Beyond Performance: Explaining Non-Intuitive Deep Reinforcement Learning Actions in Elastic Optical Networks

Louma Mehyeddine[(1)], Carlos Natalino[(2)], Alaa Amro[(1)], Jean Pierre Asdikian[(3)], Ihab Sbeity[(1)], Guido Maier[(3)], Paolo Monti[(2)], Sebastian Troia[(3)], Omran Ayoub[(4)]

[(1)] Lebanese University, Lebanon [(2)] Chalmers University of Technology, Sweden, [(3)] Politecnico di Milano, Italy, [(4)] University of Applied Sciences and Arts of Southern Switzerland, Switzerland

Abstract *We develop a Deep Reinforcement Learning (DRL) agent for the RMSA problem and improve the Shapley Value for Explaining Reinforcement Learning (SVERL) explainability framework by integrating policy sensitivity and feature interdependence for the RMSA problem. We then explain the proactive rejection of lightpath requests. © 2025 The Author(s)*

Introduction

As optical networks grow in scale and complexity[1], the demand for intelligent control mechanisms has never been greater. Reinforcement Learning (RL) has emerged as a powerful solution to this challenge[1]–[4]. In particular, Deep Reinforcement Learning (DRL) models have shown great promise in tackling the RMSA problem by learning policies that dynamically allocate spectrum and route lightpaths to minimize blocking probability and maximize throughput[5]–[9]. Despite their advantages, DRL models often behave as *black boxes*, making decisions that are potentially optimized from a theoretical standpoint but difficult for human operators to interpret. For instance, an RL agent may reject a lightpath request even when sufficient resources are available, or select a longer route over a seemingly better one. While such actions may serve a long-term optimization goal, their logic remains opaque, raising concerns around trust, accountability, and operational insight[10]–[12].

This work presents an explainable DRL-based framework to extract actionable insights into the agent's learned policies and behavior in the RMSA context. We focus on interpreting the agent's decision-making process, particularly counter-intuitive actions such as *proactive rejection*, where a request is denied despite sufficient resources.

To achieve this, we first develop and implement a DRL agent for the RMSA problem. To interpret the agent's decisions, we then adapt an existing RL-specific explainability framework, Shapley Value for Explaining Reinforcement Learning (SVERL)[13], to quantify features' contributions to the probability of the DRL agent taking any of the actions. Specifically, SVERL addresses two key limitations of existing XRL methods: *(i) Policy-awareness*: Traditional feature attribution methods often ignore how the DRL agent's policy changes when certain input features are hidden. SVERL allows to re-evaluate the agent's policy after masking each feature, ensuring that the resulting explanations reflect how the agent would actually behave with incomplete information. For instance, if the link utilization of a specific fiber is hidden, our adaptation of SVERL evaluates the new action the DRL agent would take, rather than assuming the original action remains unchanged; *(ii) Feature dependency*: Input features in RMSA, such as link utilization, modulation formats, and path lengths, are often interdependent. Masking one feature (e.g., available bandwidth) can affect the interpretation of another (e.g., modulation feasibility). Our adaptation of SVERL accounts for these dependencies by modeling their joint influence on the DRL agent's decisions, preventing misleading attributions that arise when features are treated as independent. The results show that the explanations reveal strategic decisions for proactive rejections, but the conclusions are topology-dependent.

Related Work

Recent works explore the interpretability of RL agents in RWA and RMSA problems. For instance, the behavior of DRL agents has been visualized through the distribution of allocated services across spectrum and links, revealing high-level policy trends but offering limited attribution to specific input features or internal reasoning[3]. Multi-objective DRL frameworks have been proposed to highlight trade-offs between throughput, transmitter usage, and availability. However, their outcome-level Pareto analyses do not explain how individual state features influence decisions[14]. Graph-based approaches incorporating attention mechanisms, such as Graph Attention Networks, provide implicit indicators of influential components but lack clarity on causal feature-policy relationships[15]. Other work has studied the effect of modifying the observation space to identify key state features without directly explaining their

role in agent decision-making[16]. Surrogate models combined with SHAP have also been used to interpret the importance of features. Still, these explanations are decoupled from the actual policy and overlook the sequential, probabilistic nature of DRL[17]. In contrast, our work builds on the SVERL framework[13] to directly attribute feature importance by analyzing how targeted feature masking affects the agent's action probabilities. This approach preserves the original policy, captures feature dependencies, and enables more faithful, nuanced explanations to explain non-intuitive behavior, such as *proactive rejection*.

