# Playing Games with Implicit Human Feedback

**Duo Xu***, **Mohit Agarwal***, **Faramarz Fekri and Raghupathy Sivakumar**
Georgia Institute of Technology
{dxu301, mohit, fekri, siva}@ece.gatech.edu

## Abstract

We consider the following central question in the field of Deep Reinforcement Learning (DRL): *How can we use implicit human feedback to accelerate and optimize the training of a DRL algorithm?* State-of-the-art methods rely on any human feedback to be provided explicitly, requiring the active participation of humans (e.g., expert labeling, demonstrations, etc.). In this work, we investigate an alternative paradigm, where non-expert humans are silently observing (and assessing) the agent interacting with the environment. The human's intrinsic reactions to the agent's behavior is sensed as implicit feedback by placing electrodes on the human scalp and monitoring what are known as event-related electric potentials. The implicit feedback is then used to augment the agent's learning in the RL tasks. We develop a system to obtain and accurately decode the implicit human feedback, specifically error-related event potentials (ErrP), for state-action pairs in an Atari-type environment. As a baseline contribution, we demonstrate the feasibility of capturing error-potentials of a human observer watching an agent learning to play several different Atari-games using an electroencephalogram (EEG) cap, and then decoding the signals appropriately and using them as an auxiliary reward function to a DRL algorithm with the intent of accelerating its learning of the game. Building atop the baseline, we then make the following novel contributions in our work: (i) We argue that the definition of ErrP is *generalizable* across different environments; specifically we show that ErrP of an observer can be learned for a specific game, and the definition used as-is for another game without requiring re-learning of the error-potentials. (ii) In order to improve ErrP data efficiency, we propose a new learning framework to combine recent advances in DRL into the ErrP-based feedback system, allowing humans to provide implicit feedback only prior to the start of RL agent training. (iii) Finally, we scale the implicit human feedback (via ErrP) based RL to reasonably complex environments (games) and demonstrate the significance of our approach through synthetic and real user experiments.

## Introduction

Deep Reinforcement Learning (DRL) algorithms have now beaten human experts in Go (Silver et al. 2017), taught robots to become parkour masters (Heess et al. 2017), and enabled truly autonomous vehicles (Wang, Jia, and Weng 2018). However, current state-of-the-art RL agents equipped with deep neural networks are inherently complex, difficult and time-intensive to train. Particularly in complex environments with sparse reward functions (e.g., maze navigation), the DRL agents need an inordinate amount of interaction with the environment to learn the optimal policy. Human participation can potentially help DRL algorithms by accelerating their training and reducing the learning costs without compromising final performance. This potential has inspired a several research efforts where either an alternative (or supplementary) feedback is obtained from the human participant (Knox 2012). Such approaches despite being highly effective, severely burden the human-in-the-loop demanding either expert demonstrations (Ross, Gordon, and Bagnell 2011) or explicit feedback (Christiano et al. 2017).

In this paper, we investigate an alternative paradigm 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 reward functions that can be used by the DRL algorithms. 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 if 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/negativity (ErrP or ERN) (Scheffers et al. 1996).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.

As a baseline contribution, we demonstrate the feasibility of capturing error-potentials of a human observer watching an agent learning to play several different Atari-games, and then decoding the signals appropriately and using them as an auxiliary reward function to a DRL algorithm. We show that a *full access* approach, inquiring human feedback on every state-action pair visited by RL agent, can significantly speedup the training convergence of DRL algorithm. We contend that while obtaining such implicit human feedback through EEG is less burdensome, it is still a time-intensive

*equal contribution

task for the subject and the experimenter alike. This, combined with the noisy EEG signals and stochasticity in inferring error-potentials, raises significant challenges in terms of the practicality of the solution.

In this context, we first argue that the definition of ErrPs is generalizable across different environments. We show that ErrPs of an observer can be learned for a specific game, and the definition used as-is for another game without requiring re-learning of the ErrP. This is notably different from previous approaches (Chavarriaga and Millán 2010; Salazar-Gomez et al. 2017), where the labeled ErrPs are obtained in the same environment (where the RL task is performed). For any new and unseen environment, it does not require the human to go through the training phase again, and assumes no prior knowledge about the optimal state-action pairs of the environment.

