CHALMERS
UNIVERSITY OF TECHNOLOGY

Bridging the Trust Gap in AI-Driven Optical Networks with Structured Explainability

Bridging the Trust Gap in AI-Driven Optical Networks with Structured Explainability

Kiarash Rezaei*, Omran Ayoub†, Carlos Natalino*, Paolo Monti*

* Department of Electrical Engineering, Chalmers University of Technology, 412 96 Gothenburg, Sweden
{kiarashr, carlos.natalino, mpaolo}@chalmers.se
†University of Applied Sciences and Arts of Southern Switzerland, 6928 Lugano, Switzerland
omran.ayoub@supsi.ch

***Abstract*—AI/ML models can automate optical-network decisions, yet operators distrust their opaque outputs. Existing explainability methods help but remain hard to interpret and not directly actionable. We propose the *Capture–Characterize–Communicate* (3C) framework, which formalizes explainability as an end-to-end pipeline, i.e., from capturing model behavior, through local explanations, to human-readable decision guidance. The framework is demonstrated on two optical-network problems: explainable RL for RMSA and LLM-augmented explainability for QoT estimation, where it produces auditable, operator-facing explanations.**

***Index Terms*—Optical networks, network automation, explainable AI, SHAP, reinforcement learning, large language models.**

I. Introduction

Artificial intelligence & machine learning (AI/ML) models are increasingly recognized as promising tools for addressing key operational challenges in optical networks. Two representative use cases are resource assignment, which allocates network resources under dynamic traffic conditions, and quality-of-transmission (QoT) estimation, which predicts whether a candidate lightpath will meet its performance target. In both domains, recent work has demonstrated strong model performance in simulation environments [1], [2].

Despite these results, AI/ML models are rarely deployed in production optical networks. The core barrier is a trust deficit: operators cannot inspect *why* a model made a particular decision, making it difficult to verify whether individual actions are sound, comply with operational policies, or build the confidence needed for autonomous operation [3]. Post-hoc explainability techniques such as SHAP [4] can help bridge this gap, yet they present significant drawbacks: their outputs (e.g., bar charts, force plots, attribution tables) are hard to interpret without specialized knowledge, require substantial expertise in both optical networking and the explainability method itself, and are not easily actionable in an operational context where decisions must be made quickly and confidently.

To overcome these limitations, we propose the *Capture–Characterize–Communicate* (3C) framework, which formalizes explainability as a *decision-support instrument*: structured evidence paired with concise, operator-facing narratives tied to specific operational decisions. Rather than leaving operators to interpret raw explainable AI (XAI) artifacts, the framework prescribes a three-step pipeline (i.e., from logging model behavior, through local explanation methods, to human-readable explanation cards) that makes each decision-scoped, grounded, and auditable.

To showcase its benefits, the 3C framework is applied to two optical network use cases: explainable reinforcement learning (XRL) for routing, modulation format, and spectrum assignment (RMSA) [5], and LLM-augmented explainability for QoT estimation [6], demonstrating how the framework structures actionable explainability in practice.

II. The 3C Framework

The proposed 3C framework organizes actionable explainability into three steps (Fig. 1):

Capture. At decision time, the system logs model inputs, outputs, and relevant operational context (e.g., network load, requested bandwidth, path candidates). This creates a traceable and reproducible snapshot that anchors all subsequent analysis.

Characterize. Local explanation methods such as SHAP values, feature interactions, surrogate models, are applied to the captured snapshot to quantify *which* input features drove the decision and *how* they interacted.

Communicate. The quantified evidence is translated into an operator-facing *explanation card*: a short narrative stating the decision rationale, the dominant factors, and a recommended verification action. The card is stored with references to its originating snapshot and XAI artifacts, establishing a full provenance chain.

The 3C framework is most effective when explanations are *(i) decision-scoped*, i.e., tied to a single operational action rather than aggregate model behavior; *(ii) grounded* in measurable artifacts; and *(iii) auditable* over time, enabling post-incident review and trend analysis.

III. Use Cases

We deploy the 3C framework on two use cases: RMSA resource allocation under dynamic traffic [5], and QoT estimation for lightpath provisioning [6].

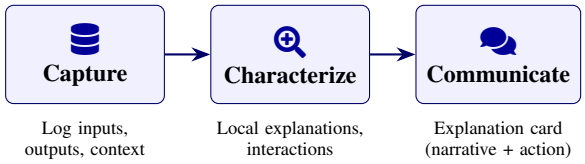

Fig. 1. The 3C framework: each operational decision is captured, characterized via local XAI methods, and communicated as a human-readable explanation card linked to its originating artifacts.

