Probabilistic Policy Blending for Shared Autonomy using Deep Reinforcement Learning

Saurav Singh[1] and Jamison Heard[1]

***Abstract*— Technologies in machine learning and artificial intelligence have come a long way in decision making and system automation, but still faces difficult challenges in semi-automation and human-in-the-loop frameworks. This work presents a probabilistic policy blending approach for shared control between a human operator and an intelligent agent. The proposed approach assumes that the agent can control a system and the human operator needs to communicate the system's intended goal. A comparative study is presented between different arbitration functions that are used to blend the human and agent's actions. The proposed approach can achieve a variable level of assistance to the human operator successfully within discrete action space using the Lunar Lander game environment developed by OpenAI. Furthermore, human physiological data have been analyzed while the human interacts with the system and the agent using different arbitration functions. A correlation between the physiological data, arbitration level, and task performance was observed.**

***Index Terms*— arbitration, blending, DQN, human-robot team, human-in-the-loop, reinforcement learning, shared autonomy, workload**

I. Introduction

Technologies in machine learning and artificial intelligence have made significant advances in decision making and system automation in the past few decades. Big data and improvements in simulation environments have paved the way for advancements in the field of deep learning and reinforcement learning. Many robotic systems that rely on these learning methodologies have characteristics such as high sensitivity to changes in sensor data, low response time, robust control, and consistent performance. These agents lack the ability to solve novel problems and identify intended system goals. This warrants humans being "in the loop" and share system control with the agent. However, human actions are noisy and sub-optimal, which increases the difficulty of developing effective shared-control frameworks. Sharing control of a robotic arm with noisy human inputs and AI inputs, for example, can result in an ineffective system. Determining how to mediate control of a robotic system between participating teammates effectively is still being investigated by the research community.

The human is responsible for communicating the system's intended goal explicitly or implicitly using their actions if the goal representation is too complex. The intelligent agent is responsible for predicting the intended goal and drive the system to that goal. The human supervisor must remain vigilant and correct any agent actions that stray away from the goal. Novel situations like the misinterpretation of the environment due to noisy sensor readings may require the human to take full system control, which may result in an overloaded workload state and reduce team performance [1]. High task performance may be maintained with the agent in full control; however, this will result in the human being underloaded, disengaged with the system, and unable to mitigate undesired system states if agent autonomy fails [2]. A symbiotic relationship is needed between the human and agent in order to promote fluent team collaborations [3]. This requires the agent to understand the human's current state. One potential solution to this problem is to use the human's physiological signals to estimate the human's current state and adapt the agent's behavior accordingly to achieve a more fluent team collaboration.

Many attempts have been made to formalize shared control in robotics, where the control of a robot can be shared between the human operator and the intelligent agent [6]. Existing research recognizes the critical role of shared control in creating a more effective collaboration between the human and the agent [3], [5]–[7], [9]–[11]. However, the previous studies on shared control within discrete action space rely on a learned model or agent policies for arbitration and lack the flexibility of changing the arbitration function when needed [7], [16], [19]. This paper proposes a flexible probabilistic policy blending approach that arbitrates between the raw human actions and intelligent agent's actions based on the assumption that each teammate has an associated probability that the action suggested by that teammate is optimal.

Further, this study explores the relationship between the human's workload, physiological signals, and the proposed arbitration approach. The analysis provides insight into using the human workload state to adapt the arbitration function in order to maintain a nominal human workload state and increase the human-agent team performance. The proposed approach is validated using the Lunar Lander game environment [4], where the goal is to land a rocket on a randomly positioned landing pad. This environment provides an episodic event-based game that is generally considered difficult for humans, but intelligent agents can solve the environment. The goal of the environment (rocket's x-position with respect to the landing pad) is hidden from the agent; thus, requiring the human and agent to collaborate. The approach was validated using data from six research lab members who were not familiar with the experiment. The **key contributions** of this research are:

- A probabilistic policy blending approach that can provide a varying level of arbitration.
- Presented the analysis of the effects of different arbi-

[1]Saurav Singh and Jamison Heard are with Rochester Institute of Technology, Rochester, NY, 14623, USA {ss3337,jrheee}@rit.edu

tration functions on human perceived workload, physiological data, and task performance.

The rest of the paper is organized as follows: Section II presents related work in the field of shared autonomy, Section III proposes the probabilistic policy blending approach for shared autonomy, while laying down the details of the experimental design used to validated the proposed approach. Section IV presents the experimental results and Section V discusses the findings and concludes this paper.

II. Related Work

Many attempts have been made to formalize shared control in the past few decades, from binary control where the system has full or no automation to a fine-grained degree of control based on a specific arbitration function. Goertz first proposed the use of robot manipulators for handling nuclear material via teleoperation in the 1960s [8]. This marked the beginning of years of research on shared control of robotic systems with varying degrees of automation. One common shared control model is a fully autonomous takeover by the intelligent agent after a trigger event, such as a goal prediction exceeding a confidence threshold or a user command.

Instead of a full autonomous takeover, the actions suggested by the human and the intelligent agent may be arbitrated to improve team performance. S. Srinivasa et al. [5], [6] developed a policy-blending arbitration method to combine actions proposed by the human and the intelligent agent for problems where the assistance is provided by the agent over the distribution of goals rather than a single goal. The policy blending method maneuvered the robot close to the goal distribution, which allowed the agent to predict the intended user goal based on the user-specified trajectory. Most shared control strategies can be generalized with different arbitration functions for full autonomy takeover [9] or virtual fixtures [14]. S. Javdani [19] exploited hindsight optimization as a shared control approach. The goal prediction problem was formalized as a partially observable Markov decision process. This formulation did not wait for user input when goal confidence was below a threshold.

