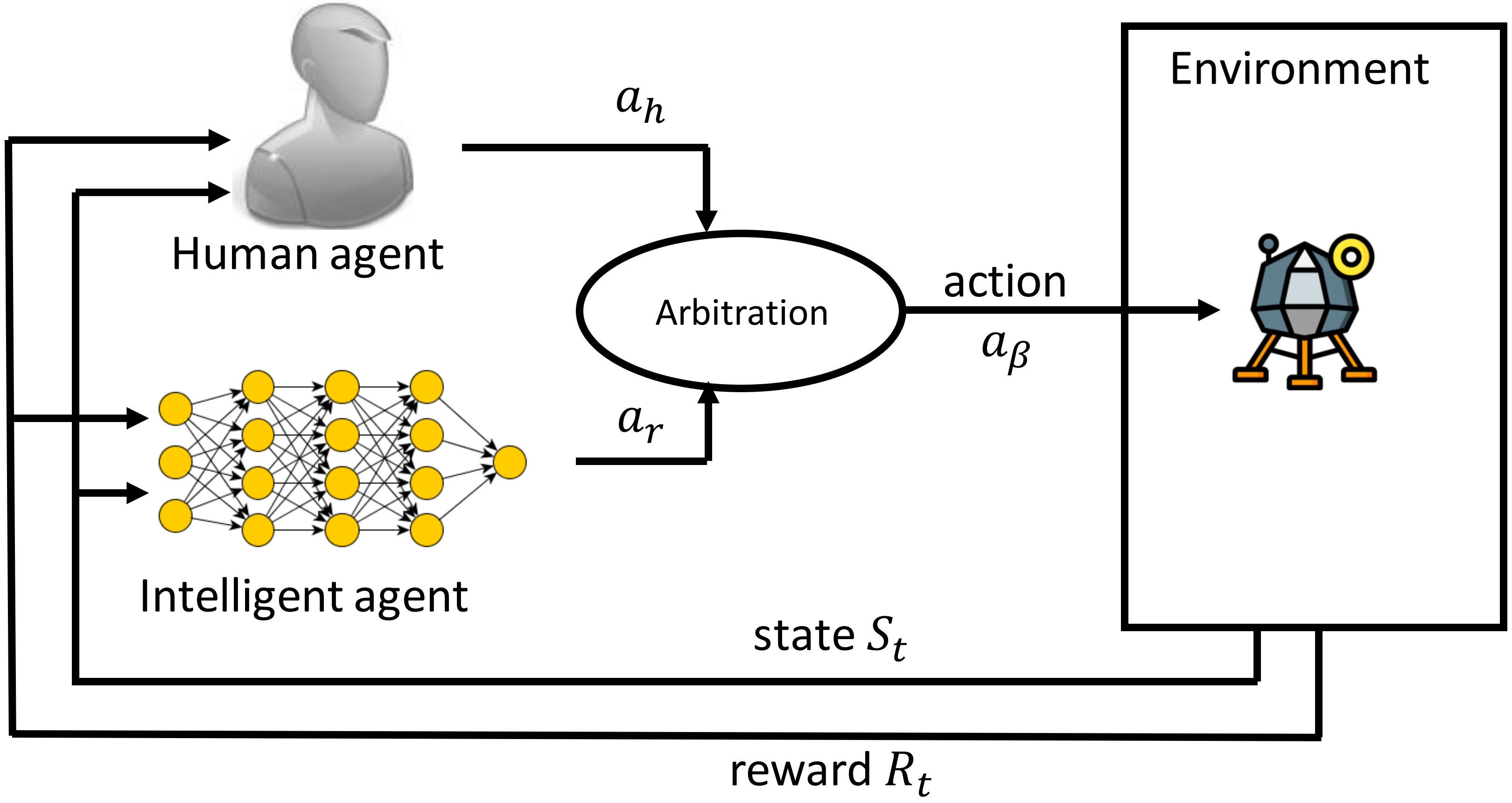
Probabilistic Policy Blending for Shared Autonomy using Deep Reinforcement Learning

Saurav Singh
Rochester Institute of Technology
Rochester, NY, USA
ss3337@rit.edu

Jamison Heard
Rochester Institute of Technology
Rochester, NY, USA
jrheee@rit.edu

