Accelerating Reinforcement Learning using EEG-based Implicit Human Feedback

Duo Xu*, Mohit Agarwal*, Ekansh Gupta, Faramarz Fekri, Raghupathy Sivakumar

Georgia Institute of Technology

Atlanta, GA, 30332

Emails: {dxu301, mohit, egupta8, fekri, siva}@ece.gatech.edu

Abstract—**Providing Reinforcement Learning (RL) agents with human feedback can dramatically improve various aspects of learning. However, previous methods require human observer to give inputs explicitly (e.g., press buttons, voice interface), burdening the human in the loop of RL agent's learning process. Further, providing explicit human advise (feedback) continuously is not always possible or too restrictive, e.g., autonomous driving, disabled rehabilitation, etc. In this work, we investigate capturing human's intrinsic reactions as implicit (and natural) feedback through EEG in the form of error-related potentials (ErrP), providing a natural and direct way for humans to improve the RL agent learning. As such, the human intelligence can be integrated via implicit feedback with RL algorithms to accelerate the learning of RL agent. We develop three reasonably complex 2D discrete navigational games to experimentally evaluate the overall performance of the proposed work. And the motivation of using ErrPs as feedbacks is also verified by subjective experiments. Major contributions of our work are as follows, (i) we propose and experimentally validate the zero-shot learning of ErrPs, where the ErrPs can be learned for one game, and transferred to other unseen games, (ii) we propose a novel RL framework for integrating implicit human feedbacks via ErrPs with RL agent, improving the label efficiency and robustness to human mistakes, and (iii) compared to prior works, we scale the application of ErrPs to reasonably complex environments, and demonstrate the significance of our approach for accelerated learning through real user experiments.**

I. INTRODUCTION

AI systems are increasingly applied to real-world tasks that involve interaction with humans. And humans are often in the loop of the RL agent's learning process. Self-driving cars learn with humans ready to intervene in dangerous situations. Facebook's algorithm for recommending trending news stories has humans filtering out inappropriate content. Therefore RL with *human-in-the-loop* has inspired several research efforts where either an alternative (or supplementary) feedback is obtained from the human participant, such as human rankings or ratings [22], human robot interaction and rehabilitation engineering for the disabled [37], [41], or the learning is performed through human demonstrations [46]. Such approaches with explicit human input despite being highly effective, severely burdens the human interacting with RL agent. Further, it is difficult or even impossible to obtain the explicit human feedback in various situations, e.g., autonomous driving, disabled users, etc.

In this work, we investigate an alternative paradigm to obtain the human feedback in an implicit manner (by tapping directly into the intrinsic brainwaves) that substantially increases the richness of the reward functions, while not severely burdening the human-in-the-loop. We study the use of electroencephalogram (EEG) based brain waves of the human-in-the-loop to generate the auxiliary reward functions to augment the learning of RL agent. Such a model will benefit from the natural rich activity of a powerful sensor (the human brain), but at the same time not burden the human since the activity being relied upon is *intrinsic*. This paradigm is inspired by a high-level error-processing system in humans that generates error-related potential (ErrP) [56], [10], a negative deflection in the ongoing EEG signals. When a human recognizes an error made by an agent, the elicited ErrP can be captured through EEG to inform agent about the sub-optimality of the taken action in the particular state. Human feedback obtained in this manner is direct and fast while being natural and easy for humans. This widens the applicability of such RL-human interactive systems where the RL agents are deployed in the real-world environment, and increased latency of human feedback could create unwanted situations. Further, obtaining large amount of explicit feedback is infeasible due to the increased cognitive load [51]. Additionally, EEG-based feedback allows disabled users to provide the feedback, where explicit communication pathway is not available.

Previous works have [18], [54] demonstrated the benefit of error-potentials in a very simple setting (i.e., very small state-space, and two actions), and used ErrPs as the only reward. As a baseline contribution, we scale the feasibility of capturing error-potentials (of a human observer watching an agent learning to play games) to reasonably complex environments, and then experimentally show that decoded ErrPs can be appropriately used as an auxiliary reward function to a RL agent. In order to validate the motivation of using ErrPs as feedbacks, we make a case for using ErrPs by specifying the advantages it offers over other brain potentials and perform user studies to show that obtaining human feedback implicitly through ErrPs outperforms explicit human labeling in terms of labeling accuracy, latency of feedback and user comfort. We also show that the *full access* approach, inquiring human feedback on every state-action pair visited by RL agent, can significantly speedup the learning of the RL agent.

Despite the accelerated performance of *full access* approach, it is not scalable to complex environments with many state-

action pairs. Since the *full access* relies on the implicit human feedback on each and every state-action pair, it would be extremely inconvenient, impractical, and time consuming for the end-user (even when the user is providing the feedback without any explicit actions). Furthermore, the EEG-based implicit feedback is stochastic, (i.e., the error rate of decoding human feedback is not perfectly zero) which could possibly diverge the training of the RL agent when the feedback has relatively high error rate.

In this context, we first argue that the definition of ErrPs can be learned in a zero-shot manner across different environments. We experimentally validate that ErrPs of an observer can be learned to decode for a specific game, and the definition can be used as-is for another game without requiring re-learning of the ErrP decoding. This is notably different from previous approaches [18], [54], where the labeled ErrPs are obtained in the same environment the RL agent is trying to solve. We contend that previous approaches are not practical, since ErrP decoder cannot be trained and tested in the same environment.

We develop a framework to integrate deep RL (DRL) model with the implicit human feedback mechanism (via ErrP) in a practical, sample-efficient manner. Our proposed framework allows humans to provide their feedback implicitly prior to the agent training, reducing the cognitive load on humans, and hence the cost of human supervision. In the presented framework, prior to the training of RL agent, we present randomly generated demonstrations to a human for giving feedback (implicitly via ErrP), and learn an auxiliary reward function to reflect the human decision and intelligence hidden behind ErrP labels. We then pass this auxiliary reward to the RL agent to accelerate the learning process in sparse-reward environments.

Similar previous work lies in the streamline of human-agent interaction via reward shaping [12], [14], [39], [59], [64], [67]. However, the stochasticity (or erroneous nature) in the human feedback was not specifically accounted in previous methods. Hence, false ErrP labels (errors due to the collection and decoding of brainwaves) make the training less robust and unstable. Thus, we learn the auxiliary reward from human feedback in a way of being robust to wrong ErrP labels. It is assumed that human feedbacks coming from the optimal policy according to human intelligence. In order to tackle wrong feedbacks, we model that optimal policy as a soft-Q policy [33] and learn the corresponding Q function via maximum likelihood with ErrP labels, where the probabilistic modeling has higher robustness than deterministic one. Then in order to make the learned Q function more compatible with the state space, we introduce a baseline function to smoothen that. Finally, at the RL agent side, the environmental reward (sparse) and the auxiliary reward learned from human feedback are combined to form the received reward.

We present results of real ErrP experiments to evaluate the acceleration in learning, and sample efficiency, of the proposed frameworks. We show that such implicit feedback approach can accelerate the training of RL agent by 2.25x, while reducing the number of queries required by 75.56%. In summary, the novel contributions of our work are,

1) We demonstrate the zero-shot learning of error-potentials over various visual-based RL problems (discrete grid-based navigation games, studied in this work), enabling the estimation of implicit human feedback in new and unseen environments without re-training of ErrP decoder. We also verify the superiority of ErrPs over manual feedback by subjective experiments, in terms of labeling accuracy, latency of feedback and user comfort.

2) In order to reduce the sample complexity of ErrP labels, we propose a new framework of integrating human feedback into RL via reward shaping. It is a novel approach specifically considering robustness against mistakes in human feedback. We first generate a set of random trajectories by Monte Carlo Tree Search (MCTS), balancing exploration and exploitation. Then we collect ErrP labels in experiments by demonstrating these trajectories to human observers. By learning optimal Q function with decoded labels, we derive an auxiliary reward function to augment the learning of the following RL agent.

3) We scale the implicit human feedback (via ErrP) based RL to reasonably complex environments. With subjective analysis of ErrP decoding errors and ablation study, we demonstrate the significance of our approach through experiments on various human subjects.

Our work demonstrates the potential of intuitive human robot interaction, facilitating robotic control by implicit human feedback in the form of ErrPs. We believe the contribution presented in this work, i.e., zero-shot learning of ErrPs and RL framework to reduce the human cognitive load, would inspire such implicit human feedback system to be deployed in practical robotic applications, such as autonomous driving or end-user applications for disabled, where explicit human feedback is not available.

II. RELATED WORK

The impact of feedback provided by a human to an agent in RL settings has been investigated by multiple researchers. A survey of recent research in using human guidance for deep RL tasks is presented in [69]. We summarize related work in some of these techniques that are most relevant to us. In addition to rewards from environment, reward shaping learns an auxiliary reward function to accelerate the learning process of the agent [16], [17], [61]. [39] presented a framework called TAMER (Training an Agent Manually via Evaluative Reinforcement) that enabled shaping (interactively training an agent via an external signal provided by a human). Then the author extended this work to enable human feedback to augment an RL agent that learned using an MDP reward signal [40], [41]. Recently an architecture called Deep-TAMER [64] has extended the TAMER framework to environments with high-dimensional state spaces. DQN-TAMER [5] modeled other characteristics of human observers, such as facial expressions, from which human reward was inferred.

Human preference [19], [66] is another approach to communicate complex goals to allow systems to interact with real-world environments in a meaningful way. This allowed the RL agent to directly learn from expert preferences. However,

this approach is limited by assumptions on the existence of a (total) order among the set of trajectories. The author proposed a framework called Human-Agent Transfer (HAT) [59]. It directly used demonstrations provided by a human operator to synthesize a baseline policy, which is to guide the learning of the agent. CHAT [63] extended HAT to consider uncertainty in summarizing demonstrations and further improve the performance.