A. Explainable RL for RMSA

We trained a deep reinforcement learning agent for deciding the routing selection on an elastic optical network simulator with first-fit spectrum assignment under three environment variants differing in reward shaping and action masking [5]. To explain the learned policy, the agent was rolled out, state-action pairs were collected, and a supervised surrogate classifier was fit to mirror the agent's decisions. SHAP was then applied to the surrogate to produce per-feature attributions.

The analysis reveals how reward design and action masking change the relative importance of features such as spectrum availability and path length. For instance, when no action masking is applied, spectrum availability is the most relevant feature for the decision, but when action masking is performed, the feature becomes one of the least relevant. These insights can be used to support concrete improvements: redesigning reward functions, adding guardrails that flag anomalous feature-importance patterns, and constructing debugging checklists. Within the deployed 3C pipeline: the rollout constitutes *Capture*, surrogate training and SHAP attribution constitute *Characterize*, and the derived operational guidelines constitute *Communicate*.

B. LLM-Augmented XAI for QoT

We trained a regression model for QoT estimation, predicting whether a candidate lightpath meets its bit-error-rate target [6]. SHAP values and SHAP feature interaction values were computed per prediction, identifying the top-k contributing features and top-m pairwise interactions. These were embedded into a structured prompt, and an LLM generated a human-readable explanation.

A human-centered evaluation compared LLM-generated explanations against a SHAP-only baseline on correctness, scope, and usefulness. Including feature interactions improved perceived *usefulness* by 12%, while 97.5% of explanations were rated as accurate. The structured prompt constrains the LLM output, reducing hallucination risk and ensuring traceability to XAI artifacts. Operationally, non-specialist staff can act on a two-sentence rationale without inspecting SHAP plots or understanding feature-interaction mechanics. Within the deployed 3C pipeline: the per-prediction XAI artifacts constitute *Capture*, SHAP value and interaction analysis constitute *Characterize*, and the LLM-generated explanation constitutes *Communicate*.

C. Discussion

The two deployments reveal distinct, complementary design insights. The XRL study shows that structured feature attributions can reveal policy flaws invisible to aggregate metrics, while the LLM study shows that narratives grounded in those attributions are perceived as more useful than raw XAI plots. Together, they highlight two general principles: *Characterize* benefits from interaction-aware methods, and *Communicate* benefits from language generation constrained by XAI evidence. For deployment, explanation cards can be integrated into network management dashboards alongside safety hooks, e.g., triggering escalation when feature-importance distributions contradict expected behavior. The provenance chain embedded in each card further enables longitudinal auditing: operators can trace how model reasoning evolved as traffic patterns or network topology changed.

IV. Conclusion

This paper proposed the 3C (*Capture–Characterize–Communicate*) framework, which turns black-box model outputs into structured, operator-actionable decision support, as demonstrated on explainable RL for RMSA and LLM-augmented explainability for QoT estimation.

While the current deployments target single-model, per-decision explanations, scaling the 3C pipeline to production settings raises several open challenges: *(i)* extending explainability to multi-modal decision support that fuses heterogeneous data sources (structured telemetry, logs, images) and may itself include LLM components, where the evidence chain is harder to trace; *(ii)* moving from local, model-level explanations toward system-wide explainability, where a single operational decision results from the composition of multiple models and policies; and *(iii)* shifting emphasis from explaining individual inferences to explaining end-to-end decisions, i.e., what triggered the action, which checks were applied, and what uncertainties remain. Addressing these challenges will be essential for realizing trustworthy autonomous optical network operations.

Acknowledgment: This work was supported by the Celtic-Next Flagship SUSTAINET-Advance project funded in Sweden by VINNOVA (2025-02987) and in Switzerland by Innosuisse (No. 119.588 INT-ICT).

References

[1] Y. Teng *et al.*, "DRL-assisted QoT-aware service provisioning in multi-band elastic optical networks," *Journal of Lightwave Technology*, vol. 43, no. 19, pp. 9090–9101, 2025.

[2] P. Lechowicz *et al.*, "Toward better QoT estimation: An ML architecture with link-level embedding layers," *IEEE Networking Letters*, vol. 7, no. 2, pp. 122–125, 2025.

[3] O. Ayoub *et al.*, "Towards explainable artificial intelligence in optical networks: the use case of lightpath QoT estimation," *Journal of Optical Communications and Networking*, vol. 15, no. 1, pp. A26–A38, 2023.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[5] O. Ayoub, C. Natalino, and P. Monti, "Towards explainable reinforcement learning in optical networks: The RMSA use case," in *International Conference on Transparent Optical Networks (ICTON)*, 2024.

[6] K. Rezaei *et al.*, "Generative explainability for next-generation networks: LLM-augmented XAI with mutual feature interactions," in *International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2025.