Engineering Trust in AI-Driven Optical Networks: Challenges in Moving from Monitoring to Autonomy

Carlos Natalino *, Omran Ayoub †, Paolo Monti *

*Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden
†University of Applied Sciences and Arts of Southern Switzerland, Lugano, Switzerland

***Abstract*—Artificial intelligence and machine learning (AI/ML) are widely recognized as primary enablers for next-generation optical network automation. However, despite significant algorithmic advances, the deployment of these technologies remains largely confined to monitoring tasks and supervised decision-making. In this paper, we explore recent achievements in applying AI/ML to optical networks, highlighting successes in quality-of-transmission estimation, failure management, and sensing. We then analyze the discrepancy between high algorithmic predictive performance and the limited adoption of fully autonomous control loops. Specifically, we investigate the technical factors that inhibit operator trust, including model opacity, lack of generalization to dynamic and evolving conditions, insufficient performance characterization, and limited performance guarantees. The paper concludes by identifying research efforts that would bring us close to transitioning AI/ML from a passive tool into a trusted, autonomous controller.**

***Index Terms*—Machine learning, trust, autonomy, uncertainty.**

I. Introduction

The optical transport layer must now operate at petabit-per-second scales to support large language model training clusters, 5G/6G mobile infrastructure, and hyperscale data center interconnects [1]. To manage the resulting complexity, which spans spatial division multiplexing (SDM), multi-band switching, and dynamic edge-to-cloud connectivity, artificial intelligence (AI) and machine learning (ML) have emerged as indispensable enablers. A rich body of literature now demonstrates that ML models can estimate quality of transmission (QoT) with sub-decibel accuracy, detect soft failures with near-perfect recall, and optimize resource allocation beyond the reach of analytical heuristics [2].

A central paradox, however, persists. High algorithmic/predictive performance in controlled laboratory and simulation settings coexists with extremely limited autonomous deployment in production networks [3]–[5]. We refer to this discrepancy as the *trust deficit*, that is, the gap between what AI *can* predict and what operators *will allow* it to control autonomously. Borrowing from the autonomous vehicle domain, the TM Forum defines six autonomy levels (L0 to L5) ranging from fully manual operation to full autonomy [3]. Most optical network deployments remain at L1 and L2, where AI assists human decisions but does not execute them independently. At L3 and L4, the system would autonomously execute decisions within defined boundaries, with human oversight shifting from approval to exception handling. At L5, the network would fully self-manage without human intervention. The transition to these higher autonomy levels demands not merely better models, but fundamentally different guarantees about *when* and *how* those models may fail.

In this context, we argue that the community needs a shift in the metrics it optimizes for. Before surrendering control, operators need assurance that AI can recognize the limits of its own knowledge. While accuracy and mean error measure what AI *knows*, metrics such as conformal coverage, violation rates, and calibration error quantify what AI *does not know*.

The contributions of this paper are threefold. First, we provide a structured survey of AI/ML applications across QoT estimation, failure management, and network optimization, accompanied by a quantitative analysis of the performance metrics employed by the community. Second, we systematically deconstruct the trust deficit by identifying both non-technological barriers and four technical inhibitors, namely model opacity, distribution shift, insufficient uncertainty quantification (UQ), and the absence of formal safety guarantees. Finally, we present a trust engineering framework comprising explainable AI (XAI), UQ via conformal prediction (CP), and safe reinforcement learning (RL) with formal verification as concrete pathways toward autonomous control. The ambition is that by raising awareness of the limitations of traditional measures, the community can start moving towards more network-specific ways to assess the performance of AI/ML models. Ultimately, the transition from AI-assisted monitoring to autonomous control hinges on verifiable trust.

II. State of the Art: AI/ML in Optical Networks

Over the past decade, AI/ML techniques have permeated virtually every layer of optical network operation, from QoT estimation and failure management to resource optimization. This section surveys some significant results across these

This work was supported in part by the Swedish Research Council (VR) BRAIN project (grant agreement no. 2022-04798) and by Vinnova (Sweden's Innovation Agency) CELTIC-NEXT SUSTAINET-Advance project (grant agreement no. 2025-02987).

978-3-903176-78-2

domains before examining the performance metrics that the community predominantly employs to evaluate these advances.

A. Quality of Transmission Estimation

Traditionally, QoT estimation has relied on analytical models such as the Gaussian Noise (GN) model, which require precise parameterization of the physical layer [6]. In practice, optical networks accumulate changes over years or decades of operation through fiber splicing, component replacements, and environmental variations, making accurate parameterization increasingly difficult to maintain. When precise parameters are unavailable, operators resort to conservative margins, typically 1.5 to 2.3 dB, to ensure service reliability at the cost of spectrum waste [7]. Notably, practical accuracy is bounded as much by input-parameter uncertainty as by the model itself, a distinction directly tied to trust [8]. We refer to [9] for a comprehensive survey on this topic.