We present a framework to combine recent advances in DRL into the implicit human feedback mechanism (via ErrP) in a practical, sample-efficient manner. This reduces the cost of human supervision sufficiently allowing the DRL systems to train. Our proposed framework allows humans to provide their feedback implicitly before the agent starts training. Based on the human feedback obtained during pre-training, a quality (Q) function is learned over these imperfect demonstrations to provide the supplementary reward to the RL agent. We present results from real ErrP experiments to evaluate the acceleration in learning, and sample efficiency, in the proposed frameworks. In summary, the novel contributions of our work are,

1. We demonstrate the generalizability of error-potentials over various Atari-like environments (discrete grid-based navigation games, studied in this work), enabling the estimation of implicit human feedback in new and unseen environments.
2. We propose a learning framework to combine recent advances in DRL into ErrP based feedback system in a practical, sample-efficient manner. Taking advantage of recent approaches in learning from imperfect demonstrations, we only need human ErrP labels on the demonstrations given initially, reducing the number of inquiring human feedback without performance degradation.
3. We scale the implicit human feedback (via ErrP) based RL to reasonably complex environments and demonstrate the significance of our approach through synthetic and real user experiments.

## Related Work

(Daniel et al. 2015; El Asri et al. 2016; Wang, Liang, and Manning 2016) studied RL from human rankings or ratings, however rely on explicit human feedback, and assume that the feedback is noiseless. Demonstrations have been commonly used to improve the efficiency of RL (Kim et al. 2013; Chemali and Lazaric 2015; Piot, Geist, and Pietquin 2014), and a common paradigm is to initialize RL algorithms with good policy or Q function (Nair et al. 2018; Hester et al. 2018; Gao et al. 2018). In this work, we use rely on implicit feedback from non-expert humans (via ErrPs) which is inherently noisy.

(Chavarriaga and Millán 2010; Iturrate, Montesano, and Minguez 2010; Salazar-Gomez et al. 2017) 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 combine with the recent advances in DRL in a sample-efficient manner for reasonably complex environments.

## Preliminaries and Setup

**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 $0 \leq \gamma \leq 1$. Here the random variable $Z(s,a)$ denotes the accumulated discounted future rewards starting from state $s$ and action $a$. 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., $\mathcal{L}(Q) = E_\pi\big[\big(Q(x,a) - r - \gamma Q(x',\hat{a})\big)^2\big]$, and temporal difference (TD) (Tesauro 1995) update where the transition tuple $(x,a,r,x')$ consists of a consecutive experience under behavior policy $\pi$. Modern techniques in DRL such as DQN (Mnih et al. 2015) and the target network (Van Hasselt, Guez, and Silver 2016) are also adopted here.

**Bayesian Deep Q Network** We introduce the DQN model adopted in this paper. *Bayesian DQN* 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 $\phi_\theta(x) \in R^d$, parameterized by neural network with weights $\theta$ (Osband, Russo, and Van Roy 2013). Here by utilizing the DQN architecture and imposing Gaussian distributions over $\omega_a$, the Bayesian linear regression (BLR) (Rasmussen 2003) can give us the posterior of $\omega_a$ as below

$$\omega_a \sim \mathcal{N}(\bar{\omega}_a, Cov_a), \;\; \bar{\omega}_a := \frac{1}{\sigma_\epsilon^2} Cov_a \Phi_a^\theta y_a,$$
$$Cov_a := \left(\frac{1}{\sigma_\epsilon^2}\Phi_a^\theta \Phi_a^{\theta T} + \frac{1}{\sigma^2} I\right)^{-1}, a \in \mathcal{A} \qquad (1)$$

where we construct disjoint replay buffer $\mathcal{D}_a$ corresponding to experience with action $a$, and a matrix $\Phi_a^\theta \in R^{d \times |\mathcal{D}_a|}$, vector $y_a$, i.e., the concatenation of state features and target values in set $\mathcal{D}_a$. Therefore the posterior of Q value can be the Gaussian distribution as below,

$$Q(x,a) \sim \mathcal{N}(\bar{\omega}_a^T \phi_\theta(x), \;\; \phi_\theta(x)^T Cov_a \phi_\theta(x)) \qquad (2)$$

## System Setup and Data Collection

We consider a setup where a non-expert human is silently observing (and assessing) a computer agent (driven by RL) interacting with an environment. The human's intrinsic reactions to the agent's behavior is sensed as implicit feedback by placing electrodes on the human scalp and monitoring what are known as event-related potentials (ErrPs).

The implicit feedback is then used to augment the agent's learning in the RL tasks. We develop a system to obtain and accurately decode the implicit human feedback (specifically error-related event potentials) for state-action pairs in an Atari-type environment.

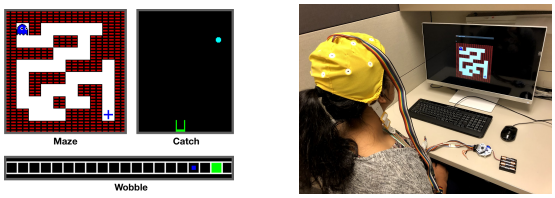
**Game Environments** We have developed three discrete-grid based navigation games in OpenAI Gym emulating *Atari* framework (Brockman et al. 2016), namely (i) Wobble, (ii) Catch, and (iii) Maze, shown in Fig. 1(a). We use the default Atari dimensions (i.e., 210x160 pixels). The source codes of the games can be found in the public repository[1], and can be used with the OpenAI Gym module.