Framework and Methodology

Our framework integrates three primary components: the EON representing the environment, a DRL Agent, and an Explainer. The DRL agent is responsible for the RMSA decisions, while the explainer clarifies the agent's actions by highlighting the influence of features on the agent's policies.

RL Agent. We employed a Proximal Policy Optimization (PPO)[18] RL agent to address a modified version of the RMSA problem within the *DeepRMSA* environment introduced in[19]. The *DeepRMSA* environment provides a comprehensive modeling of the network topology, defines the observation and action spaces, and specifies the reward function to guide the agent's learning process. The observation contains the following features: *(0)* bit_rate_gbps, *(1)* source and (*node_1*), *(2)* target node (*node_2*), for path $x \in 0..k-1$ *(3)* starting index for the request (*initial_index*_x), *(4)* number of required slots (*req_slots*_x), *(5)* number of free slots in the first free spectrum block (*free_block*_x), *(6)* total number of free slots across path x (*path_free_slots*_x), and *(7)* average number of slots across all free spectrum blocks across path x (*avg_free_slots*_x). We adopt k=3 shortest paths, resulting in 18 features. The action space comprises $k+1$ alternatives corresponding to selecting one of the k shortest paths or rejecting the request. The reward function assumes a value of 1 if the request is successfully provisioned, 0 otherwise. The PPO agent was implemented using *Stable Baselines3*[20], with a policy neural network with five layers, each one with 128 neurons. Key hyperparameters for the agent included a discount factor $\gamma = 0.95$, a learning rate of 10^{-4}, and a value function loss coefficient (`vf_coef`) of 0.1. This design enables the agent to learn an effective policy for dynamically selecting routes and allocating spectrum resources in optical networks based on the current network state and service demands.

Explainer. We adapt SVERL for use with DRL agents. SVERL explains decisions by evaluating the impact of masking input features on agent performance, accounting for how such masking may alter the learned policy and expected return. We introduce several key modifications to SVERL. Since DRL policies encoded in neural networks are not directly interpretable, we first execute the trained agent over multiple episodes to collect state distributions under a converged policy, storing full environment states to support feature masking and consistent environment reinitialization for reproducibility. We then derive action probabilities from the actor network and value estimates from the critic, building explicit policy and value tables for these states. SVERL's operations were adapted for tensor-based Shapley computations, and we format outputs as SHAP-compatible objects to enable visual explanations. To extract explanations, we track the visited states of the DRL agent during and after training, then analyze the cases of *proactive rejection*. The collected states and their policies are input to SVERL, which computes the conditioned policy for each possible coalition.

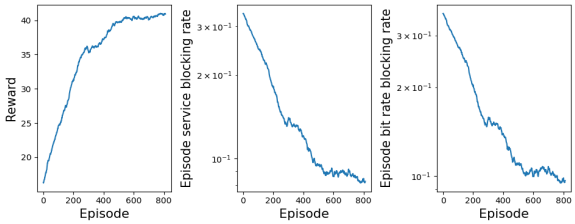

Fig. 1: Reward and blocking rates achieved by the RL agent.

Experimental Results and Interpretations

We consider two network instances. The first is the NSF topology[2], with 320 frequency slots in each link, and a non-uniform traffic profile. The second one is the German topology[21] with 320 frequency slots and a uniform traffic profile.

