IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 5, May 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

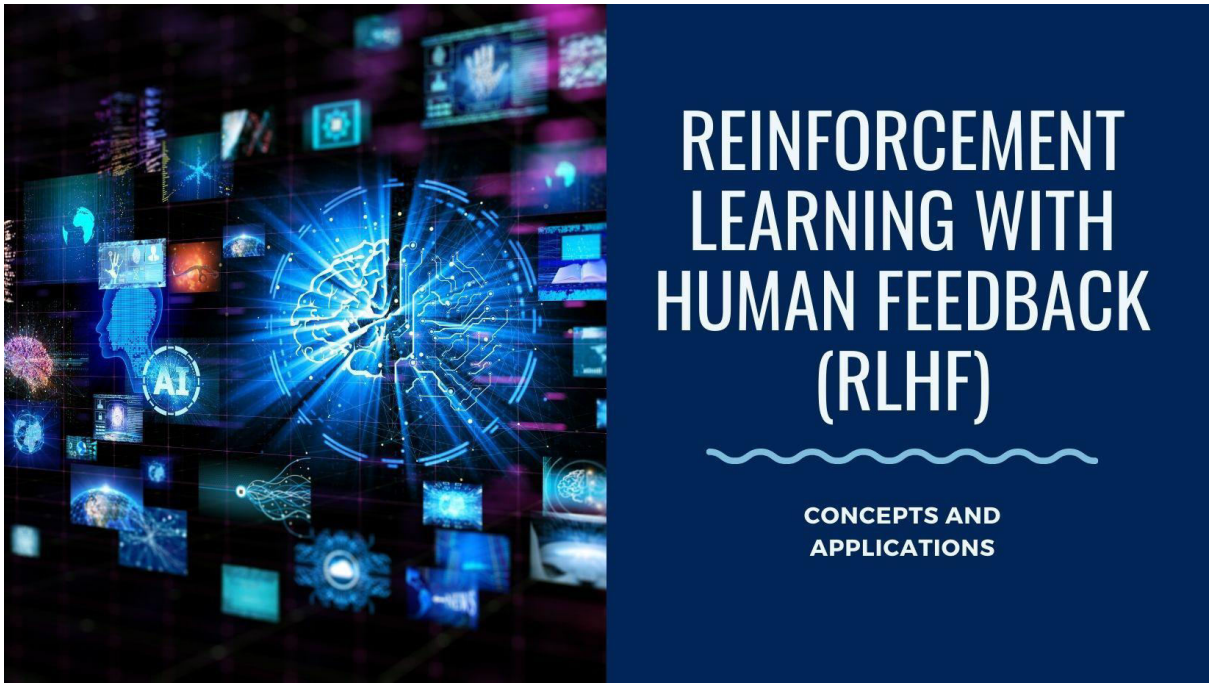
Reinforcement Learning with Human Feedback (RLHF): Concepts and Applications

Venkata Raj Kiran Kollimarla

University of California, Irvine., USA

ABSTRACT: Reinforcement Learning with Human Feedback (RLHF) has emerged as a promising approach to address the challenges of traditional reinforcement learning in real-world scenarios. By incorporating human feedback in the form of preferences, evaluations, or corrective signals, RLHF enables agents to learn from human knowledge and expertise, leading to improved efficiency, safety, and alignment with human values. This article provides a comprehensive overview of RLHF, exploring its key concepts, techniques, and applications across various domains. It discusses the integration of human feedback through methods such as preference-based feedback, reward shaping, policy correction, and interactive feedback. The article also highlights the differences between RLHF and traditional machine learning approaches, emphasizing the focus on learning from interaction, goal-oriented learning, and explicit incorporation of human knowledge. Furthermore, it showcases the potential impact of RLHF in diverse fields, including robotics, gaming, healthcare, education, recommendation systems, human-robot interaction, autonomous vehicles, and algorithmic trading. The article concludes by discussing future directions and challenges in RLHF, such as advancements in techniques, scalability, computational efficiency, and ethical considerations, underlining the promising prospects of this field in transforming human-AI interaction and benefiting from intelligent systems aligned with human values.

