Towards Reinforcement Learning from Neural Feedback: Mapping fNIRS Signals to Agent Performance

Julia Santaniello[1], **Matthew Russell**[1], **Benson Jiang**[1], **Donatello Sassaroli**[1], **Robert Jacob**[1], **Jivko Sinapov**[1]

[1]Tufts University, Department of Computer Science
{julia.santaniello, benson.jiang, donatello.sassaroli, jivko.sinapov}@tufts.edu, {mrussell, jacob}@cs.tufts.edu

Abstract

Reinforcement Learning from Human Feedback (RLHF) is a methodology that aligns agent behavior with human preferences by integrating user feedback into the agent's training process. This paper introduces a framework that guides agent training through implicit neural signals, with a focus on the neural classification problem. Our work presents and releases a novel dataset of functional near-infrared spectroscopy (fNIRS) recordings collected from 25 human participants across three domains: Pick-and-Place Robot, Lunar Lander, and Flappy Bird. We train multiple classifiers to predict varying levels of agent performance (optimal, suboptimal, or worst-case) from windows of preprocessed fNIRS features, achieving an average F1 score of 67% for binary and 46% for multi-class classification across conditions and domains. We also train multiple regressors to predict the degree of deviation between an agent's chosen action and a set of near-optimal policy actions, providing a continuous measure of performance. Finally, we evaluate cross-subject generalization and show that fine-tuning pre-trained models with a small sample of subject-specific data increases average F1 scores by 17% and 41% for binary and multi-class models, respectively. Our results demonstrate that mapping implicit fNIRS signals to agent performance is feasible and can be improved, laying the foundation for future Reinforcement Learning from *Neural* Feedback (RLNF) systems.

Dataset — https://github.com/your-profile/fNIRS2RL

Code — https://github.com/your-profile/NeuroLoop-Classification/tree/aaai26

Introduction

Recent advances in Reinforcement Learning from Human Feedback (RLHF) have become crucial for training and fine-tuning state-of-the-art systems (Wu et al. 2023; Mosqueira-Rey et al. 2022). Specifically, RLHF has seen significant success by addressing limitations that plague common Reinforcement Learning (RL) techniques. Integrating human feedback has been incredibly beneficial for aligning agent behavior with human preferences. However, many feedback methods require active participation and/or expert demonstration for evaluative feedback to be most effective (Christiano et al. 2017; Li et al. 2019; Retzlaff et al. 2024). In addition, most approaches are limited to explicit feedback, which may lack insight into the nuances of human decision-making and internal assessments. Techniques that retrieve richer feedback may require tasks that impose greater cognitive effort on users (Lindner and El-Assady 2022). These drawbacks can be inconvenient for the evaluator or limit accessibility to certain user groups. They may also contribute to higher mental workload and unnatural, shallow feedback (Casper et al. 2023). These limitations may make it difficult to develop adaptive AI technologies that fully align with human preferences while being available to all user groups.

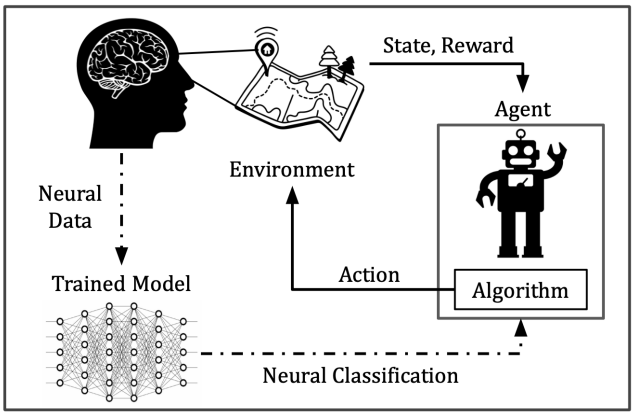

Figure 1: NEURO-LOOP: A high-level diagram of the proposed framework. This paper addresses the neural classification problem with the intent to apply trained models to the proposed pipeline.

Implicit human feedback systems present a valuable alternative by allowing agents to learn from natural human responses. Modern methods focus on facial expression or gesture classification to adapt agent behavior at little to no additional cognitive cost to the teacher (Cui et al. 2021). However, some of these procedures still require explicit instruction and conscious physical effort to adjust gestures and expressions. Further, reliance on private training data poses challenges for data collection and distribution.

To address these limitations, we propose *NEURO-LOOP*, a fully implicit neural feedback framework that leverages fNIRS signals to directly guide agents' training using

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

RLHF techniques. This framework employs passive Brain-Computer Interfaces (BCI) to align agent behavior with user intent through passive observation. We employ a functional near-infrared spectroscopy (fNIRS) device to record brain activity. fNIRS is a non-invasive neuro-imaging tool that can track sustained cognitive states such as mental workload, attention, and decision-making. This technology measures the hemodynamic response, or change in blood flow, in the prefrontal cortex (PFC) over time.

This paper addresses the neural classification problem as the first step toward integrating implicit fNIRS signals into RLHF frameworks. Our contributions are outlined below:

- We design and implement a controlled experimental protocol to investigate whether functional near-infrared spectroscopy (fNIRS) signals reflect varying levels of agent performance during human-agent interactions. We publicly release a novel dataset comprised of synchronized fNIRS recordings and agent transition variables.
- We demonstrate that machine learning can distinguish multiple levels of agent performance, extending past binary classification and enabling finer-grained feedback while maintaining low cognitive workload for users.
- We evaluate model performance and generalization between participants, suggesting that cross-subject transfer remains a challenge, but subject-specific calibration using limited data is feasible and improves performance.

