effect on co-adaptation performance. In this regard, misclassifications of error trials seem to have a weaker effect on the overall co-learning system (human and agent) than misclassifications of non-error trials. In summary this implies that efforts to improve ErrP-decoder performance, in particular non-error decoding rates, are generally worthwhile but supposedly not the main technical factor for improving co-adaptation performance in the given scenario.

IV. DISCUSSION AND FUTURE WORK

The model-based approach presented in this work allows the systematic investigation of the integrated human-agent system in complement to systems engineering efforts and future empirical studies and as such provides an avenue towards the design of co-adaptive systems (agent and BCI)

that promote optimal performance in interaction with human subjects, while aligning to subjects' individual preferences and expectations.

The framework to which the model currently applies is constrained by the following prerequisites and may need adaptation and refitting of parameters when being applied to different interaction scenarios: (1) The agent's behavior is described by sequences of discrete actions (2). These sequences encode the agent's underlying goals/intentions. (3) These goals are unambiguously disclosed to the human interaction partner at some point, e.g. explicitly through feedback or implicitly through a confirmative action concluding the action sequence. (4) The human interaction partner is intrinsically motivated to infer the agent's goals to adapt future behavior, e.g. to optimize joint performance. Extensions of the model towards a wider range of HRI scenarios are planned for future works.

Within the constraints of the experimental paradigm, the model proved useful to investigate the effect of hypothetical variation of ErrP-decoder performance. Results suggested that in the given interaction scenario, non-error decoding rates play the more crucial role than error decoding rates. This rather surprising observation could be related to subjects' learning style: subjects seem to have put more emphasis on success than failure trials in the process of learning the agent's behavior (suggested by higher success- than failure learning rates, see Table I). In practice this suggests that co-adaptation performance can be improved by biasing ErrP-decoders towards emphasizing TNR at the cost reducing TPR. A similar observation was reported by Llera et al. in 2011 [17], although in a complementary perspective. In their work, they explored the usability of ErrPs for online adaptation of a brain-computer interface (BCI) classifier for a binary choice task. Both simulations and empirical data indicated a more negative effect on the adaptation process resulting from incorrect decoding of ErrPs in response to the BCI output matching subjects intentions and practically no influence resulting from incorrect decoding of ErrPs in response to the BCI output mismatching subjects intentions. This supports both our results and post-hoc those reported by Llera et al. in 2011. Whether and to what extent this observation can be generalized across further interaction scenarios is subject to future research.

V. CONCLUSIONS

This work proposed a computational model for human decision making in the context of ErrP-based mediation of human-agent co-adaptation. The model can be employed for the simulation-based study of both human behavioral and technical factors influencing co-adaptation. This was here exemplarily demonstrated for hypothetical variations of ErrP-decoder performance. Our findings suggested a more critical role of non-error decoding rate than error-decoding rate in the co-adaptation process; an observation which was found consistent with previous research by others. This means that, in practice, interaction performance will benefit in particular from improving the non-error decoder. The proposed computational model enables the prediction of human decision making and learning in the context of ErrP-based neuro-adaptive HRI. As such it allows for simulation of future empirical studies, and thereby provides a means for accelerating progress along this line of research in a resource-saving manner.

ACKNOWLEDGMENT

We thank Ana Alves-Pinto and Pablo Lanillos for helpful feedback and comments on this work.

REFERENCES

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

[2] Stefan K Ehrlich and Gordon Cheng. A feasibility study for validating robot actions using eeg-based error-related potentials. *International Journal of Social Robotics*, pages 1–13, 2018.

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

[4] Andres F Salazar-Gomez, Joseph DelPreto, Stephanie Gil, Frank H Guenther, and Daniela Rus. Correcting robot mistakes in real time using eeg signals. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6570–6577. IEEE, 2017.

[5] Su Kyoung Kim, Elsa Andrea Kirchner, Arne Stefes, and Frank Kirchner. Intrinsic interactive reinforcement learning–using error-related potentials for real world human-robot interaction. *Scientific reports*, 7(1):17562, 2017.

[6] Stefan K Ehrlich and Gordon Cheng. Human-agent co-adaptation using error-related potentials. *Journal of Neural Engineering*, 15(6):066014, 2018.

[7] George Townsend, Bernhard Graimann, and Gert Pfurtscheller. Continuous eeg classification during motor imagery-simulation of an asynchronous bci. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 12(2):258–265, 2004.

[8] Stefanos Nikolaidis, David Hsu, and Siddhartha Srinivasa. Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, 36(5-7):618–634, 2017.

[9] Thomas Cederborg. Artificial learners adopting normative conventions from human teachers. *Paladyn, Journal of Behavioral Robotics*, 8(1):70–99, 2017.

[10] Eliana Vassena, Clay B Holroyd, and William H Alexander. Computational models of anterior cingulate cortex: At the crossroads between prediction and effort. *Frontiers in Neuroscience*, 11:316, 2017.

[11] Matthew M Botvinick, Todd S Braver, Deanna M Barch, Cameron S Carter, and Jonathan D Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

[12] Clay B Holroyd and Michael GH Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, 109(4):679, 2002.

[13] William H Alexander and Joshua W Brown. Medial prefrontal cortex as an action-outcome predictor. *Nature Neuroscience*, 14(10):1338, 2011.

[14] M Andrea Pisauro, Elsa Fouragnan, Chris Retzler, and Marios G Philiastides. Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous eeg-fmri. *Nature communications*, 8:15808, 2017.

[15] Timothy EJ Behrens, Mark W Woolrich, Mark E Walton, and Matthew FS Rushworth. Learning the value of information in an uncertain world. *Nature Neuroscience*, 10(9):1214, 2007.

[16] Michael X Cohen and Charan Ranganath. Reinforcement learning signals predict future decisions. *Journal of Neuroscience*, 27(2):371–378, 2007.

[17] Alberto Llera, Marcel AJ van Gerven, Vicenç Gómez, Ole Jensen, and Hilbert J Kappen. On the use of interaction error potentials for adaptive brain computer interfaces. *Neural Networks*, 24(10):1120–1127, 2011.