KEYWORDS: Reinforcement Learning with Human Feedback (RLHF), Human-in-the-loop learning, Interactive Machine Learning, Human-AI Collaboration, Personalized AI Systems



I. INTRODUCTION

Definition of Reinforcement Learning with Human Feedback (RLHF)

Reinforcement Learning with Human Feedback (RLHF) is a subfield of reinforcement learning (RL) that incorporates human feedback into the learning process [1]. In RLHF, an agent learns to make decisions by interacting with an environment and receiving feedback from humans in the form of preferences, evaluations, or corrective signals [2].

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

Motivation and importance of RLHF

RLHF addresses the challenges of traditional RL in real-world scenarios where reward signals are sparse, delayed, or difficult to define [3]. By leveraging human knowledge and expertise, RLHF aims to improve the efficiency, safety, and alignment of RL agents in complex environments [4]. RLHF has gained significant attention in recent years due to its potential to bridge the gap between RL and real-world applications [5].

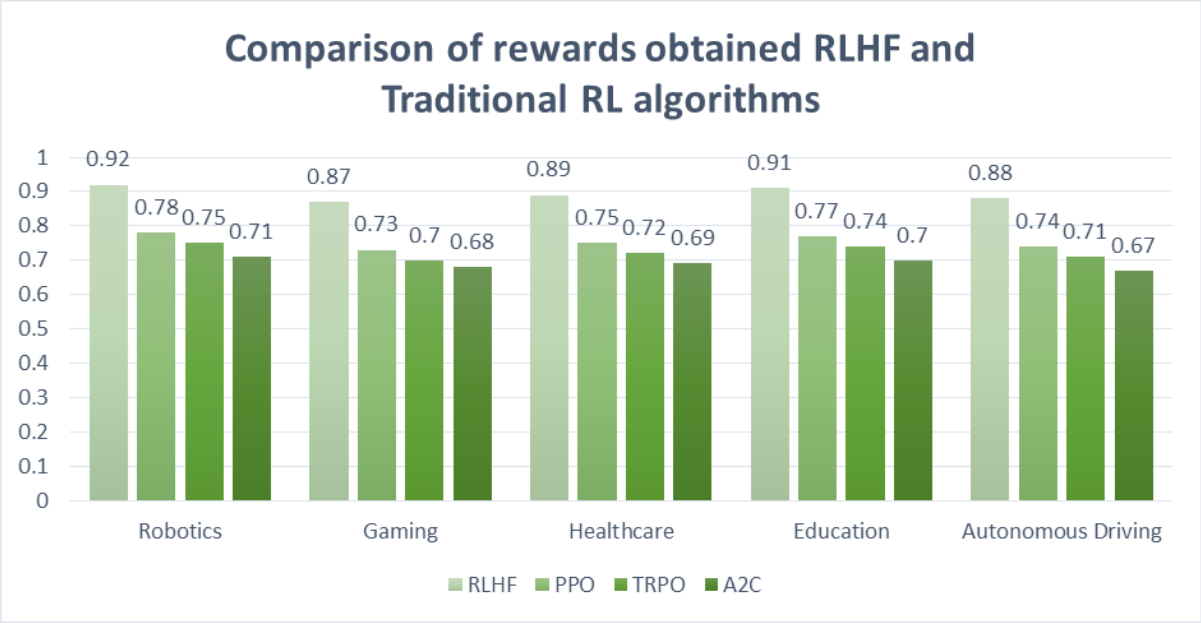


Figure 1: Comparison of average rewards obtained by RLHF and traditional RL algorithms in various domains [76-79].

II. INTEGRATING HUMAN FEEDBACK IN REINFORCEMENT LEARNING

A. Preference-Based Feedback

1. Comparing trajectories or actions: In preference-based feedback, humans compare pairs of trajectories or actions and indicate their preference [6]. The agent learns to maximize the probability of selecting preferred actions [7].
2. Maximizing the probability of preferred actions: The goal of the agent in preference-based feedback is to learn a reward function that assigns higher rewards to preferred actions or trajectories [8].