We further compare model performance and user workload across both *passive* (observation) and *active* (physical demonstration) tasks, establishing a benchmark for future research in Reinforcement Learning from *Neural* Feedback (RLNF).

Related Work

Reinforcement Learning from Human Feedback (RLHF) allows agents to learn from human preferences through a range of approaches. These methods can be described as having either implicit or explicit feedback mechanisms.

Explicit Feedback: Explicit feedback is the most common modality used to shape an agent's policy to align with human expectations. Methods in this category often provide binary or scalar evaluative signals, such as preference labels or human-provided rewards, to shape the agent's policy or learning process (Griffith et al. 2013; Knox and Stone 2011). Past research indicates that evaluative metrics with wider ranges have greater training success (Yu et al. 2023). These interactions are often shorter-term due to increased cognitive workload and sustained physical demand. Imitation learning is a related approach that often requires expert demonstration or active participation for the agent to learn from user examples effectively (Hussein et al. 2017).

Implicit Feedback: Implicit feedback leverages unobtrusive human responses to enhance reinforcement learning agent behavior. These feedback techniques often include gaze tracking (Veeriah, Pilarski, and Sutton 2016), expression recognition (Arakawa et al. 2018), or gesture classification (Cui et al. 2021) to shape agent policies. One drawback is that some of these procedures require explicit instruction, as participants must consciously and physically adjust their gestures or expressions to provide meaningful feedback. We propose using neural signals to communicate evaluative feedback directly with minimal instruction.

Figure 2: Setup: Participants sat 24 inches in front of a computer screen. The fNIRS device is a headband that shines pulsating infrared light into the PFC to detect changes in blood flow.

ErrPs for Human-Robot Interaction: Error-related potentials (ErrPs) are neural signals that occur when the brain perceives an error. These signals, recorded via EEG, have been explored as a method for improving human-robot interaction. Previous research has applied ErrP signals to intervene in robotic decision-making by providing a binary indicator of error or dissatisfaction. Some work has even integrated ErrPs into RLHF algorithms, demonstrating improved training efficiency (Vukelić et al. 2023; Agarwal, Venkateswaran, and Sivakumar 2020; Xu et al. 2021). EEG signals are often transient and susceptible to artifacts (Rihet, Clodic, and Roy 2024). We explore the use of fNIRS as a complementary signal for brain-driven RLHF systems. Compared to other neuro-imaging devices, fNIRS produces signals that offer more tolerance to physical movement, greater portability, and higher spatial resolution (Pinti et al. 2018). Using the fNIRS system described by Giles (Blaney et al. 2020), we collected phase and intensity signals, making fNIRS suitable for long-horizon naturalistic settings outside the laboratory.

fNIRS for Human-Centered Technologies: Recent advancements have made it possible for brain signals to be used in human-centered technologies (Ayaz and Dehais 2021). When applied to adaptive frameworks, fNIRS data often monitors cognitive workload, allowing systems to adjust to user strain. Past work has explored applications like unmanned aerial vehicles (UAV) (Afergan et al. 2014), and human-robot systems (Solovey et al. 2012; Roy et al. 2020).

Motivation: fNIRS is quite under-explored in reinforcement learning and human-in-the-loop machine learning. fNIRS also offers beneficial characteristics and features that EEG generally lacks. Prior research has correlated fNIRS signals to reward-based gameplay events like video game performance (Li et al. 2018) and physical game performance (Bronner et al. 2012). Other work links the PFC to reward-based decision-making during vicarious and personal video game play (Morelli, Sacchet, and Zaki 2015), suggesting hidden potential for alignment with agent

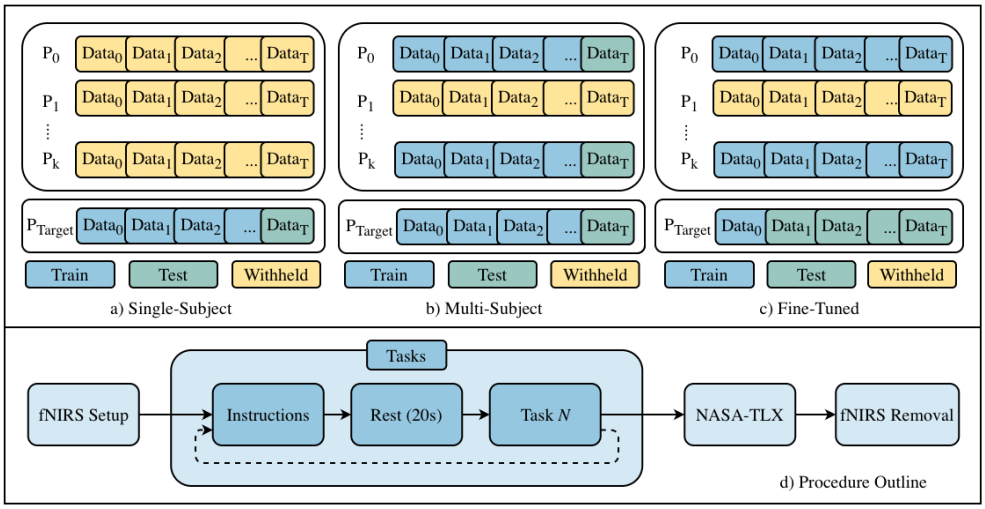

Figure 3: Training Paradigms: Diagram of the three training paradigms and Experimental Protocol. (a) Single-subject models are trained on a set of data from one participant and are evaluated using withheld data from the same participant. (b) Multi-subject models are trained on a set of participants and are evaluated using withheld data from the same set. (c) Fine-tuned models are multi-subject models calibrated with a fraction of a target participant's data.

evaluation metrics. Therefore, we propose mapping fNIRS signals to various degrees of agent performance, a first step toward creating an fNIRS-driven RLHF system.