'Potential functions' is also used in Potential-based reward shaping (PBRS) methods to accelerate the learning process, while preserving the identity of optimal policies [46], [65], [68]. The potential function was designed to encode 'rules' of the environment of the RL agent. However, potential functions will typically need to be pre-specified. This has restricted the use of PBRS to tabular / low-dimensional state spaces.

In previous reward-shaping work mentioned above, human feedback is explicit, requiring active human labeling or attention, and the mistakes in human feedback are not specifically tackled. Here we propose to read implicit human feedback from error-potential hidden in human brain waves, and deal with wrong feedback in a robust approach. Recently, there is a long line of papers studying reinforcement learning from human feedback, such as [21], [22], [62], [66], [19]. However, they are only about explicit human feedback or labeling, and they all assume human feedback is noiseless. In this work, we use reward function learned by imitation learning to augment the following RL agent.

Numerous works [15], [36], [35] have studied a high-level error-processing system in humans generating the error-related potential/negativity (ErrP or ERN).

Interaction, response, and feedback ErrPs have been heavily investigated in the domain of choice reaction tasks, where human is actively interacting with the system [55], [11], [50], [26], [27] and the error is made either by the human or by the machine. [38] demonstrated the use of ErrP signals in an interactive RL task, when the human is actively interacting with the machine system. [26] explored the ErrPs when human is silently observing the machine actions (and does not actively interact). Works at the intersection of ErrP and RL [18], [54] demonstrate the benefit of ErrPs in a very simple setting (i.e., very small state-space), and use ErrP-based feedback as the only reward. Moreover, in all of these works, the ErrP decoder is trained on a similar game (or robotic task), essentially using the knowledge that is supposed to be unknown in the RL task. In our work, we use labeled ErrPs examples of very simple and known environments to train the ErrP decoder, and integrate ErrP with DRL in a sample-efficient manner for reasonably complex environments.

III. PRELIMINARIES AND SETUP

A. RL and the Q-function

Definitions: Consider a Markov Decision Process (MDP) problem M, as a tuple $< \mathcal{X}, \mathcal{A}, P, P_0, R, \gamma >$, with state-space $\mathcal{X}$, action-space $\mathcal{A}$, transition kernel P, initial state distribution P_0, accompanied with reward function R, and discounting factor $0 \leq \gamma \leq 1$. In this work, we only consider MDP with discrete actions and states. In model-free

RL method, the central idea of most prominent approaches is to learn the Q-function by minimizing the Bellman residual, i.e., $\mathcal{L}(Q) = \mathcal{E}_\pi \big[\big(Q(x, a) - r - \gamma Q(x', \hat{a}) \big)^2 \big]$, and temporal difference (TD) update where the transition tuple (x, a, r, x') consists of state, action, reward and next state respectively, i.e., a consecutive experience under behavior policy π.

Bayesian Deep Q Network: The Q function model adopted in this paper is *Bayesian DQN* [6]. It is a neural architecture where the Q-function is approximated as a linear function, weighted by $\omega_a, a \in \mathcal{A}$, of the feature representation of states $\phi_\theta(x) \in \mathcal{R}^d$, parameterized by neural network with weights θ. The weights ω_a follow the Gaussian distribution from Bayesian linear regression.

B. Games and EEG

Using a game as a proxy for a real-life environment is beneficial in the context of human assisted RL algorithms. Games are a fertile ground for the definition, understanding, and improvement of RL algorithms in a low overhead and speedy fashion. Games have now evolved to help understand the world around us and make optimal strategies to tackle various difficult and high-risk real-world situations. For example, Foldit is an online puzzle video game about protein folding. The users of the game helped to solve the structure of a protein-sniping enzyme critical for reproduction of the AIDS virus. A curious planet with four stars was discovered through another game, Planet Hunter, along with the discovery of 40 other planets with the potential of having life-forms [60]. Motivated by these studies, we use games as environments for gathering ErrP data from humans in order to accelerate our RL algorithm.

Specifically, we use electroencephalogram (EEG) signals to generate the implicit feedback that can be used by the RL algorithm. Electroencephalography (EEG) is a mechanism to detect electrical activity in a human brain using small, metal discs (electrodes) attached to the scalp. Brain cells communicate via electrical impulses and EEG helps in recording this activity. EEG is growing to be a bonafide and easy to use [3] input modality in several applications such as communication [1] [31], lifestyle [2], RL [4] etc. and due to the wider availability of EEG headsets off-the-shelf, access to a user's EEG data is easier than it has ever been. We use EEG because error potentials, which are an outcome of a high level error-performance system, manifest themselves as a negative deflection in the EEG signal activity of a human [56], [10], [47]. This is then detected and used into the reward function of the reinforcement learning algorithm. We also make the case for using error potentials over manual labeling in section IV.

C. System Setup and Data Collection

1) BCI for implicit feedback: We designed and developed an experimental protocol, where a machine agent plays a computer game, while a human silently observes (and assesses) the actions taken by the machine agent. These implicit human reactions are captured by placing raw electrodes on the scalp of the human brain in the form of EEG potentials. The electrode cap (BIOPAC CAP-100C) was attached with

the OpenBCI Cyton[1] platform, which was further connected to a desktop machine over the wireless channel. In the game design (developed on OpenAI Gym), we open a TCP port, and continuously transmit the current state-action pair using the TCP/IP protocol. We used OpenViBE software [52] to record the human EEG data. OpenViBE continuously listens to the TCP port (for state-action pairs), and timestamps the EEG data in a synchronized manner. We recruited a total of five human subjects (mean age 26.8 with standard deviation of 1.92, 1 female) using standard procedures with their consent. For each subject-game pair, the experimental duration was less than 15 minutes. The agent took action every 1.5 seconds during the experiment. The University Institutional Review Board reviewed and approved all the research protocols for the user data collection.

2) The Games: We have developed three discrete grid-based navigational games in OpenAI Gym *Atari* framework, namely Wobble, Catch, and Maze (Fig. 1(a)).

Wobble: Wobble is a simple 1-D cursor-target game, where the middle horizontal plane is divided into 20 discrete blocks. At the beginning of the game, the cursor appears at the center of the screen, and the target appears no more than three blocks away from the cursor position. The action space constitutes moving to the left or right. The game is finished when the cursor reaches the target. Once the game is finished, a new game is started with the cursor in place.

Catch: Catch is a simplistic version of *Eggomania*[2] (Atari 2600 benchmark), where we display a single egg on the screen at a time. The has a 10x10 grid, where the *egg* and the *cart*, both occupy one block. The action space of the agent consists of *NOOP* (no operation), *left* and *right*. At the start of the game, the horizontal position of the egg is chosen randomly. At each time step, the *egg* falls one block in the vertical direction.

Maze: Maze is a 2-D navigational game, where the agent has to reach a fixed target (shown with a *plus* symbol). The screen is divided into 10x10 square blocks. The action space consists of four directional movements. The only reward here is the result of the episode, i.e., win or lose. If an agent moves, but hits a wall, a quick blinking of the agent is displayed, to render the action taken by the agent.

## IV. The Case for Error Potentials

We make the case for using error potentials (ErrP) by first, touching upon some of their characteristic advantages that have been documented in academic literature, and then by providing experimental evidence of the superiority of ErrPs over manual or explicit labeling. We show that using ErrPs as feedbacks reduces the cognitive burden of the subjects, and provides better trade-off between latency and correctness of the feedback.
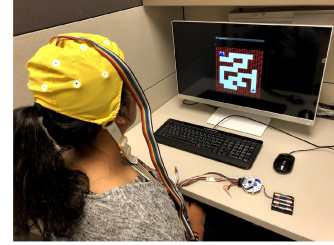
A. The usefulness of ErrPs for error detection

To begin with, relying on error-potentials provides two primary benefits:
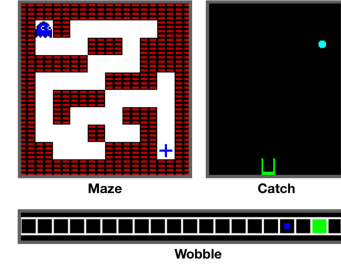
(a) Game Environments

(b) Experiment Bench
Fig. 1: Experimental framework

(a) **Provides a generalized notion of error-detection:** Error-potentials are elicited when a user is presented with an incongruent (or erroneous) stimulus in a diverse set of tasks [29] implying that the error-processing system is generic (i.e., not specific to the task or sensory organ). Error-potentials are observed across a wide variety of input modality (e.g., audio [23], visual [25], somatosensory [45], etc.). This is in contrast to other elicited potentials in the brain which cater to the stimuli of a specific category. For instance, the P600, N300, P300, and N200 are elicited when a subject is presented with syntactic anomalies in sentences [49], semantically inconsistent word and picture pairs [44], interruption of a stimulus with another divergent stimulus [58], and detection of mismatch in a stimulus [28] respectively. Thus, the generalized mechanism for eliciting ErrPs is one of the characteristic advantage that it offers, unlike other brain-potentials specific to a stimulus or modality.

(b) **Evolutionary Significance:** Error-potentials in primates are well-founded and universal (exhibiting similar behaviours across individuals) as they have an evolutionary significance due to their importance in cognition, learning, and survival. Error-potentials enable the learning process via the administration of rewards and punishments in Anterior Cingulate Cortex (ACC) [34]. In monkeys, error-potentials were generated in anterior cingulate sulcus, when monkeys made errors in a simple response task. [48] found error-recognition units in monkeys' anterior cingulate sulcus that were activated when the animals received negative feedback in the form of absence of an expected reward. Similarly, [30] found that when monkeys made errors in a simple response task, error-related potentials were generated in the anterior cingulate sulcus, thereby advocating that ErrPs link human and non human primates on the basis of error monitoring. Universality of ErrPs guarantees that it occurs naturally in humans

and the evolutionary importance of ErrPs in learning points toward them being a foundational element in the human cognition, enabling their use without worrying about individual differences based on learned behavior.