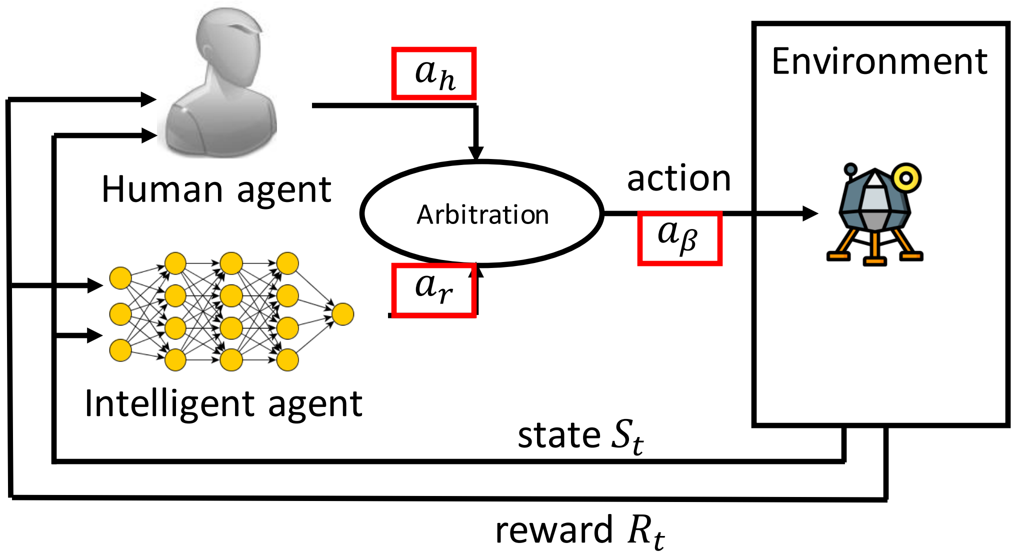
Motivation



Aim of this study:

- A probabilistic policy blending approach that can provide a varying level of arbitration.
- Study the effects of different arbitration functions on human perceived workload, physiological data, and task performance.

Methodology



Action arbitration with preferred degree of assistance:

Arbitration BL: Solo Human agent (Baseline) $a_\beta = a_h$

Arbitration AI: Solo AI (Baseline) $a_\beta = a_r$

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).

$$a_\beta = \begin{cases} a_r & ; \quad a_h = 0 \\ a_h & ; \quad a_h \neq 0 \end{cases}$$

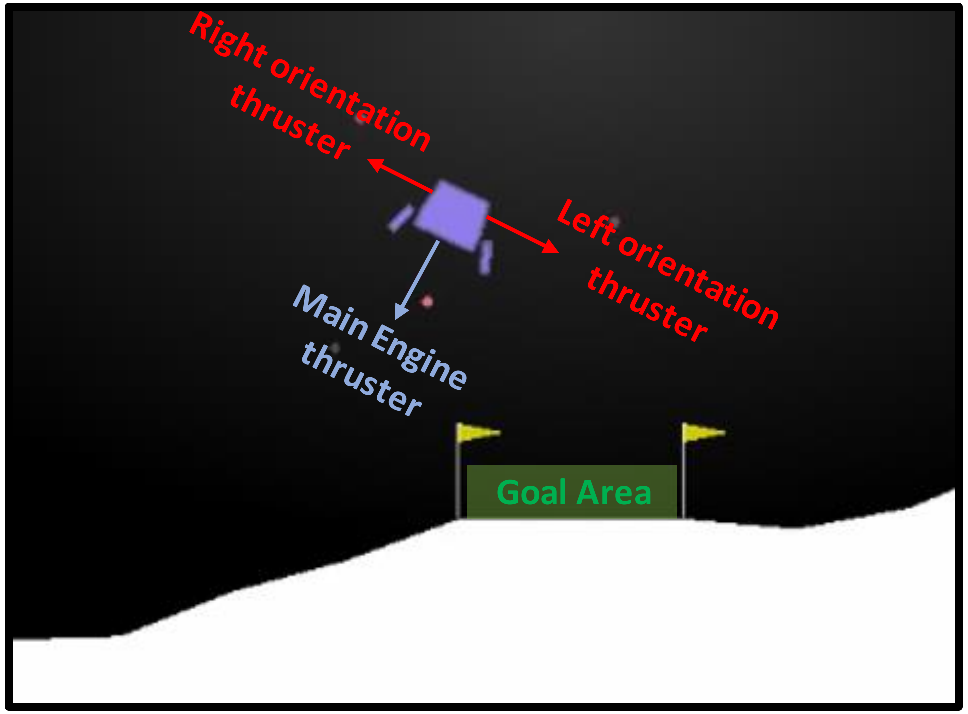
Arbitration Aβ, β $\in (\mathbf{0}, \mathbf{1})$***:*** AI's actions control the rocket with a probability of β and human agent's actions control the rocket with a probability of $1 - \beta$.