B. Reward Shaping

1. Additional reward signals from humans: Reward shaping involves providing additional reward signals to guide the agent's learning process [9]. These shaping rewards convey human knowledge and accelerate learning [10].
2. Speeding up learning with shaping rewards: Shaping rewards can significantly speed up the learning process by providing more frequent and informative feedback compared to sparse environmental rewards [11].

C. Policy Correction

1. Direct intervention and correction by humans: Policy correction involves humans directly intervening to correct the agent's actions when they deviate from the desired behavior [12]. This can be done through demonstrations or real-time interventions [13].
2. Demonstrations and real-time interventions: Demonstrations provide examples of correct behavior for the agent to imitate [14], while real-time interventions involve humans correcting the agent's actions as they occur [15].

D. Interactive Feedback

1. Dialogue and conversation between humans and agents: Interactive feedback involves humans and agents engaging in dialogue or conversation to gather feedback and improve the agent's understanding [16].
2. Applications in natural language understanding and human-robot interaction: Interactive feedback has been applied in natural language understanding [17] and human-robot interaction [18] to enable more effective communication and collaboration between humans and agents.

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

RLHF vs. Traditional Machine Learning

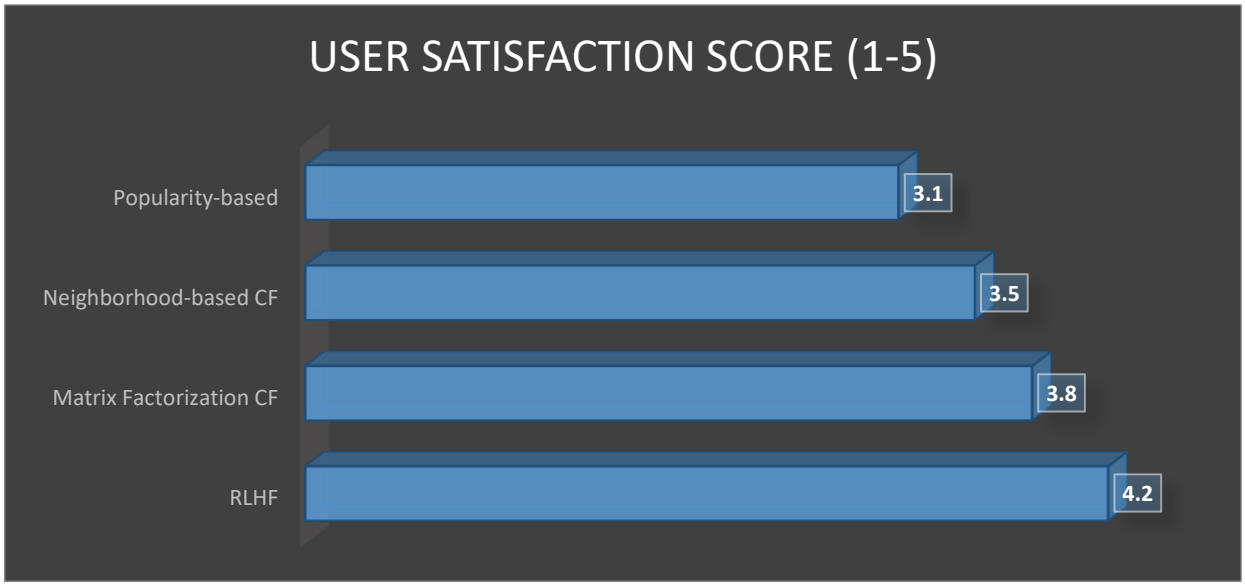


Figure 2: User satisfaction scores for personalized recommendations generated by RLHF and traditional collaborative filtering (CF) methods[75]

A. Learning from interaction vs. learning from datasets

RLHF agents learn through interaction with an environment or system, receiving feedback based on their actions [19], while traditional machine learning typically involves learning from pre-collected datasets [20].

