[1] Priya N. Parkhi
[2] Himesh Ganwani
[3] Manav Anandani
[4] Bhagyashree Hambarde
[5] Poorva Agrawal
[6] Gagandeep Kaur
[7] Pournima Pande
[8] Devika Verma[8]

Reinforcement Learning with Human Feedback: A CartPole Case Study

Abstract: - This research delves into the integration of human feedback within reinforcement learning (RL) algorithms, with a specific focus on the CartPole environment as a testbed. We present RLHFAgent, a revolutionary RL agent devised to capitalize on human guidance during training for the purpose of expediting the learning process. Through the acquisition of feedback from a human operator, RLHFAgent adapts its policy in a more efficient manner, resulting in enhanced performance when it comes to balancing the pole. Our approach involves the training of a neural network model that approximates the policy function, mapping observations to actions, and subsequently updating this model based on human feedback. By means of a series of experiments, we showcase the efficacy of RLHFAgent in learning the art of balancing the pole, as demonstrated by the consistent rise in episodic rewards and the decrease in episodic loss over the course of training episodes. These findings indicate that the incorporation of human intuition into RL algorithms can augment their ability to adapt and expedite the learning process in intricate environments. In essence, this study contributes to the ongoing endeavours aimed at bridging the gap between RL algorithms and human expertise, thereby paving the way for more efficient and effective learning strategies in both simulated and real-world scenarios.

Keywords: Reinforcement learning, human feedback, cartpole environment, neural network model.

I.INTRODUCTION

Reinforcement Learning (RL) is a fundamental paradigm in the field of artificial intelligence, empowering agents to make sequential decisions in dynamic environments. However, despite its progress, RL faces challenges, particularly in scenarios where the state and action spaces are high-dimensional or rewards are sparse. The CartPole environment serves as an example of such a challenge, as it is a classic benchmark problem in RL that requires the agent to balance a pole upright on a moving cart. Although this task may seem simple, it actually presents significant obstacles, demanding precise control and coordination to maintain stability.

The main focus of this paper is on the slow learning process of RL algorithms, which can limit their applicability in real-world settings. Traditional methods often require extensive interactions with the environment to achieve satisfactory performance, which can be problematic in domains where rapid learning is crucial.

[1] Priya N. Parkhi, Department of Computer Science and Engineering(AIML), Shri Ramdeobaba College of Engineering & Management Nagpur, India- 440013
priyaparkhi1@gmail.com
[2] Himesh Ganwani, Department of Computer Science and Engineering(AIML), Shri Ramdeobaba College of Engineering & Management Nagpur, India- 440013
himeshganwani@gmail.com
[3] Manav Anandani, Department of Computer Science and Engineering(AIML), Shri Ramdeobaba College of Engineering & Management Nagpur, India- 440013
manavanandani304@gmail.com
[4] Bhagyashree Hambarde, Department of Computer Science and Engineering(AIML), Shri Ramdeobaba College of Engineering & Management Nagpur, India- 440013
manavanandani304@gmail.com
[5] Poorva Agrawal, Computer Science and Engineering Department, Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India
poorvaagrawal3@gmail.com
[6] Gagandeep Kaur, Computer Science and Engineering Department, Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India
gagandeep.kaur@sitnagpur.siu.edu.in
[7] Pournima Pande, Yeshwantrao Chavan College of Engineering, Wanadongari, Nagpur
pournimapande@yahoo.co.in
[8] Devika Verma, Vishwakarma Institute of Information Technology, Pune
devika.verma@viit.ac.in

The motivation behind this research stems from the recognition of the value of human intuition and expertise in accelerating the learning process. [3] By incorporating human feedback, we aim to guide RL algorithms more effectively, leading to faster convergence and improved performance. This motivation is particularly relevant in complex or poorly understood environments, where human oversight can provide valuable insights.