P. Trautman [18] used a probabilistic approach for shared control encompassing the idea of optimizing the intelligent agent's action over the human's suggested actions. Katyal et al. [15] expanded upon this by developing a blending approach to control a prosthetic limb system stating that arbitrating two independent actions without evaluation of the action's quality may result in catastrophic failure. This is due to the possibility of selecting sub-optimal actions more frequently, which can be avoided by optimizing the intelligent agent's policy over the human's suggested action. However, it is not always possible to have a mechanism to evaluate the independent actions.

The developments in deep reinforcement learning have provided a model-free control approach for robot automation, where no prior knowledge of the environment dynamics is required. Members of DeepMind [15] made significant contributions to the idea of controlling a system using deep reinforcement learning. Since then, deep reinforcement learning has been used to control many systems formulated as episodic Markov Decision Processes [17]. A Markov Decision Process (MDP) is a mathematical framework that describes an environment using the tuple (S, A, T, R). S is the set of states or state space, A is the set of actions or Action space, T is the transition matrix with probabilities of transitioning from one state to another, and R is the reward received after transitioning from one state to another due to action a, i.e., reward function. MDPs follow the markov property, which states that the future states depends only on the present state, i.e. the future does not depend on the past.

Deep reinforcement learning has been applied to shared control systems. S. Reddy, et al. [7] developed a policy blending scheme for shared control using deep reinforcement learning across three scenarios: (i) unknown dynamics, known goal space, and user policy; (ii) unknown dynamics and the user policy, known goal space; (iii) unknown dynamics, user policy, and goal space, minimum assumptions are made about the task environment and user policies. This policy blending method demonstrated that a human-agent team can perform better than the sole teammates in the lunar lander environment using a user-specified arbitration function that uses the agent's Q-value function to evaluate the relative quality of actions suggested by the human and the agent.

The quality of human actions can be negatively impacted if the human is not in the optimal state [1], [2]. Estimated human states can potentially carry information on the quality of human actions. Human state-based adaptive systems are on the rise and have seen many developments in the field of Human Robot Interaction (HRI) and Human Computer Interaction (HCI). Many recent works are attempting to close the feedback loop in human-robot teaming by estimating human states using human's physiological data and adapting robots behavior based on the estimated human states. A. Darzi and D. Novak [20] used physiological measures such as electrocardiogram (ECG), respiration rate, skin conductance, and electromyography (EMG) to estimate task difficulty, enjoyment, valance, and arousal and used it to adapt the task difficulty. L. Peternel et al. [21] focused on physical muscle fatigue as the adaptation criteria in a physical human-robot co-manipulation task. EMG signals were used to estimate the physical fatigue in humans and change the robot's behavior from learning human trajectories to taking over most of the repetitive physical work in the task. J. Schwarz and S. Fuchs [22] proposed a real-time assessment of multidimensional user state (RASMUS) framework which uses workload, fatigue, motivational aspects of engagement, attention, situation awareness (SA), and emotional states of the human and the associated effects on the task performance for triggering system adaptation.

The rising availability of physiological sensors at a lower cost in recent has encouraged the HRI community to utilize physiological data for adaptation. A. Kothig et al. [23], [24] proposed a software framework called HRI Physio Lib, which provides a set of scripts, tools, and methods to provide a more standardized integration of physiological adaptation

in HRI. The framework has a working implementation with cardiovascular sensing. Many studies around shared control paradigms that incorporate multi-modal guidance feedback to humans, use workload as a performance metric. Typical modalities for guidance feedback are haptic, auditory and visual [25], [26]. However, they do not leverage most of the human workload measurements by using it to adapt the control between the human and the control system (usually an AI or a robot). The aforementioned works provide evidence that physiological data of humans in a human-robot teaming setting can be used to adapt the agent's behavior to optimize task performance. The proposed approach lays the foundation for such a shared control paradigm, where the system will be able to adapt the level of control humans have over the system to maintain the workload within the normal range. Variations in human physiological data with different arbitration functions are also analyzed so the physiological data can be used directly for adaptation in the future.

This work investigates the different aspects of arbitration in shared control. Minimum assumptions are considered, where the system dynamics, user policy, and goal space are unknown. An intelligent agent with unknown system dynamics, goal space, and user policy often do not have an optimal policy. Some work suggests that raw human actions can be used directly to optimize the agent's policy [5]–[7], [19]. However, as the problem becomes more complex, human actions can be extremely noisy due to the inconsistent and uncertain nature of humans to reliably optimize agent's policy over it, especially within a discrete action space. Additionally, the existing policy blending strategies lack the flexibility to offer varying levels of control in a shared control setting with a discrete action space. The presented approach uses an arbitrator that chooses an action from the human's proposed action set probabilistically. This method allows for a more flexible arbitration mechanism to solve discrete action space problems. The action probabilities can be influenced by the human's internal states such as workload and fatigue to make the system human aware. The presented approach studies the relationship between the action probabilities and the human physiological data. Furthermore, it builds upon the question of how to achieving varying levels of assistance from the intelligent agent within discrete space.