AI/ML offers a data-driven alternative that learns directly from available features and measured outcomes, bypassing the need for exact parameterization [10], [11]. This approach has demonstrated remarkable results, with ML-augmented controllers reducing design margins from 2.28 dB to as low as 0.15 dB in disaggregated open line system (OLS) deployments [7]. Physics-informed approaches offer a middle ground by embedding analytical models as structural priors within neural networks [6], while transfer learning (TL) reduces the required training data by 80 to 98% [12]–[14].

A key challenge for data-driven models is generalization across both spatial and temporal dimensions. Spatial generalization concerns the ability to maintain accuracy across unseen network topologies, where wide deep neural networks (DNNs) achieve a maximum prediction error of 0.40 dB on previously unseen topologies [11]. Temporal generalization concerns the ability to maintain accuracy as the physical infrastructure ages. Models trained on beginning-of-life conditions progressively lose accuracy over years of operation, requiring margin-driven adaptation mechanisms that trigger retraining based on operational performance degradation [15]. These generalization challenges highlight that high accuracy under controlled conditions does not automatically translate into trustworthy operation in evolving production networks.

B. Failure Management

Failure management in optical networks encompasses a multi-stage pipeline consisting of detection, identification, localization, and prediction. For a comprehensive treatment of ML-based failure management, the reader is referred to [16] and [17]. Here, we highlight representative results that illustrate near-perfect performance within controlled settings.

At the detection stage, unsupervised approaches serve as effective gatekeepers. Long short-term memory (LSTM) autoencoders achieve 100% recall for soft failure detection without requiring labeled training data [18], while gated recurrent unit (GRU) autoencoders operating on optical time-domain reflectometer (OTDR) traces attain F1-scores of 96.86% [19]. Once anomalies are detected, cascaded dual-stage architectures that combine unsupervised detection with supervised classification achieve 99.7% global accuracy for soft failure identification [18]. Cognitive assurance frameworks further enable proactive detection with lead times of up to 96 hours [20].

Failure localization presents a distinct challenge, particularly under dynamic network conditions. Graph neural network (GNN)-based architectures with attention mechanisms achieve 97.4% localization accuracy even under topology changes [21], and network digital twin (NDT)-assisted approaches combining artificial neural networks (ANNs) with principal component analysis (PCA) demonstrate 100% single-failure localization in experimental testbeds [22]. Nevertheless, cross-type and cross-magnitude generalization remains largely unaddressed [16]. In prognostics, bidirectional gated recurrent unit (BiGRU) models enable 4-day advance failure prediction with 99.61% accuracy [23].

These results paint an impressive picture of near-perfect accuracy at every stage of the failure management pipeline. Yet autonomous failure remediation remains elusive, and no encompassing solution addresses all pipeline stages in an integrated fashion. As we argue in Section III, the barrier is not accuracy but trust.

C. Network and Resource Optimization

Deep reinforcement learning (DRL) has emerged as a powerful paradigm for routing, modulation, and spectrum assignment (RMSA) in elastic optical networks (EONs). OpticGAI employs generative diffusion models to learn multimodal policy distributions, outperforming proximal policy optimization (PPO), advantage actor-critic (A2C), and heuristic baselines on blocking rate [24]. DRL-based approaches have also demonstrated strong benefits for multi-band EONs, where QoT-aware service provisioning must account for inter-band interference and amplifier gain variations [25].

Simulation-to-real transfer constitutes a critical milestone for practical deployment. PPO agents trained entirely on GN model simulations achieve 3600 Gbit/s throughput on a physical wavelength-division multiplexing (WDM) testbed without retraining, demonstrating that RL policies can generalize from simulation to hardware [26]. On the planning side, dual-stage approaches integrating ML-based QoT estimation with mixed-integer linear programming achieve 52.4% spectrum savings compared to margined allocation [27]. Still, the deployment of autonomous DRL control in optical networks remains challenging owing to the lack of transparent understanding of the decision process.