B. Empirical motivation for using intrinsic error-potentials over manual labeling

Experimental Methodology: We conducted an experiment in which we asked subjects to label the actions of an AI agent in a maze. If the agent took a correct action, they needed to press a certain key and if the agent performed a wrong action, they were supposed to press another key. We conducted this experiment to find differences between manual labeling and labeling using EEG experiments in terms of user comfort and labeling accuracy. The methodology of this experiment was simple. We designed a maze game and generated 3 instances of it where each instance got progressively faster (to study the impact of time pressure on mental comfort and accuracy). The first instance had a time delay of 1.5 seconds between successive actions of the agent while the second and the third had a delay of 1.0 and 0.5 seconds respectively (We use these delay values as they lie around the latency value we have used in the EEG experiments and they also help us know the variation of manual labeling accuracy with respect to latency). We made the subjects play 3 trials in each instance (thus totaling to 9 trials overall) where we randomized the order of instances to avoid biasing the users to a particular order of the game. In all these instances, the AI agent made the correct move with a probability of 0.8. Once subjects finished playing 3 trials of an instance, we redirected them to a Qualtrics survey where they had to provide their subjective feedback about the experiment. Thus, there were 3 forms that each subject had to fill (one per instance). We used Amazon's Mechanical Turk to request anonymous workers to complete this task. The Institutional Review Board approved the study protocol.

Results: We obtained a total of 281 responses (87, 91, and 103 unique user responses for the 1.5s, 1.0s, and 0.5s instances of the game respectively). On average, for the 1.5s instance of the game, we obtained a true positive rate of 56.6% and 41.5% for correct and incorrect actions of the maze agent respectively. We also obtained a feedback latency of 376ms and 540ms for correct and incorrect actions of the maze agent respectively. Note that correct and incorrect actions of the agent corresponds to the non-Errp and ErrP respectively, during EEG experiments. For the 1.0s instance of the game, we obtained a true positive rate of 49.8% and 38.8% for correct and incorrect actions of the maze agent respectively. We also obtained a feedback latency of 288ms and 456ms for correct and incorrect actions of the maze agent respectively. For the 0.5s instance of the game, we obtained a true positive rate of 34.9% and 14.1% for correct and incorrect actions of the maze agent respectively. We also obtained a feedback latency of 179ms and 207ms for correct and incorrect actions of the maze agent respectively. However some trials in these experiments had very poor labeling rate (some subjects only labeled less than 50% of the actions available). In order to prevent the results from being swayed by inert participants, we decided to separate them from the active participants.
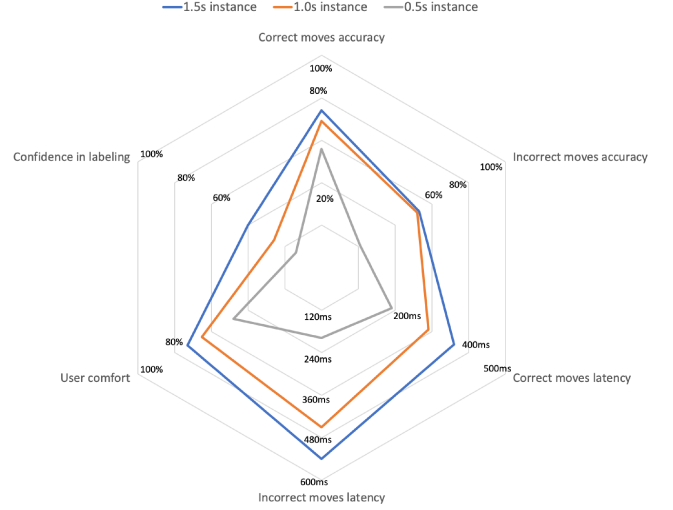

Fig. 2: Difference between instances of the maze game used for manual labeling

TABLE I: Accuracy and latency for maze game manual labeling

Time Interval (s)	Subjects	Correct Moves TPR (%)	Incorrect Moves TPR (%)	Correct Moves latency (ms)	Incorrect Moves latency (ms)
1.5	87	74.1	53.4	364	539
1.0	91	69.8	52.6	290	451
0.5	103	56.4	21.6	177	203

We decided to remove the trials for the users which had less than 50% feedback rate. In other words, we removed the trials where participants failed to provide the feedback for at least 50% of all the actions. This led to the removal of 22 users from the 1.5s instance of the game (25%), 28 users from the 1.0s instance of the game (31%), and 44 users from the 0.5s instance of the game (43%). After this filtering, for the 1.5s instance of the game, we obtained a true positive rate of 74.1% and 53.4% for correct and incorrect actions of the maze agent respectively. We also obtained a feedback latency of 364ms and 539ms for correct and incorrect actions of the maze agent respectively. For the 1.0s instance of the game, we obtained a true positive rate of 69.8% and 52.6% for correct and incorrect actions of the maze agent respectively. We also obtained a feedback latency of 290ms and 451ms for correct and incorrect actions of the maze agent respectively. For the 0.5s instance of the game, we obtained a true positive rate of 56.4% and 21.6% for correct and incorrect actions of the maze agent respectively. We also obtained a feedback latency of 177ms and 203ms for correct and incorrect actions of the maze agent respectively. These values are summarized in Table I and are contrasted pictorially in Fig 2. The results of manual labeling are also contrasted with those of implicit feedback using EEG in Fig 3.

Insights: As we can clearly see, the accuracy values for correct and incorrect actions both decrease with decrease in time interval. The labeling accuracy for correct actions seems to be more than that of incorrect actions. Both, the accuracy for correct as well as incorrect actions decreases as the time latency is decreased (thereby increasing time pressure). Even
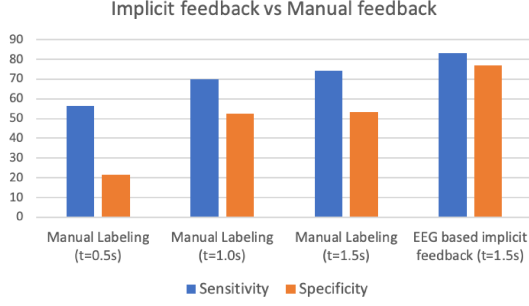
Fig. 3: Manual labeling compared with implicit feedback using EEG on maze game

the best possible accuracy for incorrect actions is about 53.4% (only marginally better than random labeling). This performs rather poorly compared to the labeling accuracy using ErrP.

Based on the qualitative survey responses, on a scale of 1 to 7, the users gave the 1.5s instance of the game an average comfort rating of 5.4 which declined to 4.9 and 3.9 for the 1.0s instance and 0.5s instance respectively. On being asked if they were able to mark all actions correctly, 40% of the subjects answered in the affirmative in the 1.5s instance of the game, which declined to 26% and 14% in the 1.0s and the 0.5s instance of the game. Across the board, the majority of the participants reported that the ideal time interval for them to correctly label all actions of the agent would be between 1.5s and 3.0s or larger. 64% of the participants in the 1.5s instance of the game reported that reducing the time interval of the game to 1.0s would decrease their labeling accuracy, and 69% of the participants reported that it would increase their mental burden. 52% of the participants in the 1.0s instance of the game reported that reducing the time interval of the game to 0.5s would decrease their labeling accuracy, and 60% of the participants reported that it would increase their mental burden. In contrast, 64% of the participants in the 1.0s instance of the game reported that increasing the time interval from 1.0s to 1.5s would increase their labeling accuracy and decrease their mental burden. 49% of the participants in the 0.5s instance of the game reported that reducing the time interval of the game further would decrease their labeling accuracy, and 53% of the participants reported that it would increase their mental burden. To summarize, the users felt increasing discomfort and cognitive burden as the time latency reduced from 1.5s to 1.0s and further to 0.5s. They also reported that the optimal time latency for comfortable manual labeling would be between 1.5s and 3s. This was also evident from the fact that more than 60% of the participants anticipated reduction in their accuracy if time latency was to be decreased from 1.5s.

V. INTEGRATING RL WITH IMPLICIT HUMAN FEEDBACK: A NAIVE APPROACH

In this section, we provide our baseline contribution, i.e., (i) we demonstrate the feasibility of capturing error-potentials of a human subject watching an RL agent learning to play several different games, and then decoding the human feedback

(judgment) on the observed state-action pair appropriately, and (ii) using them as an auxiliary reward function to accelerate the learning of the RL agent.

A. Obtaining the Implicit Human Feedback: Decoding ErrPs

In order to obtain the implicit human feedback, we need to detect the presence or absence of ErrPs inside the EEG waveform. This requires training a model that can interpret the EEG signal of a human and classify it as ErrP or non-ErrP correctly. EEG signals are inherently very noisy, and when combined with external factors like improper electrode placements, variance across users pose significant challenges in the reliable estimation of error-potentials.

We rely on the Riemannian Geometry framework for the classification of a human's intrinsic reaction [9].This framework is state-of-the-art for detecting any event-related potentials, and provides two primary advantages over other classifiers:

- The estimation algorithm operates in signal space (rather than source space), and hence minimizes the distortions due to the electrode placements.
- The spatial filtering algorithm maximizes the signal to signal plus noise ratio (SSNR) to mitigate the interference and noise.

We consider the classification of error-related potentials as a binary classification task indicating the presence (i.e., action taken by the agent is incorrect) and absence of error (i.e., action taken by the agent is correct). We bandpass filtered the raw EEG data within [0.5, 40] Hz. We then extracted Epochs of 800ms relative to pre-stimulus 200ms baseline and subjected them to spatial filtering. In spatial filtering, we compute prototype responses of each class, i.e., "correct" and "erroneous", by averaging all training trials in the corresponding classes("xDAWN Spatial Filter" [53], [20]). "xDAWN filtering" projects the EEG signals from sensor space (i.e., electrode space) to the source space (i.e., a low-dimensional space constituted by the actual neuronal ensembles in brain firing coherently). We then compute the covariance matrix of each epoch, and concatenate it with the prototype responses of the class. Further, we select relevant channels through backward elimination [7] to achieve dimensionality reduction and project the filtered signals on the tangent space [8] for feature extraction. We then normalize (using L1 norm) the obtained feature vector and feed it to a regularized regression model. We select a threshold value for the final decision by maximizing accuracy offline on the training set. We present the algorithm to decode the ErrP signals in Algorithm 1.