Figure 1 depicts the learning curves of the DRL agent over the NSF network topology. The plot shows a steady reward increase, demonstrating successful learning over time, reaching convergence after approximately 700 episodes. The blocking rate curves exhibit a consistent downward trend until reaching a stable, low level after about 800 episodes. These results suggest that the agent can efficiently manage spectrum allocation and routing decisions.

Explanations. Figures 2 and 3 show the SHAP summary plot extracted using SVERL for the case of proactive rejection across the two networks. A SHAP summary plot for a specific action (class, in our case) provides a visual summary of how each feature influences the model's prediction, either increasing or decreasing the likelihood of that action. Features on the y-axis are ordered by their overall importance (mean absolute SHAP value), while the x-axis shows the SHAP values indicating impact and direction. Each point represents a data instance: its position shows the SHAP value, and its color reflects the feature's value in that state.

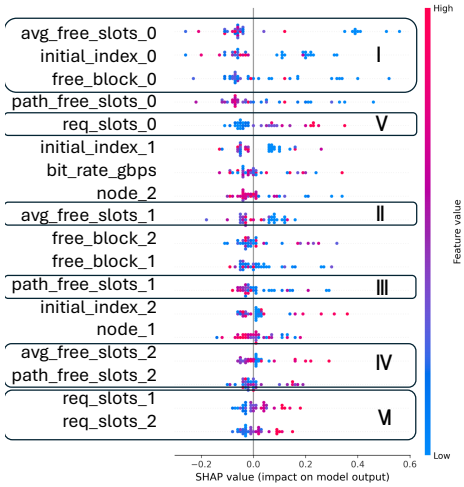

Fig. 2: Explanations of proactive rejections (NSF network topology).

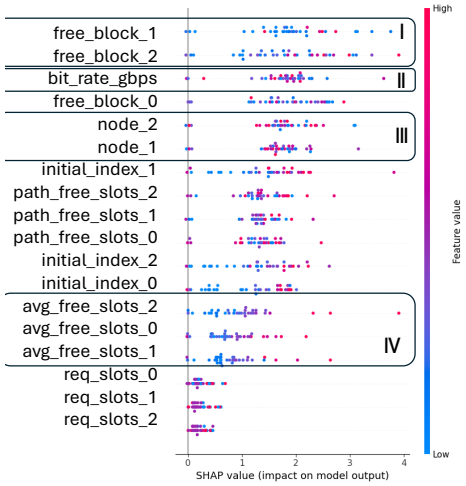

Fig. 3: Explanations of proactive rejections (German network topology).

Proactive rejection in NSF network topology are shown in Fig. 2. In inset I, we see that low values of the features corresponding to available resources on the shortest path (e.g., avg_free_slots_0) are the most influential factors toward proactive rejection (relatively high positive SHAP value), while high availability of slots pushes against proactive rejection (i.e., to accepting the lightpath request, as shown by negative SHAP values for points with medium or high values of the feature), which is intuitive. Similarly, avg_free_slots_1 and path_free_slots_1 (available slots on path 1) show a similar directional impact on agent's actions however with relatively lower SHAP value (insets II and III). Interestingly, in insets IV, we see that high availability of slots on path 3 (high values of avg_free_slots_2) is, in some cases, a significant indication to proactively reject a request. This suggests that the agent has learned a policy that considers the broader network context or potential future demands as high availability on alternative paths (path 3) might be associated with scenarios where the agent anticipates better use of those resources later. Such behavior points to a form of strategic reservation, indicating the agent's capacity to generalize beyond simple resource maximization. Finally, the explanations from insets V and VI show that the agent de-prioritizes accepting lightpath requests with relatively high number of slots (red points push toward rejection with positive SHAP value). While this indicates a preference (or, bias) for lightpaths that could potentially occupy less slots, it is important to note that this behavior is influenced by the design of the reward function, which treats all lightpaths equally regardless of the number of slots they would occupy or bit rate of these lightpath requests.