Dataset Notation

We formalize the structure of the complete dataset as $\mathcal{D}$, and the subset of data from each participant k, denoted as $\mathcal{D}^k$. Dataset $\mathcal{D}^k$ is further divided into two subsets, the neural dataset and the task dataset.

Neural Dataset: We define the neural dataset $\mathcal{N}^k$ as the set of neural channel recordings over time for some participant k. Let $\mathcal{N}_i \in \mathbb{R}^{M \times T}$ denote the neural signal matrix at instance i, where M is the number of neural channels and T is the number of timestamps for some participant. A single neural data vector at timestamp t across all channels can be denoted as $\mathbf{n}_t \in \mathbb{R}^M$. The signal at a specific timestamp t and channel m can be expressed as $n_{t,m}$, where $t \in \{0, \ldots, T\}$ and $m \in \{0, \ldots, M\}$.

Task Dataset: We define the Task dataset $\mathcal{H}^k$ as the set of agent transition variables, or learning task statistics, over time for some participant k. Let $\mathcal{H}_i \in \mathbb{R}^{P \times T}$ represent the task data matrix at instance i, where P is the number of task variables and T is the number of timestamps. We may describe a vector of learning task statistics as $\mathbf{h}_k = \{S_t, A_t, R_t, S_{t+1}, V_t, B_t, E_t\}_{t=0}^{t=T}$, where S_t is the agent's state at time t, A_t is the action chosen by the agent or human, R_t is the reward, S_{t+1} is the next state, and V_t, B_t, E_t represent various agent performance values. An instance of a task statistic data point at some timestamp for a specific task variable can be denoted as $h_{t,p} = f(H_p, T_t)$, where $p \in \{0, \ldots, P\}$. After the raw neural data has been preprocessed, we hypothesize that a relationship exists between features of the neural data and the learning task statistics outlined above, where x is a neural feature vector. We let $\hat{y} = \phi(x_i, h_i)$ denote the relationship between a vector of human neural data features at some instance i and its agent performance label.

Methodology

Participants: We recruited 25 participants for a mixed within and between-participants study. Participants were between 19 to 27 years old and were recruited through physical and virtual flyers. Fourteen participants identified as female and eleven as male. Each participant completed 3-4 conditions out of a possible six conditions, resulting in at least 10 participants per condition.

Equipment: Neural data was collected using an ISS OxiplexTS fNIRS device. This device uses pulsating infrared lasers to calculate the change in hemodynamic response under the human skull. Three OpenAI Gymnasium reinforcement learning domains were adjusted for seamless interfacing with participants: Robot Fetch and Place, Lunar Lander, and Flappy Bird.

Domains: An agent $\mathcal{A}$ operates in one of three OpenAI Gym domains. In Lunar Lander, it uses `up`, `left`, `right`, and `down` to land between a set of flags without crashing. Flappy Bird uses actions `up` and `down` to clear a sequence of pipes/obstacles. In Robot Fetch and Place, the agent uses `x-left`, `x-right`, `y-up`, `y-down`, `z-up`, `z-down`, and `gripper-close` to place a cube at a marked goal.

High-Level Experimental Procedure: A complete experimental procedure included the administration of an informed consent form, configuration of the fNIRS device, task instruction, and two post-experiment questionnaires (Figure 3). Each task consisted of a set of episodes in which an agent attempts to complete a domain-specific goal. The participant's interaction with the agent was either *passive* or *active*. A condition is any combination of domain and interaction. Participants completed 3-4 conditions that lasted between 2-5 minutes each. Conditions were randomly se-

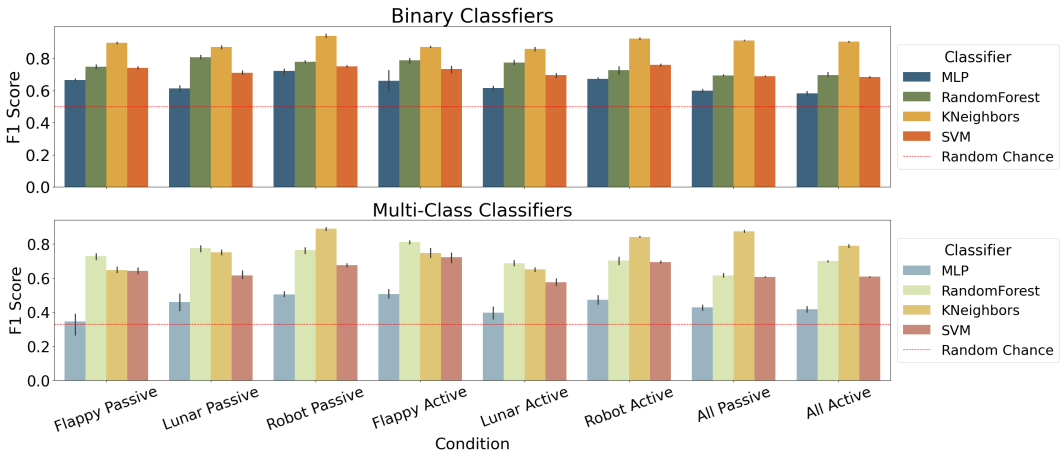

Figure 4: Multi-Subject Classification Performance: This figure illustrates classification performance (F1) for binary and multi-class models. Binary models attempted to classify optimal vs. suboptimal behavior, while multi-class models attempted to differentiate between optimal, suboptimal or worst-case agent behavior.

lected, but passive tasks were completed first. Participants were seated 24 inches from the computer screen that displayed the agent and domain. A 20-second rest/pause was taken before each task to calibrate the fNIRS device. The participant completed a NASA-TLX questionnaire for each task, and a post-task questionnaire at the end of the study.