B. The Full Access Method

A naive approach to integrate the human feedback with RL models is reward shaping with *full access*. We obtain human feedback on every visited state-action pair while RL agent is learning, and add a negative penalty to the environmental reward in case an ErrP is detected. We present the evaluation result of this method based on real ErrP data later in the evaluation section (section VII-A), validating that *full access*

Algorithm 1: Riemannian Geometry based ErrP classification algorithm

---
Input : raw EEG signals EEG
1 Pre-process raw EEG signals ;
2 Spatial Filtering: xDAWN Spatial Filter ($nfilter$) ;
3 Electrode Selection: ElectrodeSelect ($nelec$, metric='riemann') ;
4 Tangent Space Projection : TangentSpace(metric = "logeuclid") Normalize using L1 norm ;
5 Regression: ElasticNet ;
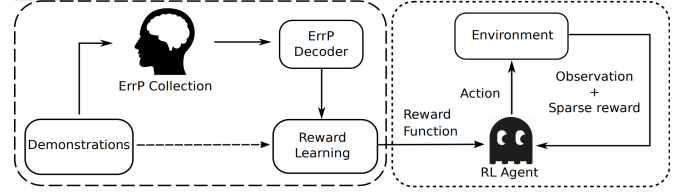6 Select decision threshold by maximizing accuracy

---

Fig. 4: Robust Reward Shaping with Human Feedback. The dashed arrow shows trajectories in $\mathcal{D} \cup \mathcal{D}_R$ are all used in reward learning

method can significantly accelerate the learning of the RL agent. However, obtaining the human feedback for every state-action pair is time-intensive and not practically feasible. In the next section, we provide our novel contributions to practically obtain and integrate the implicit feedback with the learning of RL agent.

Interaction, response, and feedback ErrPs have been heavily investigated in the domain of choice reaction tasks, where human is actively interacting with the system [55], [11], [50], [26], [27] and the error is made either by the human or by the machine. [38] demonstrated the use of ErrP signals in an interactive RL task, when the human is actively interacting with the machine system. [26] explored the ErrPs when human is silently observing the machine actions (and does not actively interact). Works at the intersection of ErrP and RL [18], [54] demonstrate the benefit of ErrPs in a very simple setting (i.e., very small state-space), and use ErrP-based feedback as the only reward. Moreover, in all of these works, the ErrP decoder is trained on a similar game (or robotic task), essentially using the knowledge that is supposed to be unknown in the RL task. In our work, we use labeled ErrPs examples of very simple and known environments to train the ErrP decoder, and integrate ErrP with DRL in a sample-efficient manner for reasonably complex environments.

VI. TOWARD SMARTER INTEGRATION OF RL WITH IMPLICIT HUMAN FEEDBACK

In this section, we propose two approaches to enable the deployment of ErrP-augmented RL into practical systems. Firstly, we show that we can learn the ErrPs of an observer in a zero-shot manner, i.e. We can train an ErrP decoder for a specific game, and use the trained decoder as-is for another game without re-training the ErrP decoder. To combat with the practical issues with obtaining ErrP labels for every state-action pairs, we propose an RL framework (motivated by imitation learning approaches) allowing humans to provide their feedback on a few trajectories prior to the learning of the RL agent. This dramatically reduces the number of feedback labels required from the human observer.

A. Robust Reward Shaping using Human Feedback

RL algorithms deployed in the environment with sparse rewards demand heavy explorations (require a large number of trial-and-errors) during the initial stages of training. In such environments, using human feedback can be very efficient for accelerating the learning process. Previous work on reward shaping with human feedback [14], [39], [59], [64], [67] build a specific model to generalize human feedback in state space, without tackling wrong feedback. Inspired by soft Q policy [33], we develop a novel framework of learning the auxiliary reward from human feedback to accelerate the training of the RL agent, with robustness to mistakes in ErrP labeling.

In this framework, we require implicit human feedback via ErrP on all state-action pairs along trajectories (demonstrations) randomly generated initially. Before RL agent starts learning, we ask the human subjects to observe a number of trajectories, and record their implicit feedbacks in the form of ErrP on corresponding state-action pair in a dataset. Then we learn the auxiliary reward function $r_a(\cdot, \cdot)$ from these trajectories labeled by human feedback. During the RL training, the learned reward function acts as a proxy for the human feedback, compensating the sparse reward from the environment. The flowchart of the proposed learning framework is shown in Figure 4. Different from the naive baseline *full access* method discussed earlier, in this approach, we require the queries for human feedback (ErrP labeling) only on trajectories generated initially, instead of querying every learning step during the training. Hence, the total number of ErrP queries are reduced significantly, further reducing the load for the human in the loop.

Trajectory Generation: Constraint by the coherence requirement in EEG experiments, the trajectories for ErrP labeling have to be complete, containing every state-action pair from the beginning to the end of the game. Further, the selected trajectories have to cover state space as much as possible, and cannot be too far away from the optimal solutions. This is essentially the trade-off between exploitation and exploration. So we propose to use Monte Carlo Tree Search (MCTS) [13], [42], [57] to generate random trajectories for ErrP experiments. It is to tackle exploration-exploitation trade-off by Upper Confidence Bound (UCB) method [42], and does not require to know the optimal solution a priori. MCTS is a general game playing technique with recent success in discrete, turn-based, and non-deterministic game domains. We choose MCTS as a trajectory sampling algorithm for its proven high-level performance, domain generality, and variable computational bounds.

In MCTS, there is one node in the tree for each state s, containing a value $Q(s, a)$ and a visit count $N(s, a)$ for each action a, and an overall count $N(s) = \sum_a N(s, a)$. Each node is initialised to $Q(s, a) = 0, N(s, a) = 0$. The value is estimated by the mean return from s in all simulations where

action a was selected in state s, and the only reward r here is the result of the game, i.e., 1 for winning and 0 for losing. At each state s of the trajectory, the action is selected to be the maximizer of the objective $Q(s,a) + c\sqrt{\frac{\log N(s)}{N(s,a)}}$, where c is to trade off between reaching the target and exploring more state space [42]. By the end of generating each trajectory, the return is back-propagated into Q values along the trajectory, i.e., $Q(s_t, a_t) := r + \gamma Q(s_{t+1}, a_{t+1})$. Only first K generated trajectories are used in ErrP experiments.

In experiments for collecting ErrPs, the human subject provides implicit feedback (via ErrP) over all the generated trajectories, labeling every state-action pair as a *positive* (if a correct action according to perceived human intelligence) or a *negative* sample. With decoded ErrP labels over trajectories as input, we propose a novel reward shaping method to incorporate ErrP labels into the reinforcement learning framework. It specifically tackles the problem of robustness against wrong ErrP labels, with details explained in the following section.

Reward Learning Since implicit human feedback via ErrP is noisy, different from previous work [14], [39], [59], [64], [67], instead of modeling human feedbacks by neural networks directly, we assume that human give feedbacks according to his probabilistic policy, modeled by a soft-Q policy with Q function $Q_{h,\theta}$ (with weights θ), under the max-entropy principle [70]. Our method is to define operators first going from demonstrations labeled by human feedbacks to the optimal Q function $Q_{h,\theta}$, and then from that Q function to the auxiliary reward function $r_a(\cdot)$. Here we assume that optimal Q function defines a soft-Q policy [32], [33], and learn it by solving a probabilistic classification problem via maximizing log-likelihood of human feedbacks. Following the principle of maximum entropy [70], the soft-Q policy giving human feedbacks and the corresponding value function can be expressed as follows,

$$
\begin{aligned}
\pi_{h,\theta}(\boldsymbol{a}|\boldsymbol{s}) &= \exp((Q_{h,\theta}(\boldsymbol{s},\boldsymbol{a}) - V_{h,\theta}(\boldsymbol{s}))/\alpha), \\
V_{h,\theta}(\boldsymbol{s}) &= \alpha \log \sum_a \exp(Q_{h,\theta}(\boldsymbol{s},\boldsymbol{a})/\alpha)
\end{aligned}
\tag{1}
$$

where α is a free parameter, tuned empirically. Define positive samples as state-action pairs with correct labels in human feedbacks while negative samples as those with wrong labels. According to the maximum entropy principle [70], the likelihood of positive and negative samples are denoted as $\pi_{h,\theta}(\boldsymbol{a}|\boldsymbol{s})$ and $1 - \pi_{h,\theta}(\boldsymbol{a}|\boldsymbol{s})$, respectively. When trajectories and human feedbacks (ErrP labels) are ready, we learn $Q_{h,\theta}(\cdot,\cdot)$ by maximizing the likelihood of both positive and negative state-action pairs in the trajectories, which is to maximize the objective (3) in Algorithm 2, where the binary variable $\text{ErrP}(s,a)$ denotes the human feedback label. To derive auxiliary reward from the learned Q function, a naive choice is the Bellman difference, i.e., $Q_{h,\theta}(\boldsymbol{s},\boldsymbol{a}) - \gamma \max_{\boldsymbol{a}} Q_{h,\theta}(\boldsymbol{s},\boldsymbol{a})$. However, due to the scarce of ErrP labels on exact state-action pairs, the function $Q_{h,\theta}$ learned by maximum likelihood may not have the shape compatible with the state dynamics of the target MDP (environments in experiments). And the derived auxiliary reward function can destabilize the learning process of RL agent.