III. Methodology

The general shared autonomy architecture for sharing the control of a system consists of a human agent, an AI agent and an arbitrator that blend the human and the AI policies (see Figure 1). Let $\beta \in [0, 1]$ be the arbitration coefficient, which is independent of human's and AI's policies. At each time step, the human and the AI suggests an action to be taken, a_h and a_r, respectively. The probabilistic policy blending approach associates a probability to the suggested actions of each agent and assumes that the AI's suggested action a_r is the optimal action with a probability of β and the human agent's suggested action a_h is the optimal action with a probability of $1 - \beta$. This can be represented as:

$$a_\beta = \begin{cases} a_r & ; p = \beta \\ a_h & ; q = 1 - \beta \end{cases} \quad (1)$$

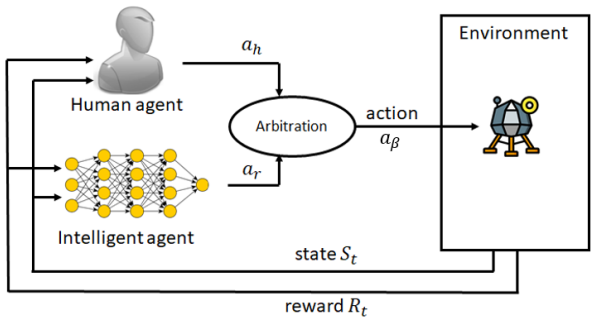

Fig. 1: The shared autonomy architecture.

where p and q are the assumed probabilities of the AI and the human agent's suggested actions, respectively, to be the near-optimal action for that time step. Setting $\beta = 0$ gives complete control to the human agent, while $\beta = 1$ gives complete control to the AI. $\beta \in (0, 1)$ gives a varying level of shared autonomy between the human and the AI. The AI will help in controlling the system while the user will have partial control over the system and communicate the intended goal implicitly through suggested actions. The user actions implicitly communicate the intended goal to the AI by partially driving the system or Markov Decision Process (MDP) towards the intended goal.

A. Task Environment

The lunar lander simulation [4] (see Figure 2) was used as the test environment for this study. The environment's goal is to land the rocket on the landing pad, indicated by the two yellow flags. The landing pad's x-position is generated within the game window randomly for each episode. The rocket starts from the top center of the screen with random initial position and velocity with respect to the landing pad. The rocket's state space consists of 8 different variables: x-position (x), y-position (y), x-velocity ($\dot{x}$), y-velocity ($\dot{y}$), angle (θ), angular velocity ($\dot{\theta}$), left leg touch down (Leg_{left}), and right leg touch down (Leg_{Right}). The goal location is hidden from the agent by removing the rocket's x-position

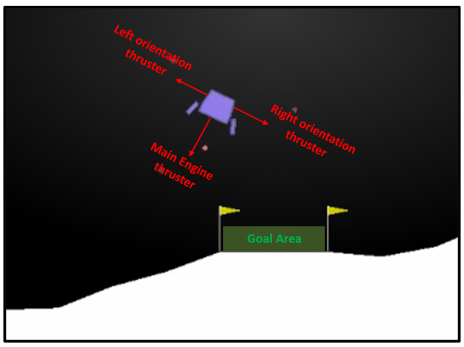

Fig. 2: Lunar Lander environment by OpenAI.

from the state-space. The goal is to accumulate the maximum rewards by landing the rocket on the landing pad. The action space consists of 6 possible discrete actions: pairs of {left, right, off} orientation thruster and {on, off} main engine thruster. The reward for descending from the top of the screen to the landing pad and reaching zero velocity after landing is approximately 100-140 points. Negative rewards occur when the rocket moves away from the landing pad; however, landing outside the landing pad is possible with no extra penalty. An episode ends if the rocket crashes or lands safely, receiving a reward of -100 or +100 points, respectively. Each leg ground contact receives +10 rewards while firing the main engine occurs a negative reward of -0.3. If an episode is running for more than 1000 steps, the episode ends and is not considered a crash or success. The reward function for each step can be represented as follows:

$$r(s) = -100.\sqrt{x^2 + y^2} - 100.\sqrt{\dot{x}^2 + \dot{y}^2} - 100.|\theta| + 10.Leg_{left} + 10.Leg_{right} \quad (2)$$

B. Experimental Setup

1) Training the AI policy: The Double Deep Q Network (DDQN) network [27] consists of 4 fully connected layers with 32, 64, 32, and 16 neurons in each layer, respectively, and 6 neurons in the output layer. The output layer provides the Q-values corresponding to each possible action. Relu and linear activation functions were used in the 4 fully connected layers and the output layer, respectively. The Adam optimizer (learning rate = 0.001) with a Mean Squared Error loss function was used to train the agent.

The agent uses the epsilon greedy policy to explore the state space during the training phase. It takes a random action (exploration) with a probability p and takes the action using the learned policy (exploitation) with a probability of q. The action suggested by the agent can be formulated as follows:

$$a_r = \begin{cases} random(A) & ; p = \epsilon \\ A[argmax(Q(a|s))] & ; q = 1 - \epsilon \end{cases} \quad (3)$$

where A is the action set and ϵ controls the ratio of exploration to exploitation and decays exponentially with each episode at a rate of 0.997.

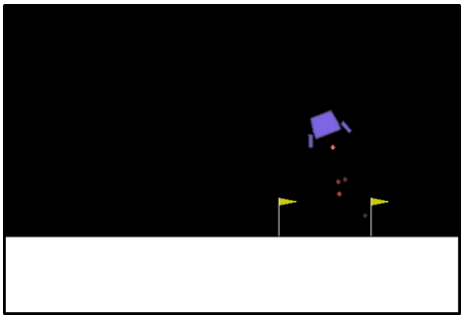

Fig. 3: Modified Lunar Lander environment for training AI.