D. Performance Metrics

A closer examination of the metrics employed across the surveyed literature reveals a telling pattern. The community predominantly relies on standard predictive metrics from two families. Classification metrics (accuracy, F1-score, area under the receiver operating characteristic curve (AUC-ROC), recall, false-positive rate) are employed for QoT feasibility and failure detection, while regression metrics (mean absolute error (MAE), mean squared error (MSE), root mean squared error

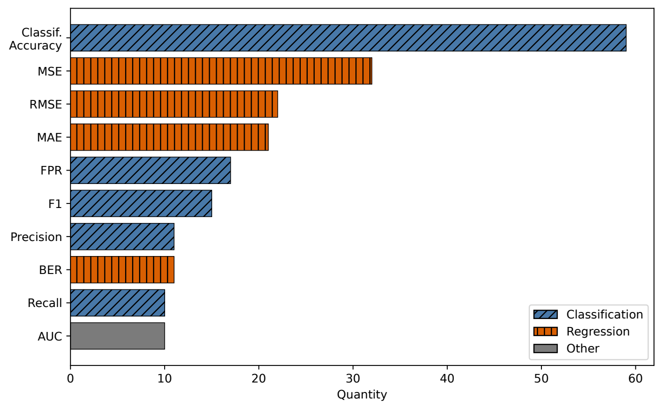

Fig. 1. Quantity of performance metrics reported across the AI/ML works surveyed in Sec. II. Classification metrics (accuracy, F1, AUC-ROC, recall, FPR) and regression metrics (MAE, MSE, RMSE, max error, R^2) are shown together. BER here is a regression *target*, not an ML evaluation metric. Point-estimate metrics dominate, while trust-relevant metrics (calibration error, conformal coverage, violation rate) are rarely reported.

(RMSE), maximum error, R^2) are used for optical signal-to-noise ratio (OSNR), Q-factor, and bit-error rate (BER) estimation. In particular, BER in Fig. 1 is a *prediction target*, whose error is measured through regression metrics. Trust-relevant metrics such as calibration error, conformal coverage, uncertainty bounds, and formal violation rates remain severely underused across both families [28], [29]. Operational metrics such as blocking probability and margin reduction (in dB) bridge algorithmic performance to network-level outcomes. The dominance of point-estimate metrics, as illustrated in Fig. 1, reveals that the community has so far focused on *predictive performance*, while *operational trustworthiness* is necessary for wide adoption. Bridging this metric gap is a prerequisite for autonomous deployment.

III. Deconstructing the Trust Deficit

The preceding section demonstrates that AI/ML models can achieve near-perfect accuracy across a wide range of optical network tasks. Yet, as noted in the introduction, most production deployments remain at low autonomy levels [3]. This section deconstructs the trust deficit by examining both the non-technological barriers that constrain adoption and the technical inhibitors that undermine confidence in autonomous operation.

A. Non-Technological Barriers

Operators today manage heterogeneous, multi-vendor infrastructure where siloed and proprietary data have accumulated over decades of incremental upgrades. Correlating AI-derived telemetry with physical topology and legacy inventory systems remains a formidable integration challenge [4]. The economic case for AI-driven automation is equally non-trivial. Field operations account for 60–70% of total telecom operational expenditure, yet translating statistical accuracy improvements into quantifiable return on investment remains an open challenge [4]. Even when AI demonstrably reduces margins by over 2 dB, savings must be weighed against the cost and risk of replacing proven engineering practices.

The convergence of photonics, telecom engineering, and AI/ML expertise required for deploying intelligent optical networks is critically scarce [4]. Without domain-knowledgeable personnel who can interpret AI outputs, operators default to conservative manual oversight. Standardization gaps further exacerbate these challenges. Although the International Telecommunication Union (ITU) has established architectural frameworks for ML in future networks [30], practical gaps persist in interoperable dataset formats, performance reporting conventions, and algorithmic bias and fairness criteria [31], [32]. Inconsistent vendor implementation of management interfaces also remains a practical barrier to zero-touch interoperability.

B. Technical Inhibitors of Trust

Beyond organizational constraints, four interconnected technical challenges erode confidence in autonomous AI operation, namely model opacity, distribution shift, insufficient UQ, and the absence of formal safety guarantees. These inhibitors are not independent. Rather, they compound one another, as an opaque model is especially dangerous when it encounters out-of-distribution data and produces overconfident predictions without any guarantee that its actions will remain safe.

Model opacity. Deep networks map high-dimensional telemetry to outputs through millions of nonlinear transformations without exposing the reasoning behind their decisions [2]. In high-stakes network environments, operators demand explainability not only for diagnostic tracing but also for regulatory accountability [4]. Lack of transparency has been identified as the single most significant technical barrier to trustworthy AI in optical networks [4], [33].

Distribution shift and concept drift. Optical networks are dynamic physical systems. Fiber aging, amplifier drift, traffic pattern changes, and component replacements continuously violate the i.i.d. assumption that training and deployment data share the same distribution [15], [34]. Consequently, QoT models trained on one topology may fail on different topologies or after physical layer changes [11], [13]. The reliance on synthetic simulator data, which encodes simplifications that degrade performance under live conditions, further compounds this problem [4], [18].