**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 for the agent is moving one step either 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 screen dimensions are divided into 10x10 grid space, where the *egg* and the *cart*, both occupies one block. The action space of the agent consists of "NOOP" (no operation), "moving left" and "moving 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 to a fixed target. The Atari screen is centered and divided into 10x10 equal-sized blocks. The agent and target occupy one block. The action space consists of four directional movements. The maze architecture is kept fixed for the purpose of this work. If an agent moves, but hits a wall, a quick blinking of the agent is displayed, to show the action taken by the agent.

**EEG experimental protocol:** 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. The electrode cap was attached with the OpenBCI[3] 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 (Renard et al. 2010) 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. A total of five human subjects were recruited using standard procedures. We recruited five human subjects (mean age $26.8 \pm 1.92$, 1 female) for collecting the EEG data. For each subject, we conducted three separate sessions over multiple days. For each subject-game pair, the experimental duration was less than 15 minutes. The agent took action every 1.5 seconds. All the research protocols for the user data collection were reviewed and approved by the Georgia Tech Institutional Review Board.




(a) Game Environments (b) Experiment Bench

Figure 1: Experimental framework

## Integrating DRL 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 observer watching an agent learning to play several different Atari-games using EEG, and then decoding the signals appropriately, and (ii) using them as an auxiliary reward function to a DRL algorithm with the intent of accelerating its learning of the game.

### Obtaining the Implicit Human Feedback: Decoding ErrPs

We rely on the Riemannian Geometry framework for the classification of human's intrinsic reaction (captured in the form of ErrPs) (Barachant and Congedo 2014; Congedo, Barachant, and Andreev 2013).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). The Riemannian Geometry based framework was first proposed in (Barachant and Congedo 2014; Congedo, Barachant, and Andreev 2013) to translate the raw EEG signals into meaningful labels[4]. The raw EEG data is bandpass filtered in [0.5, 40] Hz. Epochs of 800ms were extracted relative to pre-stimulus 200ms baseline, and were subjected to spatial filtering. In spatial filtering, prototype responses of each class, i.e., "correct" and "erroneous", are computed by averaging all training trials in the corresponding classes("xDAWN Spatial Filter" (Rivet et al. 2009; Barachant and Congedo 2014; Congedo, Barachant, and Andreev 2013)). "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). The covariance matrix of each epoch is computed, and concatenated with the prototype responses of the class. Further, dimensionality reduction is achieved by selecting relevant channels through

[1]https://github.com/meagmohit/gym-maze

[2]https://en.wikipedia.org/wiki/Eggomania

[3]http://openbci.com

[4]The authors successfully applied the framework and won multiple Kaggle challenges. E.g., https://www.kaggle.com/c/inria-bci-challenge. Later, this framework was successfully adapted in many other error-potential decoding works (Salazar-Gomez et al. 2017).

backward elimination (Barachant and Bonnet 2011). The filtered signals are projected to the tangent space (Barachant et al. 2013; 2011) for feature extraction. The obtained feature vector is first normalized (using L1 norm) and fed to a regularized regression model. A threshold value is selected for the final decision by maximizing accuracy offline on the training set. We present the algorithm to decode the ErrP signals in Algorithm 1.

**Algorithm 1:** Riemannian Geometry based ErrP classification algorithm (Barachant et al. 2013)

**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

### The Full Access Method

With the availability of implicit human feedback, we explore how the training of state-of-the-art DRL algorithms can be accelerated. A naive approach is to obtain feedback on every state-action pair while RL agent is learning (also known as *full access*). It is to add a negative penalty to the reward whenever ErrP is detected, and keep using the original reward from the environment whenever ErrP is not detected. The evaluation result of this method based on real ErrP data are presented later in the evaluation section, validating that this method can speed up the training convergence of RL agent significantly. However, obtaining the human feedback for every state-action pair is time-intensive and undesirable from a practical point of view. Ideally, an approach is desirable where the human feedback is obtained only for state-action pairs for which the learning agent has highest uncertainty.