In order to refine the reward shape and attenuate the gradient variance, we introduce another baseline function $t_\phi(\boldsymbol{s})$ only dependant on the state, to incorporate the state transition information, parametrized by ϕ. Hence, the Q function is augmented as $Q_B(\boldsymbol{s},\boldsymbol{a}) := Q_{h,\theta}(\boldsymbol{s},\boldsymbol{a}) + t(\boldsymbol{s})$. It can be proved that $Q_B(\cdot,\cdot)$ and $Q_{h,\theta}(\cdot,\cdot)$ induce the same optimal policy [46]. The baseline function $t^*(\cdot)$ can be learned by optimizing $t^* = \arg\min_\phi J_2(\phi)$, defined in (4), where the loss function $l(\cdot)$ is chosen to be l_1-norm via empirical evaluations .

For learning function $t_\phi(\cdot)$, in addition to the demonstration $\mathcal{D}$ in Figure 4, we incorporate another set of demonstrations $\mathcal{D}_R$, containing transitions that are randomly sampled from environment without the reward information. The set $\mathcal{D}_R$ is to help the learned auxiliary reward function to incorporate the state dynamics information, without the need of any human labeling. After learning both $Q_{h,\theta}(\cdot,\cdot)$ and $t_\phi(\cdot)$, for any transition tuple $(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')$, the auxiliary reward function can be represented as

$$
r_a(\boldsymbol{s},\boldsymbol{a}) = Q_{h,\theta}(\boldsymbol{s},\boldsymbol{a}) + t_\phi(\boldsymbol{s}) - \gamma \max_{\boldsymbol{a}'\in\mathcal{A}}[Q_{h,\theta}(\boldsymbol{s}',\boldsymbol{a}') + t_\phi(\boldsymbol{s}')]
\tag{2}
$$

This r_a is then used to augment the RL agent. In order to further attenuate the negative influence of wrong ErrP labels, when combining environmental reward r_e and auxiliary reward r_a, we propose a coefficient $\beta(e)$, exponentially decreasing in terms of training episodes e, i.e., $\beta(e) := ae^{-e/b}$. Finally, the RL agent receives the shaped reward in the form of $r_e(\boldsymbol{s}_t, \boldsymbol{a}_t) + \beta(e)r_a(\boldsymbol{s}_t, \boldsymbol{a}_t)$. Empirically, the best coefficient function is $\beta(e) = 3e^{-e/80}$ in experiments.

Algorithm 2: Robust Reward Shaping with Human ErrP

Input : Trajectories Given Initially

1 Conduct EEG Experiments for human ErrP to label the state-action pairs along trajectories;

2 With ErrP data collected, use Algorithm 1 to decode ErrP labels, i.e., $\text{ErrP}(\cdot,\cdot)$;

3 Initialize the Q function $Q_{h,\theta}(\cdot,\cdot)$ and baseline $t_\phi(\cdot)$;

4 Learn $Q_{h,\theta}(\cdot,\cdot)$ by optimizing θ

$$
\begin{aligned}
J_1(\theta) := \sum_{(s,a)\in\mathcal{D}} &\pi_{h,\theta}(a|s)(1 - \text{ErrP}(s,a)) \\
&+ (1 - \pi_{h,\theta}(a|s))\text{ErrP}(s,a)
\end{aligned}
\tag{3}
$$

where the relationship between $\pi_{h,\theta}$ and $Q_{h,\theta}$ is defined in (1) ;

5 Learn the baseline function $t_\phi(\cdot)$ by optimizing ϕ

$$
\begin{aligned}
J_2(\phi) := \sum_{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{s}')\in\mathcal{D}\cup\mathcal{D}_R} &l(Q_{h,\theta}(\boldsymbol{s},\boldsymbol{a}) + t_\phi(\boldsymbol{s}) \\
&- \gamma \max_{\boldsymbol{a}'\in\mathcal{A}}(Q_{h,\theta}(\boldsymbol{s}',\boldsymbol{a}') + t_\phi(\boldsymbol{s}')))
\end{aligned}
\tag{4}
$$

Then pass the auxiliary reward function r_a (2) to the RL agent ;

6 RL agent employs any RL framework using the modified reward function $r_e(s,a) + \beta(e)r_a(s,a)$.

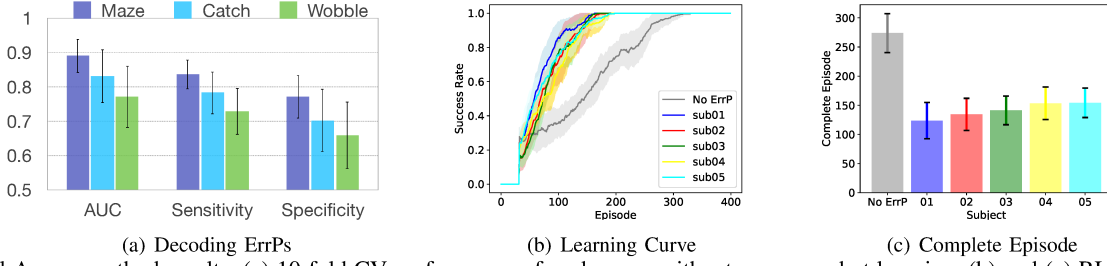(a) Decoding ErrPs   (b) Learning Curve   (c) Complete Episode

Fig. 5: Full Access method results: (a) 10-fold CV performance of each game without any zero-shot learning, (b) and (c) RL with *full access* to ErrP feedback

VII. EVALUATION

A. Naive Approach

We first validate the feasibility of decoding error-potentials using a 10-fold cross-validation scheme for each game. In this scheme, we split the state-action pairs of a game in 10-folds for training and testing of the ErrP decoder. In Figure 5(a), we show the performance of three games in terms of Area Under Curve (AUC) score, sensitivity and specificity, averaged over 5 subjects. The Maze game has the highest AUC score (0.89 ± 0.05) followed by Catch (0.83 ± 0.08) and Wobble (0.77 ± 0.09).

As discussed in section V-B, the *full access* method is the most preliminary approach to incorporate implicit human feedback (in the form of decoded error-potentials) into the DRL model. It asks the external oracle (human) for the implicit feedback in every training step, reaching the maximum number of possible queries. Hence it has the fastest training convergence rate. We use this method as a benchmark for comparing the sample efficiency of the proposed RL framework. The evaluation metric adopted here is *success rate*, which is the ratio of success plays in the last 32 episodes. The training converges and terminates at *complete episode*, when the success rate reaches to 1. The results with real ErrP data of 5 subjects are shown in Figure 5(b,c). We can see there is a significant improvement in the training convergence with all subjects when ErrP used. Here, *No ErrP* refers to the BDQN performance without integrating the human feedback. In all plots of this paper, solid lines are average values over 10 random seeds, and shaded regions correspond to one standard deviation. We use BDQN (as introduced in section III-A) as the DRL model for all experiments conducted in this paper. However, the ErrP feedback here can be used to augment any RL algorithm.

B. Evaluation of the Proposed Solution

In this subsection, we evaluate the performance of proposed approaches to practically integrate the implicit human feedback (via EEG) into the DRL algorithms. In addition, we provide subjective analysis of ErrP decoding errors and ablation study of the proposed reward shaping method.

1) Zero-shot learning of ErrPs: Error-potentials in the EEG signals are studied under two major paradigms in human-machine interaction tasks, (i) *feedback and response ErrPs:* error made by human [15], [24], [11], [50], (ii) *interaction ErrPs:* error made by machine in interpreting human intent [26]. Another interesting paradigm is *observation ErrPs*, when a human is watching (and silently assessing) the machine performing a specific task [18]. We make the case for the
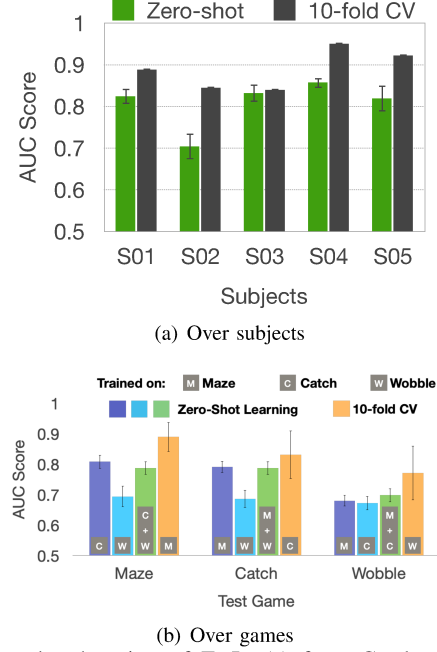

(a) Over subjects

(b) Over games

Fig. 6: Zero-shot learning of ErrP: (a) from Catch to Maze over subjects compared with 10-fold CV, (b) over all combinations of three games compared with 10-fold CV.

generalizability of these ErrPs owing to their universality across humans and other primates in section IV(b). We also observe that the manifestation of these potentials across these paradigms are found quite similar in terms of their general shape, negative and positive peak latency, and frequency characteristics[26], [18]. This prompts us to explore the consistency of the error-potentials across different environments (i.e., games, in our case) within the *observation ErrPs*.

In Figure 7, we plot the grand average waveforms across three environments (Maze, Catch and Wobble), to visually validate the consistency of potentials. We can see that the shape of negativity, and the peak latency is quite consistent across the three game environments. These indicators prompt us to experimentally explore the domain of zero-shot learning for ErrPs in evaluation section VII-B1. We show that by training error potentials on one game, we are able to cover the variability of error potentials in other games as well which suggests that error-potentials are indeed generalizable across environments, and can further be used to inform deep reinforcement learning algorithm in new and unseen environments.