Turning to the explanations derived from the DRL agent on the German topology (Fig. 3), we observe that all SHAP values across all features are positive or zero (no influence). This indicates that, in every analyzed instance of proactive rejection, each feature contributed to the agent's decision to reject, suggesting a consistent pattern in the agent's perception of the network state. Specifically, the agent appears to have frequently encountered overloaded conditions during training, leading it to adopt a rejection-biased policy, i.e., a tendency to favor rejection over acceptance without relying on specific indications from the environment. For instance, lower values (represented in blue) of free_block_1 and _2 (inset I) are associated with higher SHAP impacts, meaning that limited availability of free spectrum strongly influenced the rejection decision. However, `bit_rate_gbps` regardless of whether its values are low or high (blue or red) also correlates with increased rejection likelihood, pointing to a broader, possibly heuristic-based rationale in the agent's behavior. In inset III, we see that node_1 and node_2 play a significant role in rejection decisions, in contrast to what was observed with the NSF topology. This suggests that the agent identified the presence of potentially unfavorable node pairs in the topology, likely distant ones, where the agent has learned to prefer rejection, possibly to preserve resources for future requests between closer nodes. Finally, in inset IV, we see that avg_free_slots on paths 1, 2 and 3 contribute more to rejection when having high values than when having low values. This suggests a flaw in the model's reasoning: rather than applying a well-articulated logic for proactive rejection, the agent appears to follow a more ambiguous and inconsistent approach.

Conclusions

We presented an explainable DRL framework for RMSA. By adapting the SVERL framework we highlight how the agent's proactive rejection behavior ranges from strategic in some topologies to

overly rigid in others. These insights underscore the importance of explainability for diagnosing policies and ensuring transparent, reliable network control.

Acknowledgements

This work has been partially supported by the EUREKA cluster CELTIC-NEXT project SUSTAINET-Advance funded by the Swiss Innovation Agency, and by the European Union's Horizon Europe research and innovation program through the ECO-eNET project (10113933).

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 2018.

[2] X. Chen, B. Li, R. Proietti, H. Lu, Z. Zhu, and S. B. Yoo, "DeepRMSA: A deep reinforcement learning framework for routing, modulation and spectrum assignment in elastic optical networks", *Journal of Lightwave Technology*, vol. 37, no. 16, pp. 4155–4163, 2019. DOI: 10.1109/JLT.2019.2923615.

[3] J. W. Nevin, S. Nallaperuma, N. A. Shevchenko, Z. Shabka, G. Zervas, and S. J. Savory, "Techniques for applying reinforcement learning to routing and wavelength assignment problems in optical fiber communication networks", *Journal of Optical Communications and Networking*, vol. 14, no. 9, pp. 733–748, 2022. DOI: 10.1364/JOCN.460629.

[4] L. Xu, Y.-C. Huang, Y. Xue, and X. Hu, "Hierarchical reinforcement learning in multi-domain elastic optical networks to realize joint RMSA", *Journal of Lightwave Technology*, vol. 41, no. 8, pp. 2276–2288, 2023. DOI: 10.1109/JLT.2023.3235039.

[5] C. Hernández-Chulde, R. Casellas, R. Martínez, R. Vilalta, and R. Muñoz, "Experimental evaluation of a latency-aware routing and spectrum assignment mechanism based on deep reinforcement learning", *Journal of Optical Communications and Networking*, vol. 15, no. 11, pp. 925–937, 2023. DOI: 10.1109/ECOC52684.2021.9605919.

[6] A. Asiri and B. Wang, "Deep reinforcement learning for QoT-aware routing, modulation, and spectrum assignment in elastic optical networks", *Journal of Lightwave Technology*, vol. 43, no. 1, pp. 42–60, 2025. DOI: 10.1109/JLT.2024.3446762.