The network was trained in a modified environment with the x-position of the rocket hidden from the agent. The environment was modified (see Figure 3) to have flat ground with no obstacles in order to promote the landing behavior. The reward function was also modified to remove any effects of receiving more rewards when the rocket lands closer to the landing pad while training. This was done by changing the first term in Equation 2 to $-100|y|$. The agent was trained until it achieved an average reward of 180 (modified environment is considered solved at a cumulative score of 180) over the last 100 episodes in the training environment, which was approximately 1300 episodes.

2) Action arbitration with preferred degree of assistance: There are 4 different types of arbitration functions used for this study (including baseline experiments). Let the human agent's suggested action be a_h and let the AI's suggested action be a_r. Following are the arbitration functions used for the study with β as the arbitration coefficient:

- **Arbitration BL:** Solo Human agent (Baseline).

$$a_\beta = a_h \quad (4)$$

- **Arbitration AI:** Solo AI (Baseline).

$$a_\beta = a_r \quad (5)$$

- **Arbitration HoAI:** The human's suggested action is always used (high priority). If no suggested action exists, then the AI's action is used (full control switching).

$$a_\beta = \begin{cases} a_r & ; \text{if } a_h = 0 \\ a_h & ; \text{else} \end{cases} \quad (6)$$

- **Arbitration Aβ:** AI's actions control the rocket with a probability of β and human agent's actions control the rocket with a probability of $1 - \beta$.

$$a_\beta = \begin{cases} a_r & ; p = \beta \\ a_h & ; q = 1 - \beta \end{cases} \quad (7)$$

The arbitration coefficient $\beta \in [0, 1]$ is interpreted as the preferred assistance coefficient. Arbitration *A0.0* ($\beta = 0.0$) represents full human control, while arbitration *A1.0* ($\beta = 1.0$) gives full control to the AI agent.

3) Human subjects experiment: The system was validated using data from six research lab members (five male and one female; age: mean = 24.5 $\pm$ 0.9 years) who were not familiar with the experiment. The average video game experience of the lab members was 5.2 on a scale of 1-9. Bias in the performance data was not expected as it is difficult to bias physiological data and game performance data without practice. The data collection was performed in compliance with the university's research lab operating procedures. The within-subjects experiment manipulated the arbitration function as the independent variable, which consisted of 7 levels: *BL1*, *HoAI*, *A0.3*, *A0.5*, *A0.7*, *BL2*, *AI*. Each lab member completed a 100 episode training session before completing the seven 100-episode long trials (each trial lasted 6 to 8 minutes), corresponding to each independent variable level.

A baseline collection (*BL1*) trial occurred after the training session to measure human performance without assistance. The lab members then completed the *HoAI* trial, where the human's actions were given priority over the agent's actions.

The lab members then completed three trials corresponding to the arbitration conditions *A0.3*, *A0.5*, *A0.7*, where the conditions were counterbalanced across the lab members to mitigate ordering effects. These arbitration conditions used the presented probabilistic policy blending method (Arbitration $A\beta$). The experiment ended with another baseline (*BL2*) trial with the human completing the environment without assistance. A 5-minute break occurred between each trial in order to allow the lab member's physiological signals to return to their resting state. Post-experiment, the agent solved the environment without human intervention with the goal hidden from the agent. This provided the baseline *AI*. Physiological data was collected from each lab member throughout the experiment using the portable BioHarness BT ECG monitor. The monitor measured the lab member's heart rate, heart rate variability, respiration rate, and posture magnitude which are preprocessed on-board the sensor. The BioHarness was strapped around the lab member's chest for the entirety of the experiment.

The dependent variables consisted of the Lunar Lander environment's performance metrics: rewards, time per episode, crash rate, success rate, land on pad, rate of change of the actions per minute, workload, heart rate, heart rate variability, respiration rate, and the NASA Task Load Index [28]. The NASA Task Load Index was completed after every trial in order to measure the human's perceived workload. Performance of human-AI team on Lunar Lander with different arbitration conditions was recorded for 100 episodes and the performance of last 30 episodes was averaged over the six lab members (see Table I).

C. Research Questions

The main research questions for this study are:

- **RQ1:** How does team performance change as the value of β changes in the probabilistic policy blending approach? If a strong relationship exists, then the arbitration value may be manipulated in order to optimize the team performance based on the current task context.
- **RQ2:** Is there a relationship between the human's physiological data, workload state, and the arbitration coefficient β? This analysis will determine if a system can use internal human data (physiological and workload) as feedback to a control system in order to set the arbitration coefficient appropriately.

Two hypotheses were proposed to address these questions:

- Hypothesis **H1** states that the *rewards* and *performance* metrics of human-agent team with arbitration $A\beta$ and $\beta \in (0, 1)$ will be greater than solo agent and solo human as pilots, focusing on the research question **RQ1**.
- Hypothesis **H2.1** states that the human's perceived workload level will significantly differ between the arbitration coefficients, investigating question **RQ2**.
- Hypothesis **H2.2** states that the human's perceived workload level will have a negative moderate correlation with the arbitration coefficients β. This hypothesis also focuses on the research question **RQ2**.
- Hypothesis **H2.3** states that the *rewards* will have moderate correlations with the human teammate's physiological data, i.e., heart rate, heart rate variability and respiration rate, focusing on the research question **RQ2**.