## Towards Practical Integration of DRL with Implicit Human Feedback

In this section, we propose two approaches towards integrating human implicit feedback with recent advances in DRL and make the ErrP-agumented RL deployable into practical system. Firstly, we show that ErrPs of an observer can be learned for a specific game, and the definition used as-is for another game without requiring re-learning of the ErrP. Further, in order to utilize ErrP data more efficiently, we propose a RL framework to combine the recent advances in imitation learning into the implicit human feedback mechanism (via ErrP) to accelerate the RL agent learning. Specifically, we first obtain the implicit human feedback before the training of the RL agent. It exploits the initially given trajectories criticized by ErrP labels and learn a reward function for augmenting the following RL agent, where human with some prior knowledge is needed to specify some perturbed expert trajectories. Recently, Q function is shown to have better generalization in state-space if trained expert demonstrations perturbed by some noise (Laskey et al. 2017; Luo, Xu, and Ma 2019).

### ErrP Generalization across Environments

Error-potentials in the EEG signals is studied under two major paradigms in human-machine interaction tasks, (i) *feedback and response ErrPs:* error made by human (Carter et al. 1998; Falkenstein et al. 2000; Blankertz et al. 2003; Parra et al. 2003; Holroyd and Coles 2002), (ii) *interaction ErrPs:* error made by machine in interpreting human intent (Ferrez and Millán 2005). Another interesting paradigm is when human is watching (and silently assessing) the machine performing a specific task (Chavarriaga and Millán 2010). The manifestation of these potentials across these paradigms were found quite similar in terms of their general shape, timings of negative and positive peaks, frequency characteristics etc., (Ferrez and Millán 2005; Chavarriaga and Millán 2010). This prompts us to explore the consistency of the error-potentials across different environments (i.e., games, in our case). We restrict the score of our work to the paradigm of human acting as a silent observer of the machine actions. In Fig.2, 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 timings of the peaks is quite consistent across the three game environments studied in this work. Further, in experimental evaluation section, we show that error-potentials are indeed generalizable across environments, and can further be used to inform deep reinforcement learning algorithm in a new and unseen environments.

### Proposed Framework: Learning from Imperfect Demonstrations with Human ErrP

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. Imitation learning from a small number of demonstrations followed by RL fine-tuning is a promising paradigm to improve the sample efficiency in such cases (Večerík et al. 2017; Hester et al. 2018; Gao et al. 2018). Inspired by the paradigm of imitation learning, we develop a novel framework that can robustly learn a reward function to augment the DRL algorithms and accelerate the training of the RL agent. This reward function is derived from reward learning with imperfect demonstrations and human critique, inquiring the human feedback in the form of ErrP over a set of trajectories.

The flowchart of the proposed learning framework is shown in Fig. 3. In this framework, trajectories in the demonstration are first criticized by human ErrP in ErrP experiments, and then with ErrP labels decoded from the decoder, in the reward learning step, a quality (Q) function is learned from the trajectories where the correctness of state-action pairs is given by ErrP labeling. An alternative reward can be derived from the learned quality function, to augment the following RL agent. Here we only make queries for ErrP

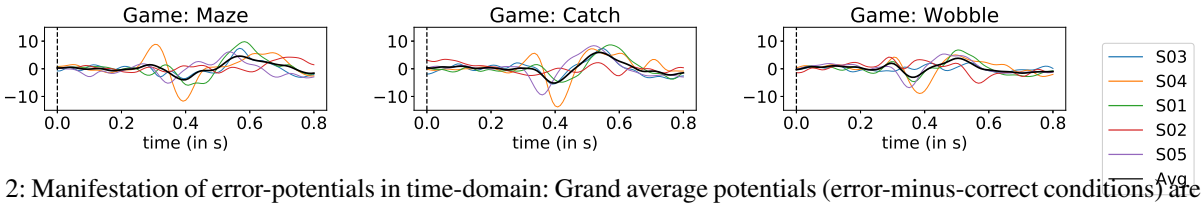



Figure 2: Manifestation of error-potentials in time-domain: Grand average potentials (error-minus-correct conditions) are shown for Maze, Catch and Wobble game environments. Thick black line denotes the average over all the subjects.

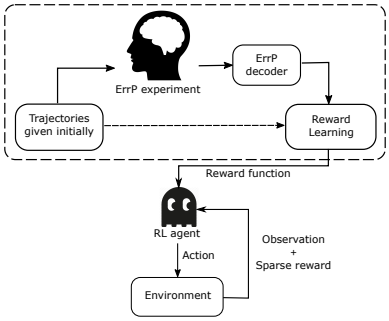



Figure 3: Learning from imperfect demonstration with implicit human feedback. The dashed arrow shows trajectories in $\mathcal{D} \cup \mathcal{D}_R$ are all used in reward learning

labeling on trajectories given initially, instead of inquiring in every training step in the *full access* method. So the number of ErrP queries needed can be reduced significantly here. These queries are made before the RL agent starts training, improving the efficiency of labeling (implicit, ErrP based) queries made to the external oracle (human). Constraint by the coherence requirement in EEG experiments, the demonstrations for ErrP labeling can only consist of complete trajectories. We assume that the trajectories in the demonstration are initially specified by human or other external algorithms, without any reward information. This is a reasonable assumption since the reward function may be unknown to humans in general cases. In ErrP experiment, the human subject provides implicit feedback (via ErrP) on state-action pairs along the trajectories, labeling every pair as a *positive* or *negative* sample corresponding to the correctness from human judgement. Based on the decoded ErrP labels and trajectories, the proposed framework learns the reward function based on maximum entropy RL methods (Ziebart 2010), which is explained with details in the following.