To tackle this challenge, we introduce RLHFAgent—a new RL agent specifically designed to learn in the CartPole environment with the aid of human feedback. RLHFAgent utilizes a neural network model to approximate the policy function, which maps observations to actions. During training, the agent interacts with the environment, seeks feedback from a human operator based on its actions, and updates its policy accordingly. This iterative process enables RLHFAgent to adapt more efficiently to the task at hand. Experimental results demonstrate the effectiveness of RLHFAgent in learning to balance the pole upright. By integrating human feedback into the training process, RLHFAgent achieves faster convergence and superior performance compared to conventional RL methods. Notably, the episodic rewards consistently increase over training episodes, indicating progressive improvement. At the same time, the episodic loss decreases as the agent refines its predictions, highlighting the effectiveness of incorporating human guidance.

Hence, this study highlights the potential of leveraging human feedback to enhance the learning efficiency and adaptability of RL algorithms. By bridging the gap between RL techniques and human intuition, RLHFAgent emerges as a promising approach for effectively tackling complex tasks. These findings contribute to the ongoing advancement of RL methodologies, with implications for a wide range of real-world applications.

II. LITERATURE REVIEW

Previous researches have delved into diverse approaches aimed at enhancing the performance of reinforcement learning (RL) algorithms. These approaches encompass deep reinforcement learning, policy gradient techniques, and actor-critic frameworks. Moreover, a number of scholarly inquiries have sought to examine how the integration of human feedback within RL frameworks can contribute to the optimization of learning efficiency. We have examined a number of notable academic papers that delve into the domain of interactive reinforcement learning with human feedback, each presenting distinct perspectives and methodologies to enhance the performance of learning.

"Human-centred Reinforcement Learning: A Survey" [1] offers a comprehensive overview of cutting-edge algorithms in human-centred RL. This survey paper thoroughly discusses various interpretations of human evaluative feedback and investigates research on agents learning from both human feedback and environmental rewards. Furthermore, it explores strategies aimed at improving the efficiency of human-centred RL algorithms. This survey sheds light on the dynamic landscape of human-centred RL, emphasizing the crucial role of human feedback in enhancing learning efficiency.

"Provably Feedback-Efficient Reinforcement Learning via Active Reward Learning" [2] introduces an RL algorithm based on active learning that explores the environment without requiring a predefined reward function. Instead, it solicits queries about task rewards from a human instructor at specific state-action pairs, ensuring a policy that is nearly optimal with high probability while effectively handling random noise in the feedback. Despite its promising approach, the paper recognizes challenges such as the design of accurate reward functions and the need for substantial human feedback.

"A Review on Interactive Reinforcement Learning from Human Social Feedback" [5] presents a comprehensive review of interactive RL methods that leverage human social feedback. It discusses various frameworks such as TAMER, VI-TAMER, and Actor-critic TAMER, along with methodologies like Deep TAMER for learning from human feedback. Moreover, the review explores transparent learning mechanisms and the combination of multiple modal inputs for agent training. Despite its breadth, the review highlights challenges such as sample efficiency and the diverse interpretations of human feedback.

Lastly, "Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance" [4] investigates RL with human teachers, employing modified action selection mechanisms and attention direction channels within a Q-Learning algorithm. By collecting data from expert training sessions and conveying rewards as feedback messages, the study aims to enhance learning performance. However, it acknowledges limitations, including assumptions about the interpretation of feedback and reliance on standard algorithms.

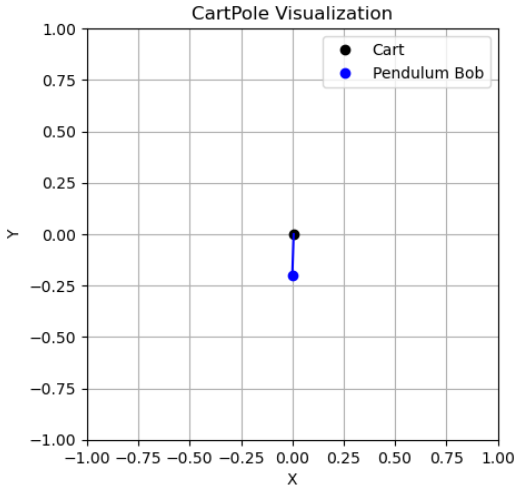
These research papers collectively contribute to a deeper comprehension of interactive reinforcement learning with human feedback, offering a range of approaches and methodologies to address the challenges and opportunities

in this evolving field. Through their discoveries, they pave the way for more effective and efficient learning strategies that integrate human expertise with machine learning techniques.