To evaluate the zero-shot learning capability of error-potentials and the decoding algorithm, we train on the samples collected from the Catch game and test on the Maze game. As Catch is a simple game, we assume the optimal action for
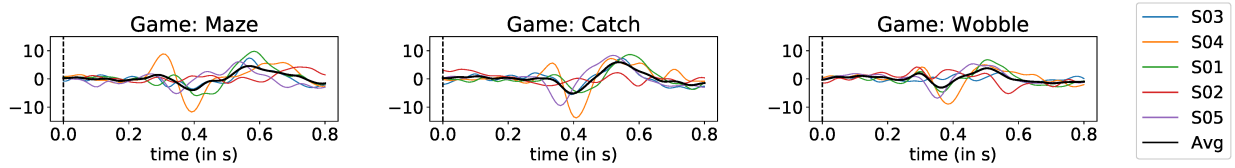
Fig. 7: Manifestation of error-potentials in time-domain: Grand average potentials (error-minus-correct conditions) are shown for Maze, Catch and Wobble game environments. Thick black line denotes the average over all the subjects.

TABLE II: Average Number of Queries on Maze Game

| Subject	01	02	03	04	05
Full access	1879.4	2072.1	2293.7	1975.4	2130.1
Proposed method	505.7	394.7	587.1	681.4	361.3

each state is already known (providing the labeled examples to train the ErrP decoder). However, the Maze game needed to be solved, hence, we do not make any assumptions about the optimality of the actions. In Figure 6(a), we provide the zero-shot learning performance and compare it against the 10-fold Cross-Validation (CV) scheme discussed in section VII-A. Further, we present the AUC score of zero-shot learning performance over all training and testing combinations in Figure 6(b). We use the Area Under Curve (AUC) as the performance metric for the decoding of error-potentials. We can see that the ErrPs recorded for Catch game, are able to capture more than 80% of the variability in the ErrPs for Maze game. Averaged over 5 subjects, the decoder performs with an AUC score of 0.8078 ($\pm$0.022) when trained on the Catch game. This compared with the performance of 0.693 ($\pm$0.034) when trained using Wobble labels. Similarly, Catch and Wobble performs with an average AUC score of 0.790 ($\pm$0.018) and 0.680 ($\pm$0.018) respectively, when trained on labels obtained through the Maze environment. These experiments validate that the error-potentials can be learned in a zero-shot manner to avoid re-training of the human feedback (via EEG) decoder.

2) Evaluation of Robust Reward Shaping with Human ErrP: For the evaluation of Algorithm 2, we generated stochastic trajectories for the Maze game by Monte Carlo Tree Search (MCTS) discussed above, where the trade-off parameter c is set to 0.5. Before training the RL agent, each human subject provided implicit feedback (via ErrP) as explained in the experimental protocol (section III-C) on every state-action pair along these trajectories. We evaluated the performance of the proposed approach with 10 and 20 initial trajectories, in the demonstrations in Figure 4, each for 5 subjects. We use the Bayesian DQN as the DRL model.

The acceleration of RL due to human feedback is shown in Figure 8(a) for 10 trajectories, where the base model is Bayesian DQN. We can see the significant acceleration in training convergence in Figure 8(a) in terms of the *success rate* for 5 subjects and compared against the case of *No ErrP, i.e. no human feedback*. Subject 01 has the highest fidelity for error-potentials, and hence, RL algorithm converges at much faster rate when relies upon the feedback obtained by Subject 01. It is evident from the results that the error-potential decoding performance is sufficient to achieve around 2x improvement in training time (in terms of the number of episodes required).
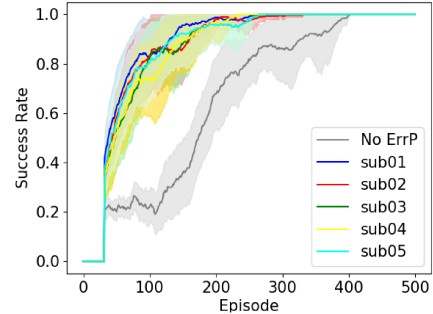
(a) 10 Trajectories
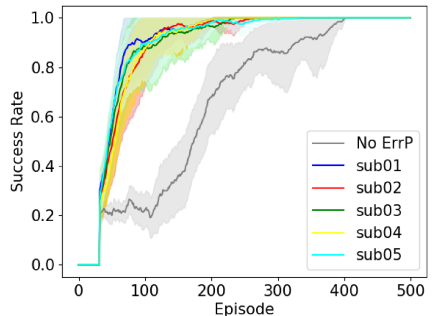
(b) 20 Trajectories

Fig. 8: Evaluation of the proposed reward shaping method.

Similarly, Figures 8(b) shows the *success rate* and convergence curve for training to complete, for 20 trajectories. Comparing Figures 8(a) and (b), we can see that the training converges at much faster rate when the number of initial trajectories are increased. Further, the learning variance also decreases with more trajectories. The comparison between Figure 8 and Figure 5(c) shows that the proposed framework learns faster than No-ErrP case, while outperforming the *full access* case, even though *full access* requires significantly larger amount of queries. We also compare the number of ErrP queries for *full access* and proposed method in Table II, according to the statistics on experiments with 20 trajectories. On an average for 5 subjects, the proposed approach makes 75.56% less queries as compared to the *full access*. As *full access* queries for feedback label at every learning step, i.e. state-action pair, while the proposed framework queries only on the trajectories given initially, the total number of queries made are significantly reduced.

3) Analysis of the dependence and subjectivity of errors: In this section, we analyze the detection accuracy of error-potentials for the Maze game, to develop insights into the char-

TABLE III: Accuracy and standard deviations per subject for ErrP and non-ErrP SAPs

Subject	ErrP SAPs mean	ErrP SAPs std dev	non-ErrP SAPs mean	non-ErrP SAPs std dev
S12	0.79	0.27	0.75	0.17
S07	0.8	0.3	0.85	0.16
S02	0.73	0.29	0.77	0.15
S08	0.6	0.25	0.56	0.14
S01	0.8	0.25	0.77	0.16
S04	0.78	0.25	0.63	0.16
S16	0.73	0.3	0.78	0.14
S03	0.65	0.25	0.61	0.13
S06	0.73	0.3	0.64	0.17
S05	0.75	0.3	0.72	0.13
S09	0.71	0.27	0.66	0.13
S15	0.67	0.31	0.65	0.1
Average	0.73	0.28	0.70	0.14

acteristics of error-potential based on the users and provided stimulations. The EEG samples recorded for the Maze experiment can be presented along two independent dimensions, (i) users and (ii) state-action pair of the agent (i.e., stimulation). Within the state-action pairs (or SAPs for short), if the action is correct, it is called a *non-ErrP SAP*, otherwise an *ErrP SAP*. Please note that the term *non-ErrP state-action pair* or *non-ErrP SAP* refers to a correct action taken by the agent given a state (due to the expected absence of an ErrP response in the brain), and does not refer to a system where we do not use implicit human feedback using EEG.

- **Experiment 1: Subjectivity over correct and incorrect actions.** For each user, we divide the EEG trials into two categories (a) *ErrP SAPs*, and (b) *non-ErrP SAPs*. For each user and category, we compute the mean and standard deviation of classification accuracy of EEG trials, and present in Table III. We can observe that the per user standard deviations for ErrP SAPs is roughly double the standard deviations for non-ErrP SAPs. The aggregate per user standard deviation across the ErrP SAPs is 0.28 and 0.14 for non-ErrP SAPs. This difference in per user standard deviations is statistically significant ($p<0.001$). We also calculate the standard deviations across our user accuracy vectors for both ErrP SAPs and non-ErrP SAPs and find that the standard deviations for the per user accuracy vectors are 0.06 and 0.08 respectively.

- **Experiment 2: Subjectivity over users.** In this experiment, for each unique state-action pair, we average the performance of EEG trials of all users. We achieved a mean and standard deviation of 0.75 and 0.13, and 0.75 and 0.07 for *ErrP SAPs* and *non-ErrP SAPs* respectively. We use Levene's test [43] to conclude that the difference in variance between these two population samples is statistically significant ($p = 0.023 < 0.05$).

- **Experiment 3: Subjectivity over states.** For each unique state in Maze game, we plot the mean and standard deviation of EEG trial performance in Fig. 9. We plot the classfier accuracy for ErrP SAPs and non-ErrP SAPs respectively based on their initial state on the maze. We can visualize that the plot corresponding the standard deviation for non-ErrP SAPs is darker (indicating lower standard deviation) compared to the plot corresponding

to the deviations for ErrP SAPs. We can also see that within a plot, there is also a gradation in the accuracy (indicated by different shades of green) implying that there is some dissimilarity among erroneous states and hence subjectivity on the user's part and diminishing the argument that erroneous vs non-erroneous scenarios are purely binary.

- **Experiment 4: Errors of commission and omission.** In this experiment, we consider only the erroneous actions (ErrP SAPs) and split the EEG trials into two categories, (i) commission errors and (ii) omission errors. We do this in order to better understand the impact on ErrPs, based on the type of error committed. A commission error is defined as an agent making an incorrect move to a new cell, while omission error refers to the incorrect action of agent by staying in the same cell grid. The total state-actions pairs for commission and omission are distributed fairly (out of 71 unique state-action pairs, 34 correspond to errors of omission and the remaining 37 correspond to errors of commission). However we observe that among the state-action pairs which had very high accuracies, state-action pairs corresponding to errors of commission are disproportionately represented. Out of the top 5 state-action pairs that have the highest accuracy, all of them represent errors of commission and out of the top 10 state-action pairs that have the highest accuracy, 9 of them signify errors of commission. This was also indicated by the fact that errors of omission had a mean accuracy of 72% whereas errors of commission had a much higher mean accuracy of 77%. This implies that the error scenarios that are the easiest to detect are likely to be errors of commission. This has certain implications that bolster the hypothesis that certain errors are indeed more "valuable" to a user than others and hence generate a far more noticeable response in the brain.

These 4 experiments collectively lead us to 2 main insights.

(a) Per subject, owing to the differences in variances, there is less variation in the non-ErrP accuracies compared to the ErrP accuracies implying that erroneous scenarios lead to more variation in the classifier accuracy and by extension, in the brain's response, than non-erroneous scenarios. This further implies that there is a gradation in error detection unlike it being a binary phenomenon which makes certain errors easier to detect and certain others more difficult to detect.