Passive Task: The *passive* condition instructed participants to observe and reflect on the performance of an autonomous agent completing a goal within its environment. The autonomous agent initially selects actions from a near-optimal policy and can be described as successful. With some probability p, the agent transitions to a non-optimal action selection state and becomes unsuccessful for the remainder of an episode. We refer to this timestep as the *point of failure*. The *degree of failure* labels all actions after the point of failure accordingly. Due to the variation in each domain, the *point of failure* and *degree of failure* are determined differently for each.

In an environment with a *discrete* action space (Lunar Lander, Flappy Bird), non-optimal actions were selected by altering the agent's chosen action distribution. We let $\mathbf{p} = [p_1, p_2, \ldots, p_n]$ be the probability distribution over n actions. Let b be the index of the best action (i.e., the action with the highest probability): $b = \arg\max(p_i)$. Let w be the index of some non-optimal action (i.e., an action without the highest probability): $w = \arg\min(p_i)$. The new probability distribution before normalizing $\mathbf{p}'$ is defined as:

$$p'_i = \begin{cases} 0 & \text{if } i = b \\ p_w + p_b & \text{if } i = w \\ p_i & \text{otherwise} \end{cases}$$

Then the array is normalized by dividing by the sum. Increasing the value of the lowest probability increases the likelihood that the worst-case and non-optimal actions are chosen.

Optimal Lunar Lander agents often landed between the flags without crashing. Suboptimal agents often landed outside of the goal location. Worst-case agents crashed. Optimal Flappy Bird agents often completed long episodes (15+ pipes cleared). Suboptimal Flappy agents completed medium-length episodes (5-15 pipes). Worst-case Flappy agents often completed short episodes (≤ 5 pipes).

In an environment with a *continuous* action space (Robot Fetch and Place), suboptimal and worst-case action selection was implemented differently. Suboptimal actions altered the robot's goal, resulting in fluid, swift movements to the wrong goal state. Worst-case actions did not change the robot's goal, but added random variance to joint positions, resulting in unnatural, abrupt movements or dropping/throwing the block.

The agent continued to navigate through various action-selection states until the end of the task. The neural data, learning task statistics, and timestamps were stored in a hash map and saved as a de-identified demonstration.

Active Task: The *active* condition instructed a participant to physically guide an agent towards its goal using a keyboard or joystick. Participants used a basic computer keyboard to guide Flappy Bird and Lunar Lander agents, and an Xbox Controller to guide the Robot arm. As they played, the participant's chosen actions, earned rewards, and seen states were saved. The neural and task data were labeled and saved identically to Passive Tasks.

Post-Experiment: After the tasks were completed and the

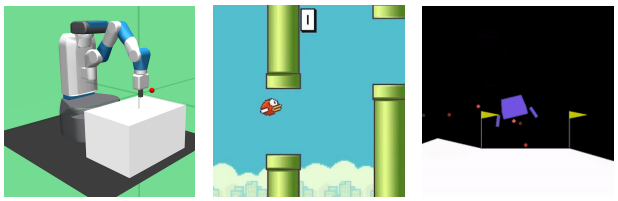

Figure 5: Agents and Domains: Robot Fetch and Place, Flappy Bird and Lunar Lander, respectively.

fNIRS device was removed from the participant, a NASA-TLX and Post-Task Questionnaire were administered. The NASA-TLX quantifies mental workload through a self-report questionnaire, offering insight into the participant's experience with each condition. The post-task questionnaire offers participants the opportunity to self-report variables that might confound results such as caffeine intake, sleep or technology familiarity.

Participant Data Summary: Two participants were excluded from data analysis due to technical difficulties or non-consent to the release of their neural data. Participants with inefficient data are also included in the dataset so researchers can explore the effect of noisy fNIRS data for future analyses (Muslimani and Taylor 2024). A Participant Data Summary is included with the dataset outlining eligibility and noteworthy disturbances during data collection.

Machine Learning Approaches

Classification Labels: To train classifiers and regressors, both discrete and continuous labels were recorded. These labels are equivalent to the calculated *degree of failure* and are assigned to all actions after the *point of failure* in an episode. Binary performance labels are denoted as $B_t \in \{0, 1\}$, where 0 represents optimal behavior and 1 represents suboptimal behavior. Multi-class performance labels are denoted as $\mathbf{V}$. At some time step t, $\mathbf{V}_t$ may be some discrete number from the set $V_t \in \{0, 1, 2\}$ where 0 represents optimal, 1 represents suboptimal, and 2 represents worst-case behavior. Continuous performance errors are denoted as $E(t) \in \mathbb{R}$. At some time step t, $\mathbf{E}_t$ may be some continuous value where lower values indicate near-optimal actions and higher values indicate suboptimal to worst-case actions. This error value was designated by the multi-policy agreement system outlined below.

Multi-Policy Action Agreement: To calculate *continuous* performance labels, we designed a multi-policy action classification system. In any reinforcement learning problem, there may be more than one optimal path to the same goal.