IV. Results

The performance of the human-agent team on Lunar Lander with different arbitration conditions was recorded for 100 episodes and the performance of last 30 episodes was averaged over the six lab members (see Table I). The results show that the average rewards for arbitration *A0.3*, *A0.5* and *A0.7* are much higher than rewards achieved by the solo human pilot (*BL1* and *BL2*) with a higher success rate and landing the rocket on the pad more frequently. Arbitration *A0.3* performed worse than solo *AI* but arbitration *A0.5* and *A0.7* still out-performed *AI* by achieving higher rewards and success rate with higher land-on-pad rate. A Kruskal Wallis test determined that rewards differed significantly between the arbitration conditions ($H(5) = 16.5490$, $p < 0.005$). The average *BL2* rewards were higher than *BL1*, due to the accumulation of experience during the experiment.

A higher rate of change of actions taken by the rocket allows the rocket to stabilize more easily. This metric is much higher with arbitration *A0.3*, *A0.5* and *A0.7* when compared to solo human and solo agent (see Table I). The solo human condition's low rate of change of actions represents higher difficulty in stabilizing the rocket. The time taken by the rocket per episode to descend with shared control between the agent and the human is greater than the solo human pilot. This is because the agent tries to stabilize the rocket while descending resulting in a slower descent. The average rate of change of actions per minute taken by the human pilot decreases as the value of arbitration coefficient β increases as the actions are dependent more on the agent's actions.

TABLE I: Evaluation of different arbitration functions for Human-agent Teaming on Lunar Lander. Mean (Standard Deviation) of last 30 episodes out of 100 episodes of each trial is shown here for six lab members. Best value of each metric is represented by a **bold values**.

Conditions	β	Reward	Time per episode	Crash Rate	Success Rate	Land on Pad	Δ Actions / min	Workload
BL1	0.0	-183.66 (55.62)	4.09 (1.07)	0.91 (0.15)	0.02 (0.02)	0.02 (0.02)	118.74 (28.26)	69.61 (7.71)
HoAI	-	-117.80 (92.81)	6.03 (1.41)	0.81 (0.15)	0.18 (0.15)	0.16 (0.13)	483.61 (100.60)	41.88 (13.27)
A0.3	0.3	-105.27 (102.07)	5.64 (1.06)	0.82 (0.13)	0.16 (0.13)	0.13 (0.13)	903.77 (47.48)	**47.22 (10.08)**
A0.5	0.5	-25.58 (103.31)	**7.21 (1.37)**	0.65 (0.20)	**0.32 (0.19)**	**0.27 (0.18)**	**1126.26 (86.69)**	**47.50 (16.78)**
A0.7	0.7	**-13.03 (38.81)**	6.73 (1.46)	**0.63 (0.15)**	0.29 (0.10)	0.22 (0.08)	1091.33 (79.85)	**47.27 (13.54)**
BL2	0.0	-123.57 (45.39)	4.57 (1.11)	0.92 (0.07)	0.07 (0.07)	0.05 (0.07)	138.46 (50.52)	69.16 (15.30)
AI	1.0	-60.14 (67.52)	6.89 (1.73)	0.67 (0.16)	0.28 (0.15)	0.19 (0.10)	853.06 (149.66)	-

TABLE II: NASA-TLX Ratings of different arbitration functions for Human-agent Teaming on Lunar Lander with baselines BL1 removed. Mean (Standard Deviation) of each trial is shown here for six lab members. Best value of each metric is represented by a **bold values**.

	HoAI	A0.3	A0.5	A0.7	BL2
Mental	30.0 (6.3)	**-20.0 (14.1)**	-14.2 (18.2)	-16.7 (14.7)	0.0 (17.6)
Physical	-15.8 (20.1)	**-11.7 (19.9)**	-10.0 (14.5)	-2.5 (17.8)	-4.2 (24.8)
Temporal	-30.8 (14.3)	-20.0 (24.9)	-20.8 (19.1)	**-30.0 (17.9)**	0.8 (10.2)
Performance	-13.3 (50.5)	-10.0 (49.0)	-23.3 (51.1)	**-24.2 (48.5)**	-8.3 (29.8)
Effort	-29.2 (8.6)	**-20.0 (14.8)**	-9.7 (15.3)	-18.3 (12.9)	8.3 (17.5)
Frustration	3.3 (41.1)	**-8.3 (38.29)**	-6.7 (40.5)	-7.5 (38.7)	15.8 (38.9)

The quality of human actions can be negatively impacted if the human is not in the normal workload state; thus, having an adverse effect on the task performance. The overall human workload was lower during the shared-control paradigms than the solo human, but a Kruskal-Wallis test found no significant difference in workload between the arbitration conditions ($H(5) = 0.07317, p > 0.85$). Amongst the 6 lab members, two lab members reported the lowest overall workload at $\beta = 0.3$, two lab members reported the lowest overall workload at $\beta = 0.5$, while the remaining lab members reported the lowest workload at $\beta = 0.7$ for arbitration $A\beta$. This outcome suggests that arbitration coefficients need to be individualized to a human. A similar trend was observed between average rewards and arbitration coefficients per lab member. Three lab members achieved the highest overall rewards at $\beta = 0.5$, while the remaining lab members achieved the highest overall rewards at $\beta = 0.7$, which again suggests that arbitration coefficients need to be individualized to a human. A negative correlation between lab member's workload and average rewards was observed with a Spearman correlation of $r_s = -0.4867$, $p = 0.0026$. This shows that human's perceive lower workload when the team performance is better. This further supports the usability of human's internal states like workload to adapt robot's behavior and improve overall team performance. Although only six lab members were recruited from our research lab due to COVID-19 restrictions, we believe that we will observe similar trends even with larger sample size.