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

Task Environment: Lunar Lander by OpenAI Gym

State Space:

$$S = \{\underbrace{x, y}_{Hidden\ Goal}, \dot{x}, \dot{y}, \theta, \dot{\theta}, Leg_{left}, Leg_{Right}\}$$



Action Space:

Orientation thruster: {left, right, off}

Main engine thruster: {on, off}

6 possible discrete actions:

$A = \{(left, on), (right, on), (off, on), (left, off), (right, off), (off, off)\}$

Reward Function:

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.

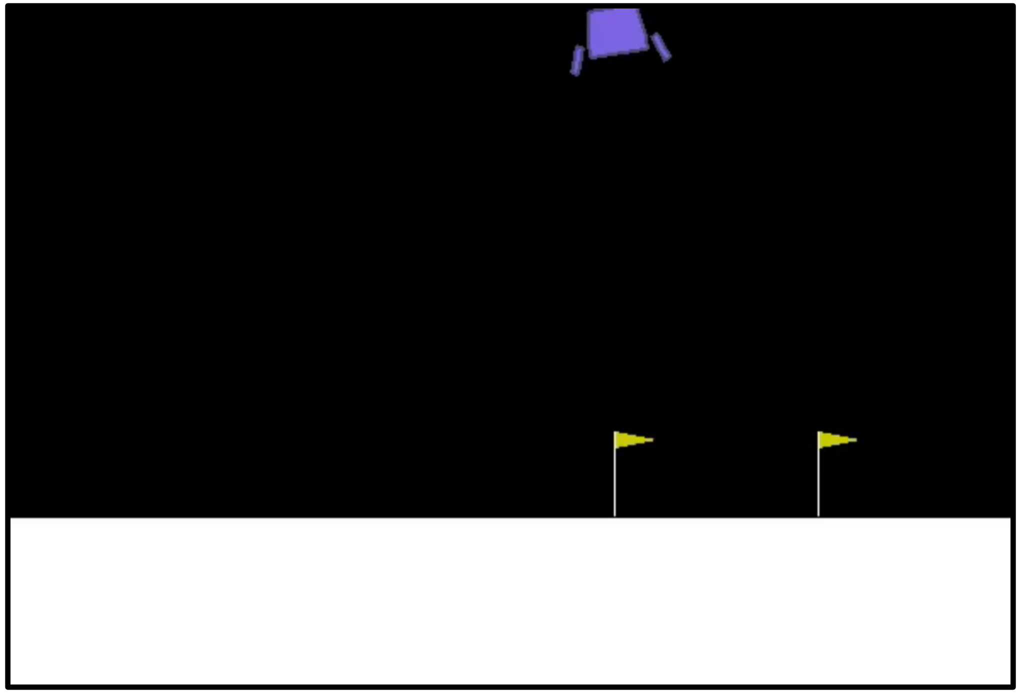
$$r(s) = -100\sqrt{x^2 + y^2} - 100|\theta| + 10 * Leg_{left} + 10 * Leg_{Right}$$

Pre-Training the RL agent

RL algorithm: Double Deep Q Network (DDQN)

State Space: $S = \{\cancel{x}, y, \dot{x}, \dot{y}, \theta, \dot{\theta}, Leg_{left}, Leg_{Right}\}$

Reward Function: $r(s) = -100\sqrt{x^2 \cancel{+} |y| y^2} - 100|\theta| + 10 * Leg_{left} + 10 * Leg_{Right}$