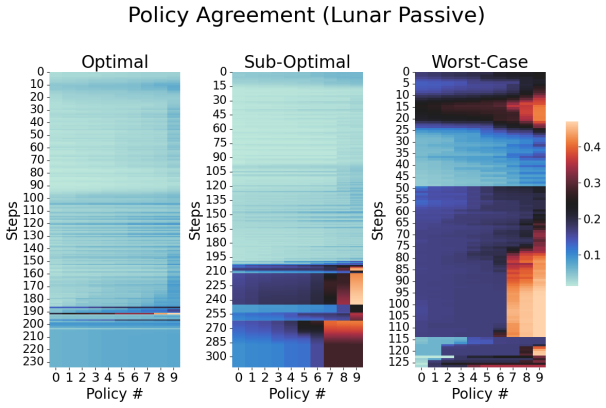

Figure 6: Multi-Policy Agreement Heatmap: These three heatmaps visually illustrate the *point of failure* and the *degree of failure* within each optimality class. Cool colors indicate low error, while warm colors indicate high error.

To avoid mislabeling agent behavior, we compare the agent's chosen action with K near-optimal policies, where $K = 10$. An error value is calculated between the agent's chosen action and each near-optimal policy's action. Then the average is calculated across the K policies. The error between the agent's discrete action and the k-th near-optimal policy $\pi_k(s_t)$ is calculated using Kullback-Leibler (KL) Divergence:

$$E_k(t) = D_{\text{KL}}\left(a_t \parallel \pi_k(s_t)\right) = \sum_{i=1}^{n} a_{t,i} \log\left(\frac{a_{t,i}}{\pi_k(s_t)_i}\right)$$

The error between the agent's continuous action and the k-th near-optimal policy $\pi_k(s_t)$ is calculated using Euclidean distance, where n is the size of the action space:

$$E_k\left(\pi_k(s_t), a_t\right) = \sqrt{\sum_{i=1}^{n} (\pi_k(s_t)_i - a_{t,i})^2}$$

We let a_t denote the agent's chosen action at time t, and let $\pi_k(s_t)$ represent the action taken by the k-th near-optimal policy at time t. The error between the agent's continuous action and the k-th near-optimal policy $\pi_k(s_t)$ is defined as $E_k(t)$. Then, the average error $\bar{E}(t)$ over K near-optimal policies is calculated and used as a continuous label. Figure 6 shows three heatmaps illustrating multi-policy action agreement over three episodes for each optimality class. Each performance category is shown to be distinct to our system and a *point of failure* can be identified within an episode.

Pre-Processing and Feature Extraction: We train a Support-Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest, and Multilayer Perceptron (MLP) using a time series classification approach. Each model was trained and tested on a set of participant data for some condition(s). We used a custom dual-slope frequency-domain fNIRS probe to suppress superficial and motion artifacts. The raw fNIRS data was calibrated using the 20-second baseline at the beginning of each trial. Intensity and phase values recorded from the left and right PFC were measured at 690 nm and 830 nm with 110 MHz modulation, and converted to oxy-/deoxy-hemoglobin. Resulting features were sampled at 5.2 Hz and band-pass filtered (0.001 - 0.2 Hz; 3rd order).

Time Classification Approach: We use a sliding window approach that extracts fixed-duration time windows, where each window was assigned a single label. In our experiments, window length varied per condition and was chosen through parameter studies. They were generally 5 to 7 seconds in length and 1 to 2 seconds in stride to address the 5 to 7 second latency of the fNIRS signal. The label assigned to an entire window was designated based on the endpoint label.

We denote our time series as: $x_{1:T}^{k,p} = [x_1^{k,p}, x_2^{k,p}, \ldots, x_T^{k,p}]$ where each subject k has neural data of length T, each terminating at an endpoint p. Each vector $x_t^{k,p} \in \mathbb{R}^F$ represents the F-dimensional data at time t. The endpoint p is associated with a label, as discussed above. Slope, mean, standard deviation, intercept, skewness, and kurtosis were

Condition	Binary Classification			Multi-Class Classification		
	Single-Subject	Cross-Subject	Fine-Tuned	Single-Subject	Cross-Subject	Fine-Tuned
Robot Passive	0.72 ± 0.005	0.54 ± 0.01	0.57 ± 0.03	0.50 ± 0.005	0.33 ± 0.01	0.41 ± 0.02
Robot Active	0.67 ± 0.005	0.53 ± 0.04	0.56 ± 0.02	0.48 ± 0.01	0.29 ± 0.02	0.35 ± 0.03
Lunar Passive	0.62 ± 0.01	0.45 ± 0.03	0.56 ± 0.02	0.46 ± 0.03	0.26 ± 0.03	0.36 ± 0.02
Lunar Active	0.62 ± 0.01	0.52 ± 0.04	0.54 ± 0.01	0.41 ± 0.02	0.27 ± 0.03	0.42 ± 0.08
Flappy Passive	0.67 ± 0.01	0.44 ± 0.04	0.52 ± 0.05	0.36 ± 0.06	0.26 ± 0.09	0.39 ± 0.12
Flappy Active	0.66 ± 0.06	0.46 ± 0.02	0.57 ± 0.03	0.51 ± 0.01	0.31 ± 0.03	0.51 ± 0.05

Table 1: Performance Across Multi-Subject Models (MLP): Average classifier performance (F1) for various levels of Model Granularity and Conditions. Binary and multi-class cross-validation saw mostly insignificant performance, but fine-tuning these models increased binary and multi-class model classification by nearly 17% and 41%, respectively. Random chance F1 performance was 0.50 for binary and 0.33 for multi-class classification.

calculated for each feature window resulting in a vector of length $F = 8 \cdot 6$.