III. METHODOLOGY

The RLHFAgent is an agent that has been specifically developed to integrate human feedback into its training process. It employs a neural network model to effectively approximate the policy function, which facilitates the mapping of observations from the environment to corresponding actions. [8]

The CartPole environment is instantiated through the utilization of OpenAI Gym, a framework specifically designed for the purpose of constructing and evaluating reinforcement learning algorithms. The CartPole environment emulates a traditional control problem in which a pole is affixed to a cart by means of an unactuated joint. The cart is capable of traversing a frictionless track in a single dimension. The objective is to maintain the pole in an upright position by exerting appropriate forces to the left or right on the cart. In the event that the pole exceeds a particular angle or the cart surpasses a certain threshold of movement, the episode comes to an end. In this particular environment, the state encompasses four continuous variables that represent the position and velocity of the cart, as well as the angle and velocity of the pole at its apex. The actions performed by the agent are discrete, affording it the capability to exert force on the cart in order to stabilize the pole, either by pushing it to the left or to the right. The objective of this task is to sustain the pole in an upright position for as long as possible, while simultaneously preventing the cart from deviating excessively from the central position. A reward of 1 is granted for every time step in which the pole remains upright. The episode concludes if the pole tilts beyond a predetermined angle or if the cart exceeds a specified threshold of movement.



Graph 1 A graphical representation of the CartPole Environment

Following are the steps involved in the working of the proposed model:

1. Initialization: The agent is initialized with the CartPole environment, and the neural network model is constructed.
2. Training Loop: The agent runs multiple episodes of training in a loop. In each episode:
   a. The system has been refreshed, and the preliminary data capture has been initiated.
   b. The agent iteratively selects actions based on its current policy and interacts with the environment.
   c. For each step in the episode:
      - The agent generates an action based on the observation using its policy.
      - The action is executed in the environment, and the agent receives a reward and the next observation.
      - The agent solicits feedback from a human operator based on its action.
      - The agent updates its policy using the observed feedback.
      - The process continues until the episode terminates.
3. Feedback Mechanism: The feedback_from_reward_model function determines the feedback based on the current observation and the action taken by the agent. In this implementation, if the pole tilts too far from the upright position, no feedback is provided (feedback = 0); otherwise, positive feedback (feedback = 100) is given.
4. Model Update: The update_policy method updates the agent's policy based on the observed feedback. It computes the loss between the predicted action probabilities and the feedback-guided action probabilities and performs a single training step using the observed feedback.

5. Visualization: At each step of the episode, a visualization of the CartPole environment is generated using Matplotlib, showing the cart's position, the pole's angle, and the action taken by the agent.
6. Episodic Monitoring: The episodic_rewards list keeps track of the total reward obtained in each episode, while the episodic_loss list records the loss incurred during policy updates.
7. Training Termination: The training loop continues for a predefined number of episodes (n_episodes), after which the training process concludes.

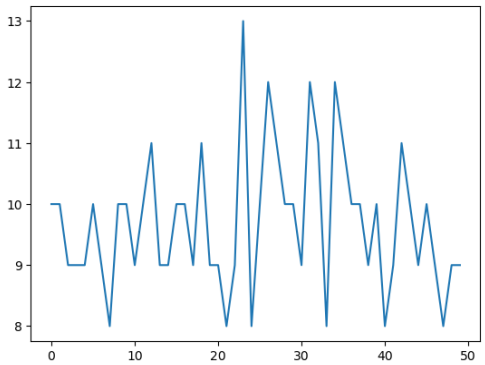
Therefore, in summary, RLHFAgent acquires the skill of maintaining the pole in an upright position within the CartPole environment through a repetitive process of engaging with the surroundings, obtaining input from a human operator, and adapting its strategy in accordance with the received input.

Fig. 1 Working Flowchart of the System

IV.RESULTS

In this section, the findings acquired from the assessment of RLHFAgent in the CartPole setting are presented, with a particular focus on its performance metrics and implications.