Table II shows the NASA-TLX Ratings for each condition averaged over the six lab members. The results in Table II show that for arbitration $A\beta$, the mean mental demand and effort is the least for *A0.3*, followed by *A0.5*, and highest for *A0.7*. This shows that taking away more control from the human does not result in a reduction in mental demand and effort. One possible explanation for this behavior is that a higher mental demand was induced due to over-engagement in a non-demanding task [29] [30]. The Spearman correlation between the mental demand and effort was found to be $r_s = 0.6681, p < 0.0001$. This again suggests that arbitration coefficients need to be individualized to a human.

Temporal demand for the task is lower with the intelligent agent as co-pilot when compared to solo human as pilot. Arbitration *A0.7* have much lower temporal demand when compared to *A0.3* and *A0.5*. This behavior was expected as a higher value of β in the arbitration $A\beta$ can help reduce the temporal demand of a task. Frustration within the arbitration $A\beta$ trials was expected to be higher after trials with a lower Success Rate. However, for arbitration $A\beta$, lab members reported the highest frustration with $\beta = 0.5$ which had the highest success rate, and lowest frustration with $\beta = 0.3$ which had the lowest success rate. The Spearman correlation between the mental demand and effort was found to be $r_s = -0.4243$, $p < 0.01$. This suggests that frustration is not only dependent on the outcome of the trials but also on other factors like how much control of the system human have, perceived mental workload and effort.

This work also investigates how human physiological data varies with different arbitration functions. The results in Table III shows that within arbitration $A\beta$ trials, the mean heart rate (HR) had an inverse relation with the mean mental demand, effort, frustration, and success rate. Arbitration *A0.5* had the highest success rate, mental demand, effort, frustration, and lowest heart rate while arbitration *A0.3* had the lowest mental demand, effort, frustration, success rate, and highest heart rate. The average respiration rate (RR) had a direct relation with the mental demand, effort, frustration, and success rate. There was not much difference observed in the mean heart rate variability (HRV) within arbitration $A\beta$ trials; however, the Spearman correlation test showed that there was a moderate correlation between the heart rate variability and the rewards with $r_s = 0.4976, p < 0.009$. The heart rate variability was moderately correlated to changes in human actions per minute with $r_s = 0.6025$, $p < 0.001$.

TABLE III: Physiological data evaluation of different arbitration functions for Human-agent Teaming on Lunar Lander. Mean (Standard Deviation) of each trial is shown here for six lab members.

Trial	β	Heart Rate	Heart Rate Variability	Respiration Rate
BL1	0.0	86.35 (10.72)	44.70 (10.96)	17.68 (6.63)
HoAI	-	83.04 (8.48)	49.81 (11.04)	18.62 (4.96)
A0.3	0.3	79.58 (10.22)	56.68 (14.88)	17.78 (2.81)
A0.5	0.5	**78.84 (10.39)**	54.00 (9.74)	**18.10 (4.20)**
A0.7	0.7	81.40 (8.44)	54.02 (9.51)	17.29 (4.23)
BL2	0.0	77.87 (7.23)	58.85 (11.23)	18.52 (5.02)

There was no correlation between the heart rate and the rewards with $r_s = 0.0229$, $p > 0.9$ and a weak correlation between the respiration rate and the rewards with $r_s = 0.3411$, $p < 0.06$. The Spearman correlation between the heart rate and rewards for individual arbitration trials were analysed to understand this behavior with $r_s = 0.9428$,

$p < 0.005$ for *A0.3*, $r_s = 0.3000$, $p > 0.6$ for *A0.5*, and $r_s = 0.8999$, $p < 0.03$ for *A0.7*. Heart-rate was highly correlated to the rewards for *A0.3* and *A0.7*, but shows weak correlation for *A0.5*. This is due to the saturation in the human-agent team performance during trial *A0.5* with the rewards being the highest amongst all the arbitration schemes. This further indicates that there exist a relationship between overall team performance and the physiological which may not be linear.

These results suggest that there is a relationship between the human physiological signals and human workload states and task performance. This can be leveraged to design an adaptive human-robot teaming system that can adapt the agent's behavior based on human states estimated using only real-time physiological data and task performance.

V. Discussion & Conclusion

Human and agents that can complete a given task individually can benefit from teaming up together if the human-agent team can achieve a higher performance than either teammate can achieve individually. The proposed approach for human-agent teaming needs to perform better than solo-human and solo-agent. Hypothesis **H1** predicted that the *rewards* and *performance* metrics (i.e., *land on pad*) of the human-agent team with arbitration $A\beta$ and $\beta \in (0, 1)$ will be greater than solo agent and solo human as pilots. The criteria for the hypothesis **H1** was met by arbitration conditions *A0.5* and *A0.7*, but not by *A0.3*; thus, the hypothesis was partially supported. Partially supporting **H1** demonstrates that there is an optimal range of β in a shared-control paradigm which is likely task environment-specific. Additionally, the results suggest that an arbitration coefficient needs to be tailored to an individual in order to achieve optimal team performance.