(b) The differences in variations in classifier accuracy between ErrP and non-ErrP SAPs diminishes when we average the accuracies over the SAPs and represent them as a function of users. This implies that the variation in the accuracy of ErrP vs non-ErrP is impacted more by differences in SAPs compared to the differences in users.

4) Robustness Evaluation: Because the generation process and decoding of brain signal are stochastic, the robustness to wrong ErrP labels is important when incorporating human feedback (via EEG) into reward shaping method. We are going to show that modeling the human policy as soft Q policy, as we
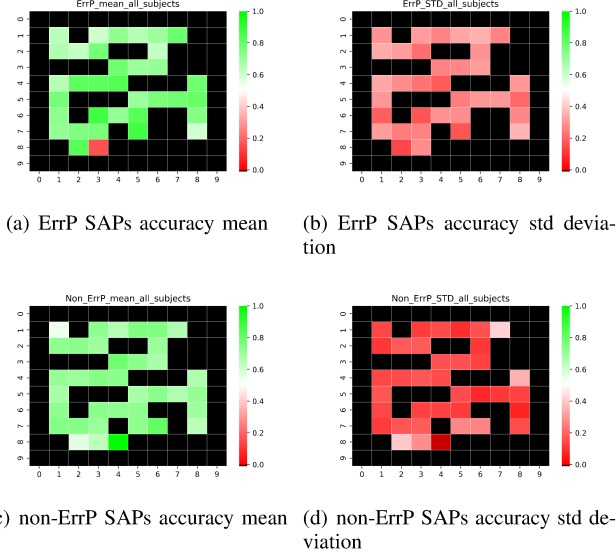
(a) ErrP SAPs accuracy mean

(b) ErrP SAPs accuracy std deviation

(c) non-ErrP SAPs accuracy mean

(d) non-ErrP SAPs accuracy std deviation

Fig. 9: Differences between ErrP and non-ErrP accuracies for each initial state over all users

did in (1), can make the learned auxiliary function r_a resist to wrong human feedback. In the comparison on robustness, the baseline method, called "*simple*", is to simply use a bootstrap neural network to generalize the binary ErrP labels across the state space, same as [67]. Both *simple* benchmark and the proposed robust reward shaping are trained on the same set of trajectories and human labels. The neural network in both methods is MLP, having two hidden layers of 64 units. And the number of bootstrap head in "simple" benchmark is set to 5. We evaluate both *simple* and the proposed methods on subject 02 and subject 07, whose accuracy are 0.71 and 0.78 respectively. The comparison result is shown in Figure 10. We can see that the proposed method performs better in both subjects with different initial trajectories. That is because the proposed method treats the human feedback in a probabilistic way, and the baseline function t can incorporate the state transition information to attenuate the influence of wrong human feedback. Moreover, the comparison of all cases shows that the performance gain of *simple* benchmark over no-ErrP method is decreased when the error probability of human label increases.

5) Ablation Study: In this section, we conduct ablation study on the proposed robust reward shaping with human feedback. We first specifically evaluate the effect of baseline function t learned from (4). Because the human feedback labels in the initial trajectories cannot cover the whole state space and some labels are wrong, the learned Q function of human $Q_h(\cdot, \cdot)$ may not be compatible with the state dynamics of the environment. Thus we introduce a baseline function only in terms of state to smoothen the learned Q function. Here the ablation evaluation on baseline function is still on subject 02 and 07, same as the section above. We realize $Q_h(\cdot, \cdot)$ and $t(\cdot)$ functions both by a two-layer MLP with 64 hidden units on each hidden layer and ReLU activation. We empirically find that it is enough to use 20 trajectories to train function $t(\cdot)$. The baseline is the proposed reward shaping method without t
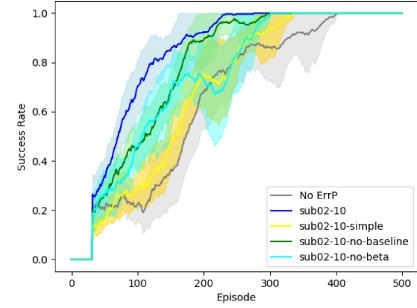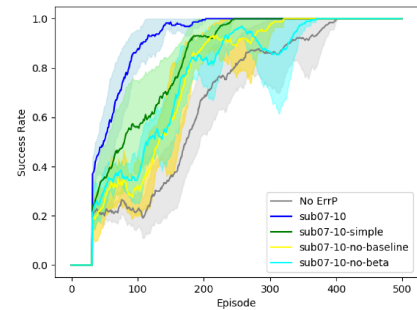

(a) Subject 02-10 Trajectories
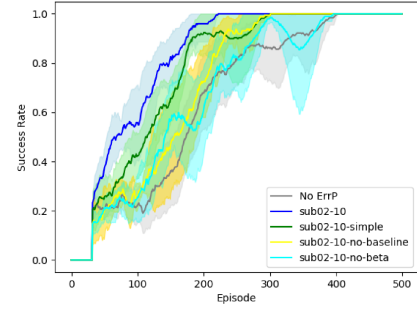
(b) Subject 02-20 Trajectories
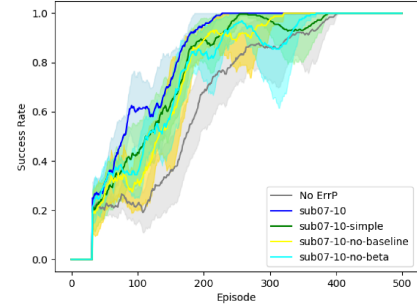
(c) Subject 07-10 Trajectories

(d) Subject 07-20 Trajectories

Fig. 10: Robustness and Ablation Study of the proposed reward shaping method. Simple: modeling human feedback simply by a neural network. No-baseline: method without function $t_\phi(\cdot)$. No-beta: method without combining coefficient $\beta(\cdot)$

in (2), corresponding to *no-baseline* curves in Figure 10, where the auxiliary reward function is only the Bellman difference of Q function between adjacent states. The comparison result is shown in Figure 10. We can see that the baseline function can improve the convergence speed in all cases, and it can even do better than *simple* method in some cases, showing the importance of the baseline function here.

In addition, we also conduct the ablation study on the combining coefficient $\beta(\cdot)$. The benchmark method is to directly sum auxiliary reward r_a and environmental reward r_e together. The coefficient in the proposed method is set to $\beta(e) = 3e^{-e/80}$. Comparison result is shown in Figure 10 where *no-beta* curves corresponding benchmarks on combining coefficients. We can see this exponentially decreasing coefficient can stabilize the training process significantly, and hence improve the convergence speed.

VIII. CONCLUSIONS AND FUTURE WORK

In this work, we investigated an interesting paradigm to obtain and integrate the implicit human feedback with RL algorithms. We first demonstrated the feasibility of obtaining implicit human feedback by capturing error-potentials of a human observer watching an agent learning to play several different visual-based games, and then decoding the signals appropriately and using them as an auxiliary reward function to help an RL agent. Then we argued that the definition of ErrPs could be learned in a zero-shot manner across different environments, eliminating the need of re-training over new and unseen environments. We validated the acceleration in learning of games through augmenting the RL agent by ErrP feedback using a naive approach, i.e., *full access* method. We then proposed a novel RL framework, improving the label efficiency and reducing human cognitive load. We experimentally showed that the proposed RL framework could accelerate the training of RL agent by 2.25x, while reducing the number of queries required by 75.56%.

REFERENCES

[1] Mohit Agarwal and Raghupathy Sivakumar. Think: Toward practical general-purpose brain-computer communication. In *HotWireless '15*, 2015.

[2] Mohit Agarwal and Raghupathy Sivakumar. Cerebro: A wearable solution to detect and track user preferences using brainwaves. In *The 5th ACM Workshop on Wearable Systems and Applications*, WearSys '19, page 47–52, New York, NY, USA, 2019.

[3] Mohit Agarwal and Raghupathy Sivakumar. Charge for a whole day: Extending battery life for bci wearables using a lightweight wake-up command. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020.

[4] Mohit Agarwal, Shyam Krishnan Venkateswaran, and Raghupathy Sivakumar. Human-in-the-loop rl with an eeg wearable headset: On effective use of brainwaves to accelerate learning. In *Proceedings of the 6th ACM Workshop on Wearable Systems and Applications*, WearSys '20, page 25–30, New York, NY, USA, 2020. Association for Computing Machinery.

[5] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shinichi Maeda. Dqn-tamer: Human-in-the-loop reinforcement learning with intractable feedback. *arXiv preprint arXiv:1810.11748*, 2018.

[6] Kamyar Azizzadenesheli, Emma Brunskill, and Animashree Anandkumar. Efficient exploration through bayesian deep q-networks. In *2018 Information Theory and Applications Workshop (ITA)*, pages 1–9. IEEE, 2018.

[7] Alexandre Barachant and Stephane Bonnet. Channel selection procedure using riemannian distance for bci applications. In *2011 5th International IEEE/EMBS Conference on Neural Engineering*, pages 348–351. IEEE, 2011.

[8] Alexandre Barachant, Stéphane Bonnet, Marco Congedo, and Christian Jutten. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomputing*, 112:172–178, 2013.

[9] Alexandre Barachant and Marco Congedo. A plug&play p300 bci using information geometry. *arXiv preprint arXiv:1409.0107*, 2014.

[10] Shlomo Bentin, Gregory McCarthy, and Charles C Wood. Event-related potentials, lexical decision and semantic priming. *Electroencephalography and clinical Neurophysiology*, 60(4):343–355, 1985.

[11] Benjamin Blankertz, Guido Dornhege, Christin Schafer, Roman Krepki, Jens Kohlmorgen, K-R Muller, Volker Kunzmann, Florian Losch, and Gabriel Curio. Boosting bit rates and error detection for the classification of fast-paced motor commands based on single-trial eeg analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 11(2):127–131, 2003.