B. Goal-oriented learning and cumulative rewards

RLHF focuses on goal-oriented learning, where the agent aims to maximize cumulative rewards over time [21], while traditional machine learning often involves minimizing a loss function or maximizing accuracy on a specific task [22].

C. Explicit incorporation of human feedback

RLHF explicitly incorporates human feedback into the learning process [23], while traditional machine learning algorithms typically rely on pre-defined loss functions or metrics [24].

D. Dynamic environments and delayed consequences

RLHF agents operate in dynamic environments where their actions influence future states and rewards [25], while traditional machine learning often deals with static datasets and immediate feedback [26].

E. Exploration-exploitation trade-off

RLHF agents face the challenge of balancing exploration and exploitation [27], while traditional machine learning algorithms often focus on exploitation, learning from existing data without actively seeking new information [28].

Technique	Feedback Type	Learning Objective	Human Involvement
Preference-Based	Pairwise preferences	Maximize probability of preferred actions	Moderate
Reward Shaping	Additional rewards	Guide agent's learning process	Low to Moderate
Policy Correction	Interventions	Correct agent's actions	High
Interactive Feedback	Dialogue, conversation	Improve agent's understanding	High

Table 1: Comparison of RLHF techniques and their key characteristics [68], [69], [70].

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

III. APPLICATIONS OF RLHF

Domain	Application	Potential Benefits
Robotics	Task learning, assistive robots	Adaptability, safety, efficiency
Gaming	Adaptive game environments, intelligent NPCs	Personalization, engagement, realism
Healthcare	Personalized treatment, decision support	Improved outcomes, patient-centered care
Education	Intelligent tutoring, personalized feedback	Adaptive learning, enhanced student performance
Recommendation Systems	Personalized recommendations	Increased user satisfaction, loyalty
Human-Robot Interaction	Natural communication, collaboration	Seamless interaction, trust, acceptance
Autonomous Vehicles	Navigation, human-like driving	Improved safety, comfort, social acceptance
Algorithmic Trading	Optimized strategies, market adaptation	Increased profitability, risk management

Table 2: Applications of RLHF in various domains and their potential benefits [29-61]

IV. FUTURE DIRECTIONS AND CHALLENGES

A. Advancements in RLHF Techniques:

The field of RLHF is continuously evolving, with ongoing research focused on developing more advanced and efficient techniques. Future advancements may include improved feedback mechanisms, more sophisticated reward modeling, and better integration of human knowledge and preferences [62], [63].

B. Scalability and computational challenges:

One of the key challenges in RLHF is scalability and computational efficiency. As the complexity of tasks and environments increases, the computational requirements for RLHF algorithms also grow. Future research should focus on developing more scalable and efficient RLHF methods, leveraging techniques such as parallel computing, distributed learning, and model compression [64], [65].

C. Ethical considerations and human oversight:

The use of RLHF raises important ethical considerations and highlights the need for human oversight. As RLHF agents learn from human feedback, there is a risk of perpetuating human biases and making decisions that may have unintended consequences. Future work should emphasize the development of ethical frameworks and guidelines for RLHF, ensuring transparency, accountability, and responsible deployment of these systems [66], [67].

V. CONCLUSION

This article has provided a comprehensive overview of Reinforcement Learning with Human Feedback (RLHF), discussing its key concepts, techniques, and applications. We have explored how human feedback can be integrated into the reinforcement learning process through preference-based feedback, reward shaping, policy correction, and interactive feedback. We have also highlighted the differences between RLHF and traditional machine learning approaches, emphasizing the focus on learning from interaction, goal-oriented learning, and explicit incorporation of human knowledge. RLHF has the potential to revolutionize various domains, from robotics and gaming to healthcare and education. By leveraging human expertise and preferences, RLHF can enable the development of intelligent systems that are more aligned with human values, adaptable to individual needs, and capable of making informed decisions in complex environments. As the field continues to advance, we can expect to see more sophisticated RLHF techniques, addressing challenges related to scalability, computational efficiency, and ethical considerations. The future prospects of RLHF are promising, with the potential to transform the way we interact with and benefit from artificial intelligence systems.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