Training Paradigms: Our models were trained and evaluated using one of three different paradigms: single-subject, multi-subject and fine-tuned training. These paradigms are based on prior work that aimed to better calibrate BCI models for specific users (Huang et al. 2021). As illustrated in 3, a single-subject model is trained on data from one participant and is validated using withheld data from the same participant. Multi-subject models were trained on data from a set of participants and evaluated using withheld data from the same set. Fine-tuned models are multi-subject models that were further trained with a fraction of a target participant's data and evaluated using withheld data from that target participant. Multi-subject and fine-tuned training data was exclusive to one condition, or a set of passive or active conditions.

To adjust for an imbalanced class distribution, we randomly down-sampled the majority class. We use a 60-20-20 train-test-validation split. Fine-tuned models use 20% of the target participant's down-sampled data to train on, and the remaining data was reserved for testing.

Cross-Validation Studies: leave-one-subject-out (LOPO), and in some cases leave-two-subjects-out, approach. From the set of participants in a condition set, one or two of those participants were completely withheld as a cross-validation testing group. These participants were excluded from the training set, allowing cross-subject evaluation of the model within the same condition.

Results

Our study evaluated the performance of binary and multi-class classification models across various task conditions, alongside cross-validation studies and a regression analysis. Below, we summarize our key findings.

Single-Subject Models: Single-subject models successfully distinguished between binary ($F1 = 0.79$) and multi-class ($F1 = 0.75$) levels of agent performance. Multi-class classifiers successfully differentiated between optimal, suboptimal, and worst-case behavior with little variance. However, cross-validation studies revealed that no single-subject model was able to predict agent performance from the neural data of another subject.

Multi-Subject Models: Binary multi-subject classifiers performed best on average, effectively learning patterns across a set of users in a specific condition, or set of conditions. Multi-class classification also showed promise, exceeding random chance in all but one condition: Flappy Passive. Like the Flappy Passive condition, some multi-class models struggled to distinguish between adjacent classes of agent performance. Task-specific challenges and variations in participant attention may have influenced performance, as the Flappy Active multi-class model performed much better than its Passive counterpart. Although model performance was generally consistent across conditions, Lunar Lander and Flappy Bird conditions showed higher variance and lower performance compared to both Robot conditions, shown in Table 1.

Cross-Validation Studies: Cross-validation studies revealed little success across conditions and participants. Only the Robot Passive condition demonstrated any notable potential for zero-shot transferability when evaluated on users within the same condition ($F1 = 0.54, \sigma = 0.01$). This finding aligns with prior work and suggests that developing generalizable BCI models, as explored in (Huang et al. 2021), remains an open challenge. These findings motivate the application of domain adaptation and deep learning techniques to improve robustness across participants and tasks (Wang et al. 2021; Eastmond et al. 2022).

Fine-tuned Models: Fine-tuned models leverage a limited amount of target participant data to calibrate multi-subject models, leading to improved cross-validation performance. With only about 20% of the downsampled target participant data, average scores increased by 16.9% and 41.3% for binary and multi-class models, respectively. These results suggest that even a small fraction of participant-specific data can help calibrate the model, boosting performance above random chance for an unseen, target participant.

Regression Performance: Most regression models performed particularly well, highlighting the ability to extract deeper complexity from fNIRS data beyond scalar feedback. Figure 7 illustrates the performance of MLP, K-Nearest Neighbors, and Random Forest regression models in the Robot Passive condition. Regressors trained on Active conditions performed marginally better on average ($R^2 = 0.81$) than Passive conditions ($R^2 = 0.77$). However, these models were unable to generalize to new participants. Finetuning with 30% of the target participant's data increased these scores, but did not raise performance to a sufficient degree

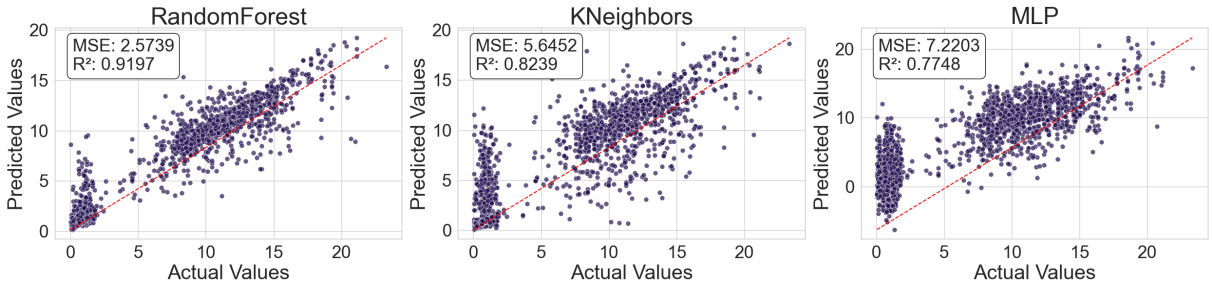

Figure 7: Regression for Multi-Subject Robot Passive Task: This figure compares actual and predicted performance values for participants passively observing the Robot condition. The model uses fNIRS data to predict a continuous performance label. All models showed strong correlation between fNIRS signals and agent performance, with MSE, R^2 values, and corresponding scatter plots.

for application in any condition other than Lunar Active (Cross-Sub $R^2 = 0.01$, Fine-tuned $R^2 = 0.23$). Regression models may need more target data to calibrate multi-subject models.

NASA-TLX Results: NASA-TLX allows participants to self-report mental workload through six different lenses: Mental, Temporal, Performance, Frustration, Physical, and overall Effort Demands. Self-reported cognitive load during Active tasks was higher for nearly every category (Figure 8). Lunar Lander Active tasks were particularly high across all categories. Passive tasks were least demanding; participants verbally described watching the autonomous Lunar Lander and Flappy Bird agent as "boring". These experiences may contribute to slightly lower Passive condition performance.