[12] Daniel S Brown, Wonjoon Goo, Prabhat Nagarajan, and Scott Niekum. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. *arXiv preprint arXiv:1904.06387*, 2019.

[13] Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of monte carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in games*, 4(1):1–43, 2012.

[14] Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. Reinforcement learning from demonstration through shaping. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[15] Cameron S Carter, Todd S Braver, Deanna M Barch, Matthew M Botvinick, Douglas Noll, and Jonathan D Cohen. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science*, 280(5364):747–749, 1998.

[16] Hyeong Soo Chang. Reinforcement learning with supervision by combining multiple learnings and expert advices. In *2006 American Control Conference*, pages 6–pp. IEEE, 2006.

[17] LI Charles and RS Christian. A social reinforcement learning agent. In *Proceedings of the fifth international conference on Autonomous agents*, 2001.

[18] Ricardo Chavarriaga and José del R Millán. Learning from eeg error-related potentials in noninvasive brain-computer interfaces. *IEEE transactions on neural systems and rehabilitation engineering*, 18(4):381–388, 2010.

[19] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[20] Marco Congedo, Alexandre Barachant, and Anton Andreev. A new generation of brain-computer interface based on riemannian geometry. *arXiv preprint arXiv:1310.8115*, 2013.

[21] Christian Daniel, Oliver Kroemer, Malte Viering, Jan Metz, and Jan Peters. Active reward learning with a novel acquisition function. *Autonomous Robots*, 39(3):389–405, 2015.

[22] Layla El Asri, Bilal Piot, Matthieu Geist, Romain Laroche, and Olivier Pietquin. Score-based inverse reinforcement learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pages 457–465. International Foundation for Autonomous Agents and Multiagent Systems, 2016.

[23] Michael Falkenstein, Joachim Hohnsbein, Joerg Hoormann, and L. Blanke. Effects of crossmodal divided attention on late erp components. ii. error processing in choice reaction tasks. *Electroencephalography and clinical neurophysiology*, 78 6:447–55, 1991.

[24] Michael Falkenstein, Jörg Hoormann, Stefan Christ, and Joachim Hohnsbein. Erp components on reaction errors and their functional significance: a tutorial. *Biological psychology*, 51(2-3):87–107, 2000.

[25] Michael Falkenstein, Jörg Hoormann, Stefan Christ, and Joachim Hohnsbein. Erp components on reaction errors and their functional significance: A tutorial. *Biological Psychology*, 51:87–107, 02 2000.

[26] Pierre W Ferrez and José del R Millán. You are wrong!—automatic detection of interaction errors from brain waves. In *Proceedings of the 19th international joint conference on Artificial intelligence*, number CONF, 2005.

[27] Pierre W Ferrez and José del R Millán. Error-related eeg potentials generated during simulated brain–computer interaction. *IEEE transactions on biomedical engineering*, 55(3):923–929, 2008.

[28] Jonathan R Folstein and Cyma Van Petten. Influence of cognitive control and mismatch on the n2 component of the erp: A review. *Psychophysiology*, 2008.

[29] William Gehring, Michael Coles, David Meyer, and Emanuel Donchin. A brain potential manifestation of error-related processing [supplement]. *Electroencephalography and clinical neurophysiology. Supplement*, 44:261–72, 02 1995.

[30] H Gemba, K Sasaki, and V.B. Brooks. 'error' potentials in limbic cortex (anterior cingulate area 24) of monkeys during motor learning. *Neuroscience letters*, 70:223–7, 11 1986.

[31] E. Gupta, M. Agarwal, and R. Sivakumar. Blink to get in: Biometric authentication for mobile devices using eeg signals. In *ICC 2020 - 2020 IEEE International Conference on Communications (ICC)*, pages 1–6, 2020.

[32] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1352–1361. JMLR. org, 2017.

[33] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290*, 2018.

[34] Clay B. Holroyd and Michael G. H. Coles. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109 4:679–709, 2002.

[35] 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.

[36] Clay B Holroyd, Sander Nieuwenhuis, Nick Yeung, and Jonathan D Cohen. Errors in reward prediction are reflected in the event-related brain potential. *Neuroreport*, 14(18):2481–2484, 2003.

[37] Iñaki Iturrate, Luis Montesano, and Javier Minguez. Robot reinforcement learning using eeg-based reward signals. In *2010 IEEE International Conference on Robotics and Automation*, pages 4822–4829. IEEE, 2010.

[38] 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.

[39] W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16. ACM, 2009.

[40] W Bradley Knox and Peter Stone. Augmenting reinforcement learning with human feedback. In *ICML 2011 Workshop on New Developments in Imitation Learning (July 2011)*, volume 855, page 3, 2011.

[41] William Bradley Knox. Learning from human-generated reward. 2012.

[42] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *European conference on machine learning*, pages 282–293. Springer, 2006.

[43] Howard Levene, II Olkin, and H Hotelling. Robust tests for equality of variances.(1960). *Contributions to probability and statistics; Essays in honor of Harold Hotelling*, pages 78–92, 1960.

[44] Mandy J Maguire, Grant Magnon, Diane A Ogiela, Rebecca Egbert, and Lynda Sides. The N300 ERP component reveals developmental changes in object and action identification. *Developmental Cognitive Neuroscience*, 5:1–9, 2013.

[45] Wolfgang H. R. Miltner, Christoph H. Braun, and Michael G. H. Coles. Event-related brain potentials following incorrect feedback in a time-estimation task: Evidence for a "generic" neural system for error detection. *J. Cognitive Neuroscience*, 9(6):788–798, November 1997.

[46] Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, pages 278–287, 1999.

[47] Sander Nieuwenhuis, K Richard Ridderinkhof, Jos Blom, Guido PH Band, and Albert Kok. Error-related brain potentials are differentially related to awareness of response errors: Evidence from an antisaccade task. *Psychophysiology*, 38(5):752–760, 2001.

[48] Hiroaki Niki and Masataka Watanabe. Prefrontal and cingulate activity during timing behavior in the monkey. *Brain research*, 171:213–24, 09 1979.

[49] Lee Osterhout, Mark D. Allen, Judith Mclaughlin, and Kayo Inoue. Brain potentials elicited by prose-embedded linguistic anomalies. *Memory & Cognition*, 30(8):1304–1312, Dec 2002.

[50] Lucas C Parra, Clay D Spence, Adam D Gerson, and Paul Sajda. Response error correction-a demonstration of improved human-machine performance using real-time eeg monitoring. *IEEE transactions on neural systems and rehabilitation engineering*, 11(2):173–177, 2003.

[51] Filip Radlinski and Thorsten Joachims. Evaluating the robustness of learning from implicit feedback. *arXiv preprint cs/0605036*, 2006.

[52] Yann Renard, Fabien Lotte, Guillaume Gibert, Marco Congedo, Emmanuel Maby, Vincent Delannoy, Olivier Bertrand, and Anatole Lécuyer. Openvibe: An open-source software platform to design, test, and use brain–computer interfaces in real and virtual environments. *Presence: teleoperators and virtual environments*, 19(1):35–53, 2010.

[53] Bertrand Rivet, Antoine Souloumiac, Virginie Attina, and Guillaume Gibert. xdawn algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering*, 56(8):2035–2043, 2009.

[54] 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.

[55] Gerwin Schalk, Jonathan R Wolpaw, Dennis J McFarland, and Gert Pfurtscheller. Eeg-based communication: presence of an error potential. *Clinical neurophysiology*, 111(12):2138–2144, 2000.

[56] Marten K Scheffers, Michael GH Coles, Peter Bernstein, William J Gehring, and Emanuel Donchin. Event-related brain potentials and error-related processing: An analysis of incorrect responses to go and no-go stimuli. *Psychophysiology*, 33(1):42–53, 1996.

[57] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Advances in neural information processing systems*, pages 2164–2172, 2010.

[58] Hillyard SA. Squires NK, Squires KC. Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalogr Clin Neurophysiol*, 1975.

[59] Matthew E Taylor, Halit Bener Suay, and Sonia Chernova. Integrating reinforcement learning with human demonstrations of varying ability. In *The 10th International Conference on Autonomous Agents and Multiagent Systems-Volume 2*, pages 617–624. International Foundation for Autonomous Agents and Multiagent Systems, 2011.

[60] Georgia tech robotarium lab. http://www.news.gatech.edu/features/robotarium-robotics-lab-accessible-all.

[61] Andrea Lockerd Thomaz, Cynthia Breazeal, et al. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. In *Aaai*, volume 6, pages 1000–1005. Boston, MA, 2006.

[62] Sida I Wang, Percy Liang, and Christopher D Manning. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*, 2016.

[63] Zhaodong Wang and Matthew E Taylor. Improving reinforcement learning with confidence-based demonstrations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3027–3033, 2017.

[64] Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[65] Eric Wiewiora, Garrison W Cottrell, and Charles Elkan. Principled methods for advising reinforcement learning agents. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 792–799, 2003.

[66] Christian Wirth, Johannes Fürnkranz, and Gerhard Neumann. Model-free preference-based reinforcement learning. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[67] Baicen Xiao, Qifan Lu, Bhaskar Ramasubramanian, Andrew Clark, Linda Bushnell, and Radha Poovendran. Fresh: Interactive reward shaping in high-dimensional state spaces using human feedback. *arXiv preprint arXiv:2001.06781*, 2020.

[68] Baicen Xiao, Bhaskar Ramasubramanian, Andrew Clark, Hannaneh Hajishirzi, Linda Bushnell, and Radha Poovendran. Potential-based advice for stochastic policy learning. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 1842–1849. IEEE, 2019.

[69] Ruohan Zhang, Faraz Torabi, Lin Guan, Dana H Ballard, and Peter Stone. Leveraging human guidance for deep reinforcement learning tasks. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pages 6339–6346. AAAI Press, 2019.

[70] Brian D Ziebart. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. PhD thesis, figshare, 2010.