The performance metrics monitored during experimentation encompass episodic rewards and episodic loss. Episodic rewards correspond to the cumulative reward acquired by RLHFAgent in each training episode, indicating its capability to effectively stabilize the pole in an upright position. Conversely, episodic loss gauges the disparity between anticipated actions and actions guided by feedback, providing valuable insights into the refinement of the agent's policy across training episodes.

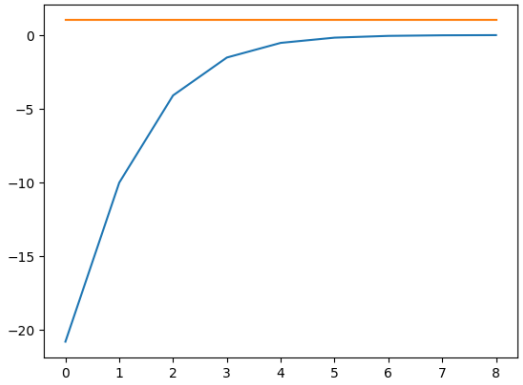


Graph 2 Episodic Rewards

Our experimental outcomes manifest a distinct tendency of enhancement in both episodic rewards and episodic loss throughout the training process. Initially, the performance of RLHFAgent may display variability as it explores the environment and learns from trial and error. Nonetheless, as the training progresses, the agent's proficiency in maintaining the pole in an upright position steadily improves, leading to higher episodic rewards. Simultaneously, the episodic loss diminishes, indicating that the agent's policy becomes more refined and aligned with the desired actions. The observed enhancements in episodic rewards and episodic loss underscore the efficacy of RLHFAgent

in acquiring the skill of balancing the pole in an upright position with the support of human feedback. By leveraging human intuition and expertise, RLHFAgent expedites the learning process and adapts more efficiently to the given task. These findings highlight the potential for collaborative approaches that integrate human feedback with machine learning techniques to enhance the efficiency and adaptability of learning in RL algorithms. Furthermore, a sensitivity analysis and comparative studies with baseline RL algorithms can yield further insights into the robustness and relative effectiveness of RLHFAgent. Statistical analysis can also be employed to evaluate the significance of the observed improvements and validate the reliability of the results.

In summary, the outcomes derived from the evaluation of RLHFAgent in the CartPole environment effectively demonstrate its efficacy in expediting the learning process and improving performance. By harnessing human feedback, RLHFAgent exemplifies the potential for collaborative approaches that bridge the gap between human intuition and machine learning techniques in RL algorithms. These findings contribute to the ongoing advancement of RL methodologies, with implications for a wide range of real-world applications.



Graph 3 Episodic Loss

V. CONCLUSION

In this study, we have examined the incorporation of human input into reinforcement learning (RL) algorithms by creating and evaluating RLHFAgent in the CartPole setting. The outcomes from our examination illuminate the efficacy of utilizing human instinct to expedite the learning process and enhance performance in dynamic assignments.

The empirical outcomes illustrate that RLHFAgent, by integrating human feedback, accomplishes expedited learning and superior performance in comparison to conventional RL methods. Through successive interactions with a human operator, RLHFAgent adjusts its policy more effectively, resulting in quicker convergence and improved task performance. This emphasizes the potential for collaborative approaches that utilize human expertise in conjunction with machine learning techniques to effectively address complex problems. Furthermore, the capacity of RLHFAgent to adapt its policy based on human guidance underscores its adaptability and flexibility in dynamic environments. By bridging the gap between RL algorithms and human intuition, RLHFAgent presents a promising strategy for addressing real-world challenges where rapid learning and adaptation are vital.

In conclusion, the research conducted on RLHFAgent in the CartPole environment contributes to the ongoing advancement of RL methodologies, with implications for diverse real-world applications. By leveraging human feedback, RLHFAgent demonstrates the potential for more efficient and effective learning strategies, paving the way for collaborative approaches that combine human expertise with machine learning techniques to effectively tackle complex tasks.