A human-agent team may achieve near-optimal performance if the human is under normal workload state [1]. The proposed approach may adapt the arbitration coefficient based on real-time human workload estimates [31] if there is a direct relationship between the arbitration coefficient β and the human workload. Hypothesis **H2.1** predicted that the human's perceived workload level will significantly differ between the arbitration coefficients and hypothesis **H2.2** predicted that the human's perceived workload level will have a negative moderate correlation with the arbitration coefficients β. The average workload did not significantly differ between the arbitration coefficients; thus, **H2.1** was not supported. However, there was a moderate negative correlation between workload and overall rewards (team performance); thus, supporting **H2.2**. This negative correlation is attributed to individual differences (e.g., skill levels and preferences), as the lab member's lowest reported workload occurred at different arbitration coefficients. This suggests that the value of β needs to be adapted for individual users to achieve optimal workload and performance. Future work will investigate the use of the proposed approach in order to maximize team performance by changing the arbitration coefficient β and adapting based on individual user's metadata (e.g., preferences, skill level, past performance) and current state (e.g., workload, stress, fatigue).

The proposed approach with lab members achieved comparable results to the simulated pilot-copilot teams on Lunar Lander in [7]. Specifically, *A0.5* and *A0.7* achieved a higher success rate than all their models but their simulated Laggy pilot-copilot team achieved slightly higher rewards than the probabilistic policy blending approach. A comparison with their real human subjects study was not possible since they modified the game environment for their study with real humans to make the vehicle's legs more resistant to crashing on impact with the ground.

One potential limitation of the proposed policy blending approach is that the quality of the actions suggested by either of the teammates is not analyzed. Instead, the approach relies on the assumed probabilities of a teammate's action being near-optimal which makes the approach indeterministic. This may lead to the system taking sub-optimal actions occasionally when the assumption does not hold true.

Relationships between the task performance, human workload states, and raw physiological data with different arbitration functions were also investigated. Since the Lunar Lander environment was a more cognitively demanding task, greater focus was given to the trends that emerged in mental demand, effort, and frustration with respect to task performance metrics such as success rate. Taking away more control from the human did not result in a reduction in mental demand workload and effort, this again suggests that arbitration coefficients need to be individualized to a human. This may be due to the human staying focused on the task even though the agent had most of the control of the rocket.

Changes in the mean heart rate and mean respiration rate of the human rate were observed in response to changing mental demand, effort, and frustration which had a direct impact on the success rate of the task. Hypothesis **H2.3** predicted that the *rewards* will have moderate correlations with the human teammate's physiological data, i.e., heart rate, heart rate variability and respiration rate. The criteria for the hypothesis **H2.3** was met with *rewards* metric showing strong correlation with heart rate, moderate correlation with the heart rate variability, and weak correlation with respiration rate; thus, the hypothesis was partially supported. For a more effective adaptive human-robot teaming system, human states should be used as a part of the adaptive strategy and leverage the most from relationships between the human physiological signals and task performance. Future work will investigate the use of physiological data to estimate human workload states and use it not only as a trigger for adaptation but also use it to choose the adaptation hyper-parameters such as the value of $beta$ in the proposed policy blending approach.

Humans can deal with novel problems and identify goals for the system better than the agent but human actions are generally sub-optimal, especially under different workload conditions. In the proposed probabilistic policy blending approach, this can be taken into account by associating a higher/lower probability of human suggested action being optimal based on human workload via the arbitration coefficient β. If a human is underloaded or overloaded, their suggested actions are less likely to be optimal. Thus, the ar-

bitration coefficient needs to be lower or higher, respectively.

In conclusion, this paper presented a flexible probability-based arbitration approach for shared control with reinforcement learning. The arbitration approach presented in this paper assumes that agent policy is not perfect and thus, does not depend on the agent policy for evaluation of the suggested actions by each teammate. The agent in this paper was trained in an easier environment with a relaxed goal and represents a scenario where the exact goal representation for the task may be too complex to model for training the agent in the first place. The proposed approach allows humans to communicate the intended goal of the system implicitly through user actions without the need to have an explicit goal representation. The human-agent team performed better than the solo human and solo agent as the pilots. The proposed arbitration approach can also be used to implement varying levels of control each teammate gets while sharing control of a system within a discrete action space. Trends in the human physiological data with respect to arbitration coefficient were studied which can be used to optimize the arbitration coefficient β in future studies. The proposed policy blending approach offers a method to fine-tune shared autonomy to an individual human and arbitrate control of a system based on human's internal states such as workload, and fatigue that can be estimated using physiological data.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

REFERENCES

[1] Wickens, C. D., Lee, J. D., Liu, Y., and Becker, S. E. G. (2004). *An Introduction to Human Factors Engineering*. Pearson Education, Inc., 2nd edition.

[2] M. S. Young and N. A. Stanton, "Malleable Attentional Resources Theory: A New Explanation for the Effects of Mental Underload on Performance," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 44, no. 3, pp. 365–375, 2002.

[3] G. Hoffman, "Evaluating Fluency in Human–Robot Collaboration," in IEEE Transactions on Human-Machine Systems, vol. 49, no. 3, pp. 209-218, June 2019.

[4] OpenAI, "A toolkit for developing and comparing reinforcement learning algorithms," Gym.

[5] A. D. Dragan and S. S. Srinivasa, "A policy-blending formalism for shared control," *The International Journal of Robotics Research*, vol. 32, no. 7, pp. 790–805, 2013.

[6] N. Roy, P. Newman, S. Srinivasa, "Formalizing Assistive Teleoperation," in *Robotics: Science and Systems VIII* , MIT Press, 2013, pp.73-80.

[7] S. Reddy, A. Dragan, and S. Levine, "Shared Autonomy via Deep Reinforcement Learning," arXiv:1802.01744 [cs.LG], Feb. 2018.