Discussion

Our results indicate that both binary and multi-class classifiers are capable of identifying performance categories with meaningful accuracy, and regression analysis further reveals that neural responses contain fine-grained evaluative signals beyond categorical labels. Our cross-subject evaluations highlight promising potential for generalization in the Robot Passive condition, suggesting that task structure and participant engagement may play critical roles in model transferability.

We identified three key limitations in this work. First, our dataset was partially imbalanced: some condition episodes were shorter, leading to fewer samples and underrepresented classes. While we addressed this via downsampling, the reduced data volume likely limited generalization and performance in these conditions. Second, self-reported NASA-TLX and post-task questionnaires revealed lower engagement and cognitive effort in certain passive conditions, which may have affected neural signal quality. Finally, while functional near-infrared spectroscopy (fNIRS) offers a non-invasive and accessible modality for Brain-Computer Interfaces (BCIs), it remains subject to signal noise, motion artifacts, and inter-subject variability. These technical challenges highlight the need for continued development of robust preprocessing techniques for neural classification tasks.

A natural next step would be to integrate these models into real-time RLHF frameworks to enable adaptive agent behavior based on implicit neural feedback. Future work should also explore personalized calibration strategies, and take advantage of multi-modal feedback (e.g., EEG, EMG, GSR) for greater insights into internal human assessments. Finally, we highlight the need to expand datasets to further support cross-subject generalization.

Conclusion

This paper demonstrates a measurable relationship between fNIRS data and agent performance, showing that fNIRS signals can be decoded into meaningful evaluative feedback. We trained models to distinguish levels of agent performance based solely on passive neural responses, and repeated our procedures using active demonstration to serve as a benchmark against explicit feedback modalities. We show that passive fNIRS signals can be mapped to various levels of agent performance, laying the groundwork for enabling seamless integration of neural feedback into RLHF systems.

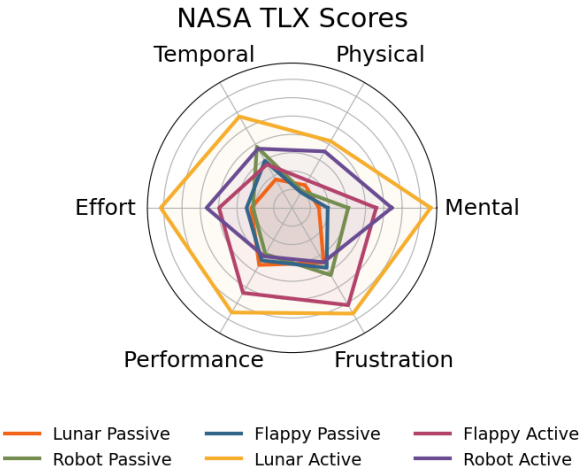

Figure 8: NASA-TLX Scores per Condition: NASA-TLX is a self-reported questionnaire that measures a participant's perceived cognitive workload. Average scores showed greater perceived cognitive workload for active tasks. Lunar Passive and Flappy Passive tasks were lowest overall.

Acknowledgments

We are grateful for the support and feedback of the MuLIP and HCI labs at Tufts University, especially Kenny Zheng, Anes Kim, Anna Sheaffer, Brennan Miller-Klugman, and Iris Yang.

References

Afergan, D.; Peck, E. M.; Solovey, E. T.; Jenkins, A.; Hincks, S. W.; Brown, E. T.; Chang, R.; and Jacob, R. J. 2014. Dynamic difficulty using brain metrics of workload. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, 3797–3806. Association for Computing Machinery.

Agarwal, M.; Venkateswaran, S. K.; and Sivakumar, R. 2020. 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*, 25–30. ACM.

Arakawa, R.; Kobayashi, S.; Unno, Y.; Tsuboi, Y.; and ichi Maeda, S. 2018. DQN-TAMER: Human-in-the-Loop Reinforcement Learning with Intractable Feedback. *Proceedings of 2nd Workshop on Human-Robot Teaming Beyond Human Operational Speeds and Robot Teammates Operating in Dynamic, Unstructured Environments*, abs/1810.11748.

Ayaz, H.; and Dehais, F. 2021. *Neuroergonomics*, chapter 31, 816–841. John Wiley & Sons, Ltd.

Blaney, G.; Sassaroli, A.; Pham, T.; Fernandez, C.; and Fantini, S. 2020. Phase dual-slopes in frequency-domain near-infrared spectroscopy for enhanced sensitivity to brain tissue: First applications to human subjects. *Journal of Biophotonics*, 13(1): e201960018.

Bronner, S.; Ono, Y.; Nomoto, Y.; Tanaka, S.; Sato, K.; Shimada, S.; Tachibana, A.; and Noah, A. 2012. Frontotemporal oxyhemoglobin dynamics predict performance accuracy of dance simulation gameplay. *NeuroImage*.

Casper, S.; Davies, X.; Shi, C.; Gilbert, T. K.; Scheurer, J.; Rando, J.; Freedman, R.; Korbak, T.; Lindner, D.; Freire, P.; Wang, T. T.; Marks, S.; Segerie, C.-R.; Carroll, M.; Peng, A.; Christoffersen, P. J.; Damani, M.; Slocum, S.; Anwar, U.; Siththaranjan, A.; Nadeau, M.; Michaud, E. J.; Pfau, J.; Krasheninnikov, D.; Chen, X.; Langosco, L.; Hase, P.; Biyik, E.; Dragan, A.; Krueger, D.; Sadigh, D.; and Hadfield-Menell, D. 2023. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*. Survey Certification, Featured Certification.