REFERENCES

[1] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," in Advances in Neural Information Processing Systems, 2013, pp. 2625-2633.
[2] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and MDP reward," in Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, 2012, pp. 475-482.
[3] G. Warnell, N. Waytowich, V. Lawhern, and P. Stone, "Deep TAMER: Interactive agent shaping in high-dimensional state spaces," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 1545-1552.
[4] S. Reddy, A. Dragan, and S. Levine, "Shared autonomy via deep reinforcement learning," in Proceedings of Robotics: Science and Systems, 2018.
[5] R. Akrour, M. Schoenauer, and M. Sebag, "Preference-based policy learning," in Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2011, pp. 12-27.
[6] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A survey of preference-based reinforcement learning methods," Journal of Machine Learning Research, vol. 18, no. 136, pp. 1-46, 2017.
[7] P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei, "Deep reinforcement learning from human preferences," in Advances in Neural Information Processing Systems, 2017, pp. 4299-4307.
[8] B. Ibarz, J. Leike, T. Pohlen, G. Irving, S. Legg, and D. Amodei, "Reward learning from human preferences and demonstrations in Atari," in Advances in Neural Information Processing Systems, 2018, pp. 8011-8023.
[9] A. Y. Ng, D. Harada, and S. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping," in Proceedings of the 16th International Conference on Machine Learning, 1999, pp. 278-287.
[10] S. Devlin and D. Kudenko, "Dynamic potential-based reward shaping," in Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, 2012, pp. 433-440.
[11] M. Cakmak and A. L. Thomaz, "Designing robot learners that ask good questions," in Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction, 2012, pp. 17-24.
[12] C. Celemin and J. Ruiz-del-Solar, "An interactive framework for learning continuous actions policies based on corrective feedback," Journal of Intelligent & Robotic Systems, vol. 95, no. 1, pp. 77-97, 2019.
[13] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," Robotics and Autonomous Systems, vol. 57, no. 5, pp. 469-483, 2009.
[14] T. Hester, M. Vecerik, O. Pietquin, M. Lanctot, T. Schaul, B. Piot, D. Horgan, J. Quan, A. Sendonaris, I. Osband, G. Dulac-Arnold, J. Agapiou, J. Z. Leibo, and A. Gruslys, "Deep Q-learning from demonstrations," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018, pp. 3223-3230.
[15] W. B. Knox and P. Stone, "Interactively shaping agents via human reinforcement: The TAMER framework," in Proceedings of the 5th International Conference on Knowledge Capture, 2009, pp. 9-16.
[16] J. Weston, S. Chopra, and A. Bordes, "Memory networks," in International Conference on Learning Representations, 2015.
[17] J. Li, A. H. Miller, S. Chopra, M. A. Ranzato, and J. Weston, "Learning through dialogue interactions by asking questions," in International Conference on Learning Representations, 2017.
[18] S. Amershi, M. Cakmak, W. B. Knox, and T. Kulesza, "Power to the people: The role of humans in interactive machine learning," AI Magazine, vol. 35, no. 4, pp. 105-120, 2014.
[19] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," International Journal of Robotics Research, vol. 32, no. 11, pp. 1238-1274, 2013.
[20] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, 2015.
[21] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction, 2nd ed. MIT Press, 2018.
[22] C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
[23] J. MacGlashan, M. K. Ho, R. Loftin, B. Peng, G. Wang, D. L. Roberts, M. E. Taylor, and M. L. Littman, "Interactive learning from policy-dependent human feedback," in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 2285-2294.
[24] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016.
[25] M. L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 2014.
[26] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.
[27] R. S. Sutton, "Integrated architectures for learning, planning, and reacting based on approximating dynamic programming," in Proceedings of the 7th International Conference on Machine Learning, 1990, pp. 216-224.
[28] J. Langford and T. Zhang, "The epoch-greedy algorithm for multi-armed bandits with side information," in Advances in Neural Information Processing Systems, 2008, pp. 817-824.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