[8] R. Goertz, "Manipulators used for handling radioactive materials," in *Human Factors in Technology, E. Bennett, J. Degan, and J. Spiegel, Eds. New York: McGraw Hill*, 1963, pp. 425–443.

[9] E. You and K. Kris, "Assisted Teleoperation Strategies for Aggressively Controlling a Robot Arm with 2D Input," *Robotics: Science and Systems VII. editor / Hugh Durrant-Whyte ; Nicholas Roy ; Pieter Abbeel. MIT Press Journals*, 2012. pp. 354-361.

[10] K. K. Hauser, "Recognition, prediction, and planning for assisted teleoperation of freeform tasks," *Autonomous Robots 35*, 2013, pp. 241-254.

[11] L. V. Herlant, R. M. Holladay and S. S. Srinivasa, "Assistive teleoperation of robot arms via automatic time-optimal mode switching," *2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Christchurch, 2016, pp. 35-42.

[12] D. Kragic, P. Marayong, M. Li, A. M. Okamura, and G. D. Hager, "Human-Machine Collaborative Systems for Microsurgical Applications," *The International Journal of Robotics Research*, vol. 24, no. 9, pp. 731–741, 2005.

[13] N. Mehr, R. Horowitz and A. D. Dragan, "Inferring and assisting with constraints in shared autonomy," *2016 IEEE 55th Conference on Decision and Control (CDC)*, Las Vegas, NV, 2016, pp. 6689-6696.

[14] D. Aarno, S. Ekvall and D. Kragic, "Adaptive Virtual Fixtures for Machine-Assisted Teleoperation Tasks," Proceedings of the 2005 IEEE International Conference on Robotics and Automation, Barcelona, Spain, 2005, pp. 1139-1144.

[15] K. D. Katyal et al., "A collaborative BCI approach to autonomous control of a prosthetic limb system," *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, San Diego, CA, 2014, pp. 1479-1482.

[16] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Human-level control through deep reinforcement learning," *Nature 518*, 2015, pp. 529–533.

[17] V. Mnih, K. Kavukcuoglu, D. Silver et al., "Playing Atari with Deep Reinforcement Learning," arXiv:1312.5602 [cs.LG], Dec. 2013.

[18] P. Trautman, "Assistive Planning in Complex, Dynamic Environments: A Probabilistic Approach," *2015 IEEE International Conference on Systems, Man, and Cybernetics*, Kowloon, 2015, pp. 3072-3078.

[19] S. Javdani, H. Admoni, S. Pellegrinelli, S. S. Srinivasa, and J. A. Bagnell, "Shared autonomy via hindsight optimization for teleoperation and teaming," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 717–742, 2018.

[20] A. Darzi and D. Novak, "Automated affect classification and task difficulty adaptation in a competitive scenario based on physiological linkage: An exploratory study," *International Journal of Human-Computer Studies*, vol. 153, Sept. 2021.

[21] L. Peternel, N. Tsagarakis, D. Caldwell and A. Ajoudani, "Adaptation of robot physical behaviour to human fatigue in human-robot co-manipulation," *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 489-494, 2016.

[22] J. Schwarz, S. Fuchs, "Multidimensional Real-Time Assessment of User State and Performance to Trigger Dynamic System Adaptation," in *Augmented Cognition. Neurocognition and Machine Learning*, D. D. Schmorrow, C. M. Fidopiastis, Eds., Springer International Publishing, 2017, pp. 383-398.

[23] A. Kothig, J. Muñoz, S. A. Akgun, A. M. Aroyo and K. Dautenhahn, "HRI Physio Lib: A Software Framework to Support the Integration of Physiological Adaptation in HRI," in *Social Robotics*, A. R. Wagner, D. Feil-Seifer, K. S. Haring, S. Rossi, T. Williams, H. He, and S. S. Ge, Shuzhi, Eds., Springer International Publishing, 2020, pp. 36-47.

[24] A. Kothig, J. Muñoz, S. A. Akgun, A. M. Aroyo and K. Dautenhahn, "Connecting Humans and Robots Using Physiological Signals – Closing-the-Loop in HRI," *2021 30th IEEE International Conference on Robot & Human Interactive Communication*, pp. 735-742, 2021.

[25] D. Powell and M. K. O'Malley, "The Task-Dependent Efficacy of Shared-Control Haptic Guidance Paradigms," in *IEEE Transactions on Haptics*, vol. 5, no. 3, pp. 208-219, Third Quarter 2012.

[26] R. Sigrist, G. Rauter, R. Riener, and P. Wolf, "Augmented visual, auditory, haptic, and multimodal feedback in motor learning: A review," *Psychonomic Bulletin & Review*, vol. 20, pp. 21–53, 2012.

[27] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, no. 1, 2016.

[28] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," Advances in Psychology, pp. 139–183, 1988.

[29] J. D. Lee, "Dynamics of Driver Distraction: The process of engaging and disengaging." *Annals of advances in automotive medicine. Association for the Advancement of Automotive Medicine. Annual Scientific Conference*, vol. 58, pp. 24-32, 2014.

[30] F. Dehais, M. Causse, F. Vachon, S. Tremblay, "Cognitive conflict in human-automation interactions: a psychophysiological study," *Applied ergonomics*, vol. 43(3), pp. 588–595.

[31] J. Heard, R. Heald, C. E. Harriott, and J. A. Adams, "A Diagnostic Human Workload Assessment Algorithm for Collaborative and Supervisory Human–Robot Teams," *ACM Transactions on Human-Robot Interaction*, vol. 8, no. 2, pp. 1–30, 2019.