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 4302–4310. Red Hook, NY, USA: Curran Associates Inc.

Cui, Y.; Zhang, Q.; Knox, B.; Allievi, A.; Stone, P.; and Niekum, S. 2021. The EMPATHIC Framework for Task Learning from Implicit Human Feedback. In *Conference on Robot Rearning*, 604–626. PMLR.

Eastmond, C.; Subedi, A.; De, S.; and Intes, X. 2022. Deep learning in fNIRS: a review. *Neurophotonics*, 9(4): 041411.

Griffith, S.; Subramanian, K.; Scholz, J.; Isbell, C. L.; and Thomaz, A. L. 2013. Policy Shaping: Integrating Human Feedback with Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 26.

Huang, Z.; Wang, L.; Blaney, G.; Slaughter, C.; McKeon, D.; Zhou, Z.; Jacob, R. J. K.; and Hughes, M. C. 2021. The Tufts fNIRS Mental Workload Dataset & Benchmark for Brain-Computer Interfaces that Generalize. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Track on Datasets and Benchmarks*.

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2017. Imitation Learning: A Survey of Learning Methods. *ACM Comput. Surv.*, 50(2).

Knox, W. B.; and Stone, P. 2011. Augmenting Reinforcement Learning with Human Feedback. *ICML*.

Li, G.; Gomez, R.; Nakamura, K.; and He, B. 2019. Human-Centered Reinforcement Learning: A Survey. *IEEE Transactions on Human-Machine Systems*, 49(4): 337–349.

Li, Y.; Zhang, L.; Long, K.; Gong, H.; and Lei, H. 2018. Real-time monitoring prefrontal activities during online video game playing by functional near-infrared spectroscopy. *Journal of Biophotonics*, 11: e201700308.

Lindner, D.; and El-Assady, M. 2022. Humans are not Boltzmann Distributions: Challenges and Opportunities for Modelling Human Feedback and Interaction in Reinforcement Learning. In *Proceedings of the Communication in Human-AI Interaction (CHAI) Workshop at IJCAI-ECAI 2022*.

Morelli, S.; Sacchet, M.; and Zaki, J. 2015. Common and distinct neural correlates of personal and vicarious reward: A quantitative meta-analysis. *NeuroImage*, 112: 244–253.

Mosqueira-Rey, E.; Hernandez-Pereira, E.; Alonso-Rios, D.; Bobes-Bascaran, J.; and Fernandez-Leal, A. 2022. Human-in-the-loop machine learning: a state of the art. *Artificial Intelligence Review*, 56.

Muslimani, C.; and Taylor, M. E. 2024. Leveraging Sub-Optimal Data for Human-in-the-Loop Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '24, 2399–2401. International Foundation for Autonomous Agents and Multiagent Systems.

Pinti, P.; Tachtsidis, I.; Hamilton, A.; Hirsch, J.; Aichelburg, C.; Gilbert, S. J.; and Burgess, P. W. 2018. The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1464: 5 – 29.

Retzlaff, C. O.; Das, S.; Wayllace, C.; Mousavi, P.; Afshari, M.; Yang, T.; Saranti, A.; Angerschmid, A.; Taylor, M. E.; and Holzinger, A. 2024. Human-in-the-Loop Reinforcement Learning: A Survey and Position on Requirements, Challenges, and Opportunities. *Journal of Artificial Intelligence Research*, 79: 1–64.

Rihet, M.; Clodic, A.; and Roy, R. N. 2024. Robot Noise: Impact on Electrophysiological Measurements and Recommendations. In *Companion of the 2024 ACM/IEEE Interna-*

tional Conference on Human-Robot Interaction, 888–891. ACM.

Roy, R. N.; Drougard, N.; Gateau, T.; Dehais, F.; and Chanel, C. P. C. 2020. How Can Physiological Computing Benefit Human-Robot Interaction? *Robotics*, 9(4): 100.

Solovey, E.; Schermerhorn, P.; Scheutz, M.; Sassaroli, A.; Fantini, S.; and Jacob, R. 2012. Brainput: enhancing interactive systems with streaming fnirs brain input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2193–2202. ACM.

Veeriah, V.; Pilarski, P. M.; and Sutton, R. S. 2016. Face Valuing: Training User Interfaces with Facial Expressions and Reinforcement Learning. In *Proceedings of the IJCAI Workshop on Interactive Machine Learning*. ArXiv:1606.02807.

Vukelić, M.; Bui, M.; Vorreuther, A.; and Lingelbach, K. 2023. Combining brain-computer interfaces with deep reinforcement learning for robot training: a feasibility study in a simulation environment. *Front. Neuroergonomics*, 4.

Wang, L.; Huang, Z.; Zhou, Z.; McKeon, D.; Blaney, G.; Hughes, M. C.; and Jacob, R. J. K. 2021. Taming fNIRS-based BCI Input for Better Calibration and Broader Use. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, UIST '21, 179–197. Association for Computing Machinery.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xu, D.; Agarwal, M.; Gupta, E.; Fekri, F.; and Sivakumar, R. 2021. Accelerating Reinforcement Learning using EEG-based implicit human feedback. *Neurocomputing*, 460: 139–153.

Yu, H.; Aronson, R. M.; Allen, K. H.; and Short, E. S. 2023. From "Thumbs Up" to "10 out of 10": Reconsidering Scalar Feedback in Interactive Reinforcement Learning. *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4121–4128.