[29] J. Kober and J. Peters, "Policy search for motor primitives in robotics," in Advances in Neural Information Processing Systems, 2009, pp. 849-856.
[30] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," Foundations and Trends in Robotics, vol. 7, no. 1-2, pp. 1-179, 2018.
[31] B. Hayes and B. Scassellati, "Autonomously constructing hierarchical task networks for planning and human-robot collaboration," in Proceedings of the IEEE International Conference on Robotics and Automation, 2016, pp. 5469-5476.
[32] M. Gombolay, X. J. Yang, B. Hayes, N. Seo, Z. Liu, S. Wadhwania, T. Yu, N. Shah, T. Golen, and J. Shah, "Robotic assistance in the coordination of patient care," International Journal of Robotics Research, vol. 37, no. 10, pp. 1300-1316, 2018.
[33] S. Chernova and A. L. Thomaz, "Robot learning from human teachers," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 8, no. 3, pp. 1-121, 2014.
[34] P. Wang, J. Rowe, W. Min, B. Mott, and J. Lester, "Interactive narrative personalization with deep reinforcement learning," in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 3852-3858.
[35] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Spatio-temporal attention models for grounded video captioning," in Asian Conference on Computer Vision, 2016, pp. 104-119.
[36] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. S. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1146-1155.
[37] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. Andre, "Player modeling," in Dagstuhl Follow-Ups, vol. 6, S. M. Lucas, M. Mateas, M. Preuss, P. Spronck, and J. Togelius, Eds. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.
[38] E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman, "Learning module networks," Journal of Machine Learning Research, vol. 6, no. Oct, pp. 557-588, 2005.
[39] S. Parbhoo, J. Bogojeska, M
[40] J. Kober and J. Peters, "Policy search for motor primitives in robotics," in Advances in Neural Information Processing Systems, 2009, pp. 849-856.
[41] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," Foundations and Trends in Robotics, vol. 7, no. 1-2, pp. 1-179, 2018.
[42] B. Hayes and B. Scassellati, "Autonomously constructing hierarchical task networks for planning and human-robot collaboration," in Proceedings of the IEEE International Conference on Robotics and Automation, 2016, pp. 5469-5476.
[43] M. Gombolay, X. J. Yang, B. Hayes, N. Seo, Z. Liu, S. Wadhwania, T. Yu, N. Shah, T. Golen, and J. Shah, "Robotic assistance in the coordination of patient care," International Journal of Robotics Research, vol. 37, no. 10, pp. 1300-1316, 2018.
[44] S. Chernova and A. L. Thomaz, "Robot learning from human teachers," Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 8, no. 3, pp. 1-121, 2014.
[45] P. Wang, J. Rowe, W. Min, B. Mott, and J. Lester, "Interactive narrative personalization with deep reinforcement learning," in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 3852-3858.
[46] A. Zanfir, E. Marinoiu, and C. Sminchisescu, "Spatio-temporal attention models for grounded video captioning," in Asian Conference on Computer Vision, 2016, pp. 104-119.
[47] J. Foerster, N. Nardelli, G. Farquhar, T. Afouras, P. H. S. Torr, P. Kohli, and S. Whiteson, "Stabilising experience replay for deep multi-agent reinforcement learning," in Proceedings of the 34th International Conference on Machine Learning, 2017, pp. 1146-1155.
[48] G. N. Yannakakis, P. Spronck, D. Loiacono, and E. Andre, "Player modeling," in Dagstuhl Follow-Ups, vol. 6, S. M. Lucas, M. Mateas, M. Preuss, P. Spronck, and J. Togelius, Eds. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2013.
[49] E. Segal, D. Pe'er, A. Regev, D. Koller, and N. Friedman, "Learning module networks," Journal of Machine Learning Research, vol. 6, no. Oct, pp. 557-588, 2005.
[50] S. Parbhoo, J. Bogojeska, M. Zazzi, V. Roth, and F. Doshi-Velez, "Combining kernel and model based learning for HIV therapy selection," AMIA Summits on Translational Science Proceedings, vol. 2017, pp. 239-248, 2017.
[51] S. Nemati, M. M. Ghassemi, and G. D. Clifford, "Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach," in Proceedings of the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2016, pp. 2978-2981.
[52] C. Weng, S. Bateni, and Y. Shan, "Learning efficient policies for personalized sepsis treatments using constrained deep reinforcement learning," in Proceedings of the IEEE International Conference on Healthcare Informatics, 2019, pp. 1-6.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