Different from regular imitation learning, here these trajectories are not exactly same as expert demonstrations. This is beneficial, because the Q function learned from imperfect trajectories can have better estimations on states unseen in the demonstration, providing better generalization in the state space (Laskey et al. 2017; Luo, Xu, and Ma 2019).

**Reward Learning** Since implicit human feedback via ErrP is noisy (hence *imperfect demonstrations*), we model the reward learning as a probabilistic maximum entropy RL problem. Following the principle of maximum entropy, given Q function $Q(\cdot,\cdot)$, the policy distribution and value function in terms of Q function can be expressed as follows,

$$V_Q(s) \quad = \quad \alpha \log \sum_a \exp(Q(s,a)/\alpha),$$

$$\pi_Q(a|s) \quad = \quad \exp((Q(s,a) - V_Q(s))/\alpha) \qquad (3)$$

where $\alpha$ is a free parameter, tuned empirically. The likelihood of positive and negative state-action pair are denoted as $\pi_Q(a|s)$ and $1 - \pi_Q(a|s)$. When demonstrations and corresponding implicit human feedback are ready, we train the Q function by maximizing the likelihood of both positive and negative state-action pairs in the demonstrations, which is to maximize the following objective,

$$\begin{aligned} J_1(Q) \quad := \quad & \sum_{(s,a)\in\mathcal{D}} \pi_Q(a|s)(1 - ErrP(s,a)) \\ & +(1 - \pi_Q(a|s))ErrP(s,a) \end{aligned} \qquad (4)$$

where the binary variable $ErrP(s,a)$ denotes the correctness of state-action pair from human feedback.

In order to refine the reward shape and attenuate the variance of learning updates, we introduce another baseline function $t(s)$. Hence, the Q function becomes $Q_B(s,a) := Q(s,a) - t(s)$. It can be proved that $Q_B(\cdot,\cdot)$ and $Q(\cdot,\cdot)$ induce the same optimal policy (Ng, Harada, and Russell 1999). The baseline function $t^*(\cdot)$ can be learned by optimizing $t^* = \arg\min_t J_2(t)$, and the objective is defined as

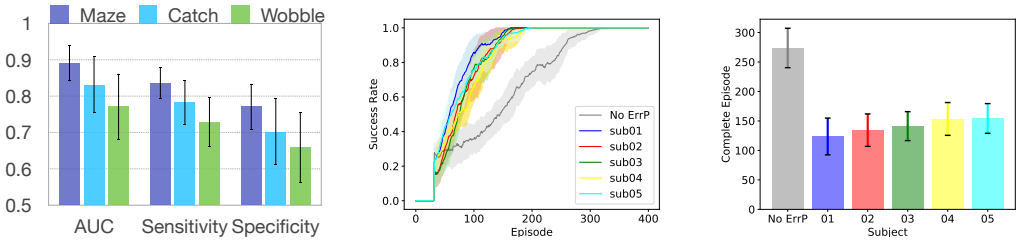
$$\begin{aligned} J_2(t) \quad := \quad & \sum_{(s,a,s')\in\mathcal{D}\cup\mathcal{D}_R} l(Q(s,a) - t(s) \\ & -\gamma \max_{a'\in\mathcal{A}}(Q(s',a') - t(s'))) \end{aligned} \qquad (5)$$

where the loss function $l(\cdot)$ is chosen to be $l_1$-norm via empirical evaluations. In addition to the demonstration $\mathcal{D}$, we incorporate another set of demonstrations $\mathcal{D}_R$, containing transitions randomly sampled from environment without reward information. The set $\mathcal{D}_R$ is to help the function $t(\cdot)$ to efficiently learn the state dynamics, and does not require any human labeling, essentially not increasing the labeling workload. After reward learning, consisting of learning Q function and baseline function, for any transition tuple $(s,a,s')$, the learned reward function can be represented as $Q_B(s,a) - \gamma \max_{a'\in\mathcal{A}} Q_B(s',a')$. We then use this reward function to augment the following RL agent.