Human Subjects Experiment

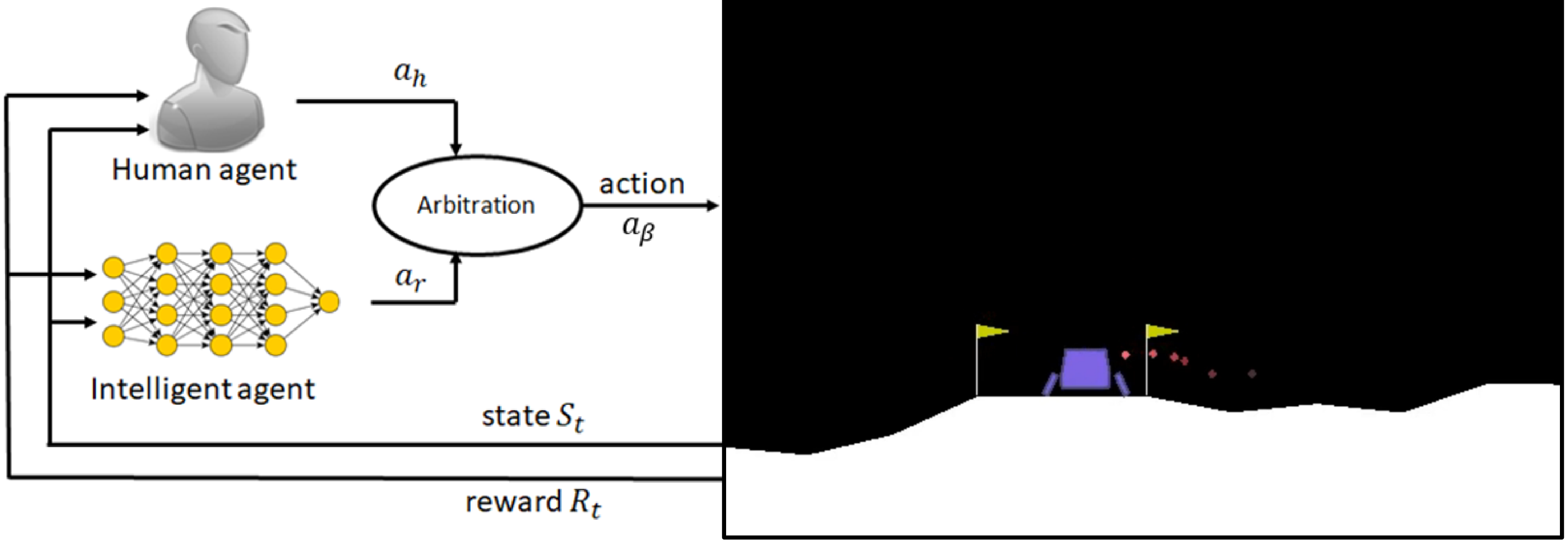
Experiment Trials:

BL1 | **HoAI** | [**A0.3** | **A0.5** | **A0.7** — Random Order] | **BL2** | **AI**

BioHarness BT
ECG monitor

NASA TLX
Task Load Index

Subjective
workload rating



Results

Mean 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 **bold values**.

Trial	β	Reward	Crash Rate	Success Rate	Land on Pad	Perceived Workload	Heart Rate	Heart Rate Variability	Respiration Rate
BL1	0.0	-183.66	0.91	0.02	0.02	69.61	86.35	44.70	17.68
HoAI	-	-117.80	0.81	0.18	0.16	41.88	83.04	49.81	18.62
A0.3	0.3	-105.27	0.82	0.16	0.13	**47.22**	79.84	56.68	17.78
A0.5	0.5	-25.58	0.65	**0.32**	**0.27**	**47.50**	**78.84**	54.00	**18.10**
A0.7	0.7	**-13.03**	**0.63**	0.29	0.22	**47.27**	81.40	54.02	17.29
BL2	0.0	-123.57	0.92	0.07	0.05	69.16	77.87	58.85	18.52
AI	1.0	-60.14	0.67	0.28	0.19	-	-	-	-

Conclusion

- Presented a flexible probability-based arbitration approach for shared control with reinforcement learning.

- 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.

- 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.

Thank You!!

Any questions

Please feel free to reach out to us at ss3337@rit.edu