[7] Y. Teng, C. Natalino, H. Li, *et al.*, "Deep-reinforcement-learning-based RMSCA for space division multiplexing networks with multi-core fibers [invited tutorial]", *Journal of Optical Communications and Networking*, vol. 16, no. 7, pp. C76–C87, 2024. DOI: 10.1364/JOCN.518685.

[8] Ciena. "Applying AI to maximize capacity and QoT in optical networks". Accessed: April 30, 2025. (2020), [Online]. Available: https://www.ciena.com/insights/articles/applying-ai-to-maximize-capacity-and-qot-in-optical-networks.html.

[9] Y. Cheng, S. Ding, Y. Shao, and C. C.-K. Chan, "PtrNet-RSA: A pointer network-based QoT-aware routing and spectrum assignment scheme in elastic optical networks", *Journal of Lightwave Technology*, vol. 42, no. 17, pp. 5808–5819, 2024. DOI: 10.1109/JLT.2024.3405587.

[10] G. A. Vouros, "Explainable deep reinforcement learning: State of the art and challenges", *ACM Computing Surveys*, vol. 55, no. 5, pp. 1–39, 2022. DOI: 10.1145/3527448.

[11] A. Heuillet, F. Couthouis, and N. Díaz-Rodríguez, "Explainability in deep reinforcement learning", *Knowledge-Based Systems*, vol. 214, p. 106 685, 2021. DOI: 10.1016/j.knosys.2020.106685.

[12] M. Arana-Catania, A. Sonee, A.-M. Khan, *et al.*, "Explainable reinforcement and causal learning for improving trust to 6G stakeholders", *IEEE Open Journal of the Communications Society*, pp. 1–1, 2025. DOI: 10.1109/OJCOMS.2025.3563415.

[13] D. Beechey, T. M. Smith, and Ö. Şimşek, "Explaining reinforcement learning with shapley values", in *International Conference on Machine Learning*, PMLR, 2023, pp. 2003–2014.

[14] S. Nallaperuma, Z. Gan, J. Nevin, M. Shevchenko, and S. J. Savory, "Interpreting multi-objective reinforcement learning for routing and wavelength assignment in optical networks", *Journal of Optical Communications and Networking*, vol. 15, no. 8, pp. 497–506, 2023. DOI: 10.1364/JOCN.483733.

[15] Z. Xiong, Y.-C. Huang, and X. Hu, "Graph attention network enhanced deep reinforcement learning framework for routing, modulation, and spectrum allocation in EONs", in *Asia Communications and Photonics Conference (ACP) and International Conference on Information Photonics and Optical Communications (IPOC)*, 2024, pp. 1–6. DOI: 10.1109/ACP/IPOC63121.2024.10810116.

[16] J. Suárez-Varela, A. Mestres, J. Yu, *et al.*, "Routing in optical transport networks with deep reinforcement learning", *Journal of Optical Communications and Networking*, vol. 11, no. 11, pp. 547–558, 2019. DOI: 10.1364/JOCN.11.000547.

[17] O. Ayoub, C. Natalino, and P. Monti, "Towards explainable reinforcement learning in optical networks: The RMSA use case", in *Optical Fiber Communications Conference and Exhibition (OFC)*, 2024, W4I.6. DOI: 10.1364/OFC.2024.W4I.6.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal policy optimization algorithms*, 2017. arXiv: 1707.06347 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1707.06347.

[19] C. Natalino, T. Magalhaes, F. Arpanaei, *et al.*, "Optical Networking Gym: An open-source toolkit for resource assignment problems in optical networks", *Journal of Optical Communications and Networking*, vol. 16, no. 12, G40–G51, 2024. DOI: 10.1364/JOCN.532850.

[20] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations", *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021. [Online]. Available: http://jmlr.org/papers/v22/20-1364.html.

[21] S. Orlowski, R. Wessäly, M. Pióro, and A. Tomaszewski, "SNDlib 1.0–survivable network design library", *Networks*, vol. 55, no. 3, pp. 276–286, 2010. DOI: 10.1002/net.20371.