## Experimental Results

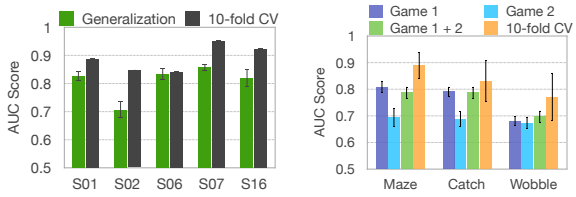
### Baseline results: Naive Approach

We first validate the feasibility of decoding ErrP signals using a 10-fold cross-validation scheme for each game. In this scheme, we train and test on the ErrP samples of the same game environment. In Fig. 4(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 ($0.89 \pm 0.05$) followed by Catch ($0.83 \pm 0.08$) and Wobble ($0.77 \pm 0.09$).




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

Figure 4: Baseline results of Naive Approach: (a) 10-fold CV performance of each game without any generalization, (b) and (c) RL with *full access* to ErrP feedback




(a) Over subjects (b) Over games

Figure 5: Generalizability 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.

The *full access* method as discussed above is the most preliminary approach to make ErrP labels augment the RL algorithm. It has the fastest training convergence rate (provides upper bound) but makes the maximum possible queries to the external oracle (human) for the implicit feedback. We use this method as a benchmark for comparing the data-efficiency of other RL augmentation methods. The results with real ErrP data of 5 subjects are shown in Figure 4. We can see there is a significant improvement in the training convergence. In this paper, "No ErrP" method refers to regular RL algorithms without the help of any human feedback. The success rate is defined as the ratio of success plays in the previous 32 episodes. The training completes when the success rate reaches to 1. In all plots of this paper, solid lines are average values over 10 random seeds, and shaded regions correspond to one standard deviation. In the evaluations of this paper, the Q network is modeled by the Bayesian deep Q network introduced above.

## Evaluation of practical solution

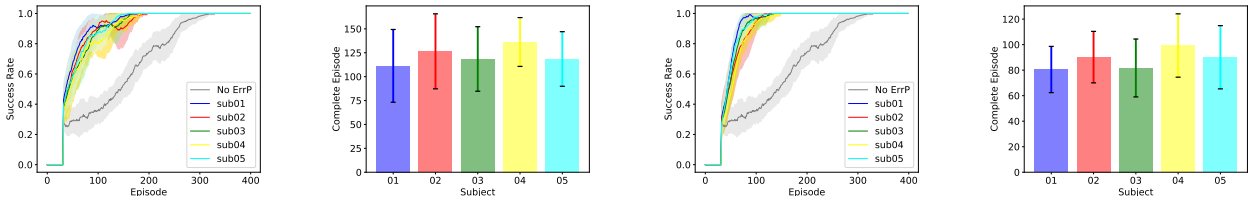
In this subsection, we evaluate the performance of three approaches to practially integrate the DRL with implicit human feedback (via ErrPs).

**Generalizability** To evaluate the generalization capability of error-potential signals and the decoding algorithm, we train on the samples collected from the Catch game and test on the Maze game. We assume that the information about state-action optimality is given for the Catch game, and thus labeled examples are obtained for the Catch game to train the ErrP decoder. However, the Maze game need to be solved, hence, we do not make any assumptions about the optimality of the actions. In Fig. 5(a), we provide the AUC score performance compared with the 10-fold Cross-Validation (CV) AUC score of Maze. We can see that the Catch game is able to capture more than 80% of the variability in the ErrPs for Maze game. To provide deeper insights into the generalizability extent, we present the AUC score of generalizability performance over all combinations in fig. 5(b). In fig. 5(b), for Maze, Game 1 and 2 refers to Catch and Wobble respectively. Similarly, for Catch, they refer to Maze and Wobble respectively, and for Wobble, they refer to Maze and Catch. In the later subsections, we experimentally show that these performance numbers are sufficient to achieve 2.25x improvement in training time (in terms of the number of episodes required).

We performed preliminary experiments to gain fundamental insights into the extent of generalizability. All the three games considered in this work, differ in terms of their action space. Wobble can move either left or right (two actions), Catch has an additional "NOOP" (3 actions), and the agent in the Maze can move in either direction (4 actions). To understand the generalizability of ErrP in terms of the actions taken by the agent, we train on the Wobble, and test on the Catch game for two groups - (i) when the agent moves in either direction, and (ii) when the agent stays in the place. We obtain an average AUC score of 0.7359 ($\pm$ 0.1294) and 0.6423 ($\pm$ 0.1451) for both groups, respectively. Through a paired t-test, we found the difference in mean statistically significant. Similarly, for the Catch game, we test two groups - (i) when *egg* is close to the *paddle*, and (ii) when *egg* is far from the *paddle*. We found the mean AUC scores of 0.71 ($\pm$ 0.1) and 0.84 ($\pm$ 0.12) for each group, respectively. The difference of the mean of both groups was found statistically significant.