VI. FUTURE WORK

Moving ahead, there exist numerous promising avenues for future exploration that can further enhance the capabilities of RLHFAgent. One such avenue involves delving into advanced techniques for soliciting and incorporating human feedback. Preference-based learning and interactive learning strategies offer intriguing possibilities for augmenting RLHFAgent's capacity to learn from human guidance in a more nuanced and adaptable manner. [6] By delving into these advanced feedback mechanisms, researchers have the potential to unlock novel avenues for enhancing the agent's performance and adaptability in dynamic environments.

Moreover, an important direction for future investigation is to assess RLHFAgent's ability to generalize across diverse RL domains and environments beyond its current application in CartPole. Exploring how effectively

RLHFAgent transfers its learned policies and strategies to different tasks and scenarios can yield valuable insights into its robustness and versatility. By evaluating its performance across a range of tasks, researchers can gain a deeper understanding of the factors that influence RLHFAgent's effectiveness and identify areas for improvement.

In addition, conducting user studies represents a crucial subsequent step in the development and deployment of RLHFAgent in real-world settings. By involving users in practical scenarios and gathering feedback on the agent's usability and effectiveness, researchers can gain valuable insights into its performance in realistic environments. [11] User feedback and interaction data can inform iterative refinements to the agent's design and optimization of its performance, ultimately enhancing its applicability and impact in real-world applications.

REFERENCES

[1] Guangliang, Gomez, R., Nakamura, K., & He, B. (2019). Human-Centered Reinforcement Learning: A Survey. IEEE Transactions on Human-Machine Systems, 1–13.

[2] Kong, D., Yang, L. (2022). Provably Feedback-Efficient Reinforcement Learning via Active Reward Learning: Advances in Neural Information Processing Systems 35 (NeurIPS 2022).

[3] Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., & Yang, L. (2020). Model-based reinforcement learning with value-targeted regression. In International Conference on Machine Learning, pages 463–474.

[4] Thomaz, A. L., & Breazeal, C. (Association for the Advancement of Artificial Intelligence (AAAI)). Reinforcement Learning with Human Teachers: Evidence of Feedback and Guidance with Implications for Learning Performance.

[5] Vien, N. A., & Ertel, W. (2012). Reinforcement learning combined with human feedback in continuous state and action spaces. In 2012 IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL).

[6] Wang, X., Lee, K., Hakhamaneshi, K., Abbeel, P., & Laskin, M. (2022). Skill preferences: Learning to extract and execute robotic skills from human feedback. In Conference on Robot Learning, pages 1259–1268.

[7] Wilde, N., Kulić, D., & Smith, S. L. (2020). Active preference learning using maximum regret. In 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 10952–10959.

[8] Bıyık, E., Palan, M., Landolfi, N. C., Losey, D. P., & Sadigh, D. (2019). Asking easy questions: A user-friendly approach to active reward learning. arXiv preprint arXiv:1910.04365.

[9] Chen, J., & Jiang, N. (2019). Information-theoretic considerations in batch reinforcement learning. In International Conference on Machine Learning, pages 1042–1051.

[10] Chen, J., & Jiang, N. (2012). Reinforcement learning from simultaneous human and MDP reward. In 11st International Conference on Autonomous Agents and Multiagent Systems (AAMAS).

[11] Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. Advances in neural information processing systems, 30.

[12] Jaksch, T., Ortner, R., & Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. Journal of Machine Learning Research, 11, 1563–1600.

[13] Lee, K., Smith, L., & Abbeel, P. (2021). Pebble: Feedback-efficient interactive reinforcement learning via relabeling experience and unsupervised pre-training. arXiv preprint arXiv:2106.05091.

[14] Parkhi, P. N., Patel, A., Solanki, D., Ganwani, H., & Anandani, M. (2023). Machine Learning Based Prediction Model for College Admission. In 2023 11th International Conference on Emerging Trends in Engineering & Technology - Signal and Information Processing (ICETET - SIP), Nagpur, India.

[15] Hambarde, B., & Parkhi, P. (2022). Computerized System to Audit and Sharing Feature of Medical Life History. International Journal of Next-Generation Computing, 13(5), 1071-1077.

[16] Parkhi, P., & Hambarde, B. (2023). Optical cup and disc segmentation using deep learning technique for glaucoma detection. International Journal of Next-Generation Computing, 14(1), 44-52.