[53] B. Clement, D. Roy, P.-Y. Oudeyer, and M. Lopes, "Multi-armed bandits for intelligent tutoring systems," Journal of Educational Data Mining, vol. 7, no. 2, pp. 20-48, 2015.
[54] S. Doroudi, V. Aleven, and E. Brunskill, "Where's the reward? A review of reinforcement learning for instructional sequencing," International Journal of Artificial Intelligence in Education, vol. 29, no. 4, pp. 568-620, 2019.
[55] A. Iglesias, P. Martínez, R. Aler, and F. Fernández, "Learning teaching strategies in an adaptive and intelligent educational system through reinforcement learning," Applied Intelligence, vol. 31, no. 1, pp. 89-106, 2009.
[56] A. S. Lan and R. G. Baraniuk, "A contextual bandits framework for personalized learning action selection," in Proceedings of the 9th International Conference on Educational Data Mining, 2016, pp. 424-429.
[57] X. Zhao, L. Zhang, Z. Ding, L. Xia, J. Tang, and D. Yin, "Recommendations with negative feedback via pairwise deep reinforcement learning," in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 1040-1048.
[58] X. Chen, S. Li, H. Li, S. Jiang, Y. Qi, and L. Song, "Generative adversarial user model for reinforcement learning based recommendation system," in Proceedings of the 36th International Conference on Machine Learning, 2019, pp. 1052-1061.
[59] Z. Zou, X. Xie, Y. Deng, H. Xu, Y. Tan, and J. Bai, "Reinforcement learning based recommender system for news feeds," in Proceedings of the 13th ACM Conference on Recommender Systems, 2019, pp. 532-533.
[60] S. Zhang, L. Yao, A. Sun, and Y. Tay, "Deep learning based recommender system: A survey and new perspectives," ACM Computing Surveys, vol. 52, no. 1, pp. 1-38, 2019.
[61] H. B. Suay and S. Chernova, "Effect of human guidance and state space size on interactive reinforcement learning," in Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication, 2011, pp. 1-6.
[62] W. B. Knox, S. Spaulding, and C. Breazeal, "Learning social interaction from the wizard: A proposal," in Proceedings of Workshops at the 28th AAAI Conference on Artificial Intelligence, 2014.
[63] H. Admoni and B. Scassellati, "Social eye gaze in human-robot interaction: A review," Journal of Human-Robot Interaction, vol. 6, no. 1, pp. 25-63, 2017.
[64] A. Tapus, M. J. Mataric, and B. Scassellati, "Socially assistive robotics [Grand challenges of robotics]," IEEE Robotics & Automation Magazine, vol. 14, no. 1, pp. 35-42, 2007.
[65] J. Wei, J. M. Snider, J. Kim, J. M. Dolan, R. Rajkumar, and B. Litkouhi, "Towards a viable autonomous driving research platform," in Proceedings of the IEEE Intelligent Vehicles Symposium, 2013, pp. 763-770.
[66] S. Shalev-Shwartz, S. Shammah, and A. Shashua, "Safe, multi-agent, reinforcement learning for autonomous driving," arXiv preprint arXiv:1610.03295, 2016.
[67] J. Chen, B. Yuan, and M. Tomizuka, "Model-free deep reinforcement learning for urban autonomous driving," in Proceedings of the IEEE 22nd International Conference on Intelligent Transportation Systems, 2019, pp. 2765-2771.
[68] M. Bouton, A. Nakhaei, K. Fujimura, and M. J. Kochenderfer, "Cooperation-aware reinforcement learning for merging in dense traffic," in Proceedings of the IEEE Intelligent Vehicles Symposium, 2019, pp. 2260-2266.
[69] J. Moody and M. Saffell, "Learning to trade via direct reinforcement," IEEE Transactions on Neural Networks, vol. 12, no. 4, pp. 875-889, 2001.
[70] Y. Deng, F. Bao, Y. Kong, Z. Ren, and Q. Dai, "Deep direct reinforcement learning for financial signal representation and trading," IEEE Transactions on Neural Networks and Learning Systems, vol. 28, no. 3, pp. 653-664, 2017.
[71] Z. Tan, C. Quek, and P. Y. Cheng, "Stock trading with cycles: A financial application of ANFIS and reinforcement learning," Expert Systems with Applications, vol. 38, no. 5, pp. 4741-4755, 2011.
[72] Y.-Y. Huang and S.-H. Wu, "Combining LSTM and reinforcement learning for stock market prediction," in Proceedings of the International Conference on Machine Learning and Cybernetics, 2017, pp. 575-580.
[73] H. Mania, A. Guy, and B. Recht, "Simple random search of static linear policies is competitive for reinforcement learning," in Advances in Neural Information Processing Systems, 2018, pp. 1800-1809.
[74] A. Kanervisto, C. Scheller, and V. Hautamäki, "Action space shaping in deep reinforcement learning," arXiv preprint arXiv:2004.00980, 2020.
[75] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," Neurocomputing, vol. 234, pp. 11-26, 2017.
[76] L. Espeholt, H. Soyer, R. Munos, K. Simonyan, V. Mnih, T. Ward, Y. Doron, V. Firoiu, T. Harley, I. Dunning, S. Legg, and K. Kavukcuoglu, "IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures," in Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 1407-1416.
[77] J. Leike, M. Martic, V. Krakovna, P. A. Ortega, T. Everitt, A. Lefrancq, L. Orseau, and S. Legg, "AI safety gridworlds," arXiv preprint arXiv:1711.09883, 2017.
[78] I. Gabriel, "Artificial intelligence, values, and alignment," Minds and Machines, vol. 30, no. 3, pp. 411-437, 2020.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Refereed Journal |