**Evaluation of Proposed Framework** In the evaluation of this framework, the trajectories given initially are generated based on optimal paths corrupted by randomly chosen wrong actions, which appear with the probability of 0.2 along the trajectory. We evaluate the performance with 10 and 20 given trajectories. Prior to training the RL agent, each subject is asked to provide feedback via ErrP on the state-action pairs along these trajectories. We conducted experiments on 5 subjects, based on Maze game. Here the Q network is modeled by Bayesian DQN. The performance of augmented RL algorithms is shown in Figure 6.

The reward function is shown to speed up the training convergence of the RL agent significantly. The number of ErrP inquiries in the proposed framework is $372.1(\pm 58.2)$, based on the statistics on empirical simulations. In full access method, ErrP labeling is needed on every trajectory generated during the training process. However, we can see that the proposed framework even outperforms the full access method, but it only needs ErrP inquiries on only 20 trajectories, which is much smaller than that in full access




(a) Learning Curve (b) Complete Episode (c) Learning Curve (d) Complete Episode

Figure 6: Evaluation of Second Framework: Learning from Imperfect Demonstrations Labeled by ErrP. Figures (a) and (b) are for the demonstration with 10 trajectories, and figures (c) and (d) are for 20 trajectories.

method.

## Conclusions and Future Work

We first demonstrate the feasibility of capturing error-potentials of a human observer watching an agent learning to play several different Atari-games, and then decoding the signals appropriately and using them as an auxiliary reward function to a DRL algorithm. Then we argue that the definition of ErrPs is generalizable across different environments. In the ideal approach, we validate the augmentation effect of ErrP labels on RL algorithms by the *full access* method. Then, in the practical approach, we propose an augmentation framework for RL agent, based on imitation learning. It is to learn a reward function from noisy demonstrations with ErrP labeling.

The demonstration of the generalizability of error-potentials is limited across the environments presented in the paper. We have considered discrete grid-based reasonably complex navigation games. The validation of the generalization to a variety of Atari and Robotic environments is the subject of future work. We plan to test our framework of integrating implicit human feedback (via ErrPs) over robotic environments, and text the *generalization* capability of error-potentials between virtual and physical worlds. As part of our future work, we also plan to investigate as to how machines can be assisted in DRL by using intrinsic EEG-based co-operations among humans and machines.

## References


Barachant, A., and Bonnet, S. 2011. Channel selection procedure using riemannian distance for bci applications. In *2011 5th International IEEE/EMBS Conference on Neural Engineering*, 348–351. IEEE.

Barachant, A., and Congedo, M. 2014. A plug&play p300 bci using information geometry. *arXiv preprint arXiv:1409.0107*.

Barachant, A.; Bonnet, S.; Congedo, M.; and Jutten, C. 2011. Multiclass brain–computer interface classification by riemannian geometry. *IEEE Transactions on Biomedical Engineering* 59(4):920–928.

Barachant, A.; Bonnet, S.; Congedo, M.; and Jutten, C. 2013. Classification of covariance matrices using a riemannian-based kernel for bci applications. *Neurocomputing* 112:172–178.

Blankertz, B.; Dornhege, G.; Schafer, C.; Krepki, R.; Kohlmorgen, J.; Muller, K.-R.; Kunzmann, V.; Losch, F.; and Curio, G. 2003. 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.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Carter, C. S.; Braver, T. S.; Barch, D. M.; Botvinick, M. M.; Noll, D.; and Cohen, J. D. 1998. Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science* 280(5364):747–749.

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

Chemali, J., and Lazaric, A. 2015. Direct policy iteration with demonstrations. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*.

Christiano, P. F.; Leike, J.; Brown, T.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, 4299–4307.

Congedo, M.; Barachant, A.; and Andreev, A. 2013. A new generation of brain-computer interface based on riemannian geometry. *arXiv preprint arXiv:1310.8115*.

Daniel, C.; Kroemer, O.; Viering, M.; Metz, J.; and Peters, J. 2015. Active reward learning with a novel acquisition function. *Autonomous Robots* 39(3):389–405.

El Asri, L.; Piot, B.; Geist, M.; Laroche, R.; and Pietquin, O. 2016. Score-based inverse reinforcement learning. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, 457–465. International Foundation for Autonomous Agents and Multiagent Systems.

Falkenstein, M.; Hoormann, J.; Christ, S.; and Hohnsbein, J. 2000. Erp components on reaction errors and their functional significance: a tutorial. *Biological psychology* 51(2-3):87–107.

Ferrez, P. W., and Millán, J. d. R. 2005. You are wrong!—automatic detection of interaction errors from brain waves. In *Proceedings of the 19th international joint conference on Artificial intelligence*, number CONF.

Gao, Y.; Lin, J.; Yu, F.; Levine, S.; Darrell, T.; et al. 2018. Reinforcement learning from imperfect demonstrations. *arXiv preprint arXiv:1802.05313*.

Heess, N.; Sriram, S.; Lemmon, J.; Merel, J.; Wayne, G.; Tassa, Y.; Erez, T.; Wang, Z.; Eslami, S.; Riedmiller, M.; et al. 2017. Emergence of locomotion behaviours in rich environments. *arXiv preprint arXiv:1707.02286*.

Hester, T.; Vecerik, M.; Pietquin, O.; Lanctot, M.; Schaul, T.; Piot, B.; Horgan, D.; Quan, J.; Sendonaris, A.; Osband, I.; et al. 2018. Deep q-learning from demonstrations. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Holroyd, C. B., and Coles, M. G. 2002. The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review* 109(4):679.

Iturrate, I.; Montesano, L.; and Minguez, J. 2010. Robot reinforcement learning using eeg-based reward signals. In *2010 IEEE International Conference on Robotics and Automation*, 4822–4829. IEEE.

Kim, B.; Farahmand, A.-m.; Pineau, J.; and Precup, D. 2013. Learning from limited demonstrations. In *Advances in Neural Information Processing Systems*, 2859–2867.

Knox, W. B. 2012. Learning from human-generated reward.

Laskey, M.; Lee, J.; Fox, R.; Dragan, A.; and Goldberg, K. 2017. Dart: Noise injection for robust imitation learning. *arXiv preprint arXiv:1703.09327*.

Luo, Y.; Xu, H.; and Ma, T. 2019. Learning self-correctable policies and value functions from demonstrations with negative sampling. *arXiv preprint arXiv:1907.05634*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

Nair, A.; McGrew, B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2018. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 6292–6299. IEEE.

Ng, A. Y.; Harada, D.; and Russell, S. 1999. Policy invariance under reward transformations: Theory and application to reward shaping. In *ICML*, volume 99, 278–287.

Osband, I.; Russo, D.; and Van Roy, B. 2013. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, 3003–3011.

Parra, L. C.; Spence, C. D.; Gerson, A. D.; and Sajda, P. 2003. 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.

Piot, B.; Geist, M.; and Pietquin, O. 2014. Boosted bellman residual minimization handling expert demonstrations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 549–564. Springer.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer School on Machine Learning*, 63–71. Springer.

Renard, Y.; Lotte, F.; Gibert, G.; Congedo, M.; Maby, E.; Delannoy, V.; Bertrand, O.; and Lécuyer, A. 2010. 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.

Rivet, B.; Souloumiac, A.; Attina, V.; and Gibert, G. 2009. xdawn algorithm to enhance evoked potentials: application to brain–computer interface. *IEEE Transactions on Biomedical Engineering* 56(8):2035–2043.

Ross, S.; Gordon, G.; and Bagnell, D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, 627–635.

Salazar-Gomez, A. F.; DelPreto, J.; Gil, S.; Guenther, F. H.; and Rus, D. 2017. Correcting robot mistakes in real time using eeg signals. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 6570–6577. IEEE.

Scheffers, M. K.; Coles, M. G.; Bernstein, P.; Gehring, W. J.; and Donchin, E. 1996. Event-related brain potentials and error-related processing: An analysis of incorrect responses to go and no-go stimuli. *Psychophysiology* 33(1):42–53.

Silver, D.; Hubert, T.; Schrittwieser, J.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Tesauro, G. 1995. Temporal difference learning and td-gammon. *Communications of the ACM* 38(3):58–68.

Van Hasselt, H.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double q-learning. In *Thirtieth AAAI conference on artificial intelligence*.

Večerík, M.; Hester, T.; Scholz, J.; Wang, F.; Pietquin, O.; Piot, B.; Heess, N.; Rothörl, T.; Lampe, T.; and Riedmiller, M. 2017. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards. *arXiv preprint arXiv:1707.08817*.

Wang, S.; Jia, D.; and Weng, X. 2018. Deep reinforcement learning for autonomous driving. *arXiv preprint arXiv:1811.11329*.

Wang, S. I.; Liang, P.; and Manning, C. D. 2016. Learning language games through interaction. *arXiv preprint arXiv:1606.02447*.

Ziebart, B. D. 2010. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Ph.D. Dissertation, figshare.