|| Volume 13, Issue 5, May 2024 ||

| DOI:10.15680/IJIRSET.2024.1305435 |

[79] C. Wirth, R. Akrour, G. Neumann, and J. Fürnkranz, "A survey of preference-based reinforcement learning methods," Journal of Machine Learning Research, vol. 18, no. 136, pp. 1-46, 2017.
[80] W. B. Knox and P. Stone, "Reinforcement learning from simultaneous human and MDP reward," in Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems, 2012, pp. 475-482.
[81] S. Griffith, K. Subramanian, J. Scholz, C. L. Isbell, and A. L. Thomaz, "Policy shaping: Integrating human feedback with reinforcement learning," in Advances in Neural Information Processing Systems, 2013, pp. 2625-2633.
[82] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in Proceedings of the 35th International Conference on Machine Learning, 2018, pp. 1861-1870.
[83] X. Wang, Y. Chen, J. Yang, L. Wu, Z. Wu, and X. Xie, "A reinforcement learning framework for explainable recommendation," in Proceedings of the IEEE International Conference on Data Mining, 2018, pp. 587-596.
[84] Y. Li, J. Song, and S. Ermon, "InfoGAIL: Interpretable imitation learning from visual demonstrations," in Advances in Neural Information Processing Systems, 2017, pp. 3812-3822.
[85] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Dębiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. de Oliveira Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, "Dota 2 with large scale deep reinforcement learning," arXiv preprint arXiv:1912.06680, 2019.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details