Insufficient uncertainty quantification. Standard DNNs output uncalibrated point estimates or softmax scores that are misinterpreted as confidence measures [29]. An overconfident wrong prediction from an autonomous agent can rapidly destabilize the network, for instance by provisioning a lightpath that violates QoT requirements or disrupts co-propagating channels. Uncertainty in ML predictions arises from two distinct sources requiring different mathematical treatments. Aleatoric uncertainty represents inherent measurement noise that cannot be reduced with more data, while epistemic uncertainty stems from data scarcity or model limitations [29], [35]. Recent works address this gap through Bayesian recurrent networks for QoT forecasting [28] and policy-driven CP [36], but their suitability for autonomous operation remains open.

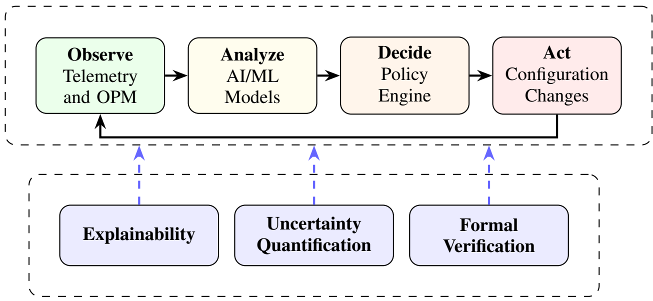

Fig. 2. Trust engineering framework for autonomous optical networks. The autonomous control loop (Observe, Analyze, Decide, Act) is augmented by a trust engineering layer comprising explainability, uncertainty quantification, and formal verification mechanisms.

Absence of formal safety guarantees. Unlike classical control theory, which offers well-established stability frameworks, standard DRL provides no formal stability guarantees [37]. An exploratory RL action in a live EON could cause routing oscillations, link congestion, or physical-layer constraint violations [38]. The inability to mathematically guarantee that an autonomous agent will operate within safe bounds is a roadblock to network orchestration at higher autonomy levels [38], [39].

IV. Engineering Trust: Pathways to Autonomous Control

Closing the trust deficit requires augmenting the autonomous control loop (observe, analyze, decide, act) with a dedicated trust engineering layer, as illustrated in Fig. 2. This layer comprises three complementary pillars. XAI renders model reasoning transparent, UQ quantifies prediction reliability, and formal verification guarantees operational safety. We examine each pillar and its recent advances in the optical networking context.

Explainable AI. Post-hoc explainability methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) decompose model outputs to quantify individual feature contributions (e.g., transmission distance, launch power, span count) to a given prediction [33]. For failure localization, attention mechanisms in GNN-based architectures provide inherent interpretability by assigning scores to network elements, thereby enabling operators to trace decisions to specific links or amplifiers [21]. Physics-informed architectures further enhance transparency by ensuring that model parameters remain physically interpretable [6], [40]. DRL decisions can similarly undergo scrutiny through policy decomposition techniques that reveal which network state features drive routing and spectrum assignment actions [41]. To standardize trust communication, model report cards [4] and natural-language explanations of feature importance [42] have been proposed.

Uncertainty quantification and conformal prediction. Bayesian approaches provide a principled foundation for UQ. Gaussian Processs (GPs) deliver fully probabilistic outputs with inherent uncertainty bounds. When combined with active learning, GPs reduce training data requirements by 33% while maintaining calibrated estimates [35], [43]. Deep quantile regression offers an alternative by providing lightpath-specific probabilistic bounds that reduce design margins by 71 to 83% and blocking probability by 19% [44].

CP has recently emerged as a particularly compelling framework owing to its distribution-free, model-agnostic nature and its ability to provide finite-sample coverage guarantees [16], [29]. Policy-driven conformal QoT frameworks raise lightpath-feasibility prediction accuracy from 92% to 99.6%, adapting to domain shifts with minimal new data [36]. Crucially, prediction intervals automatically widen for out-of-distribution (OOD) paths, thereby signaling uncertainty to the orchestrator and preventing silent failures. Complementing these methods, generative adversarial network (GAN)-based OOD detection rejects anomalous signals (AUC=0.952) before they reach the analyzer [45], [46], and deterministic rigid shifts (e.g., +0.2 dB) can convert non-conservative predictions to conservative ones with a 94% conservatism rate [7].

Safe reinforcement learning and formal verification. Embedding Lyapunov stability conditions directly into RL objectives provides formal stability certificates and interpretable visualization of the region of attraction [37]. This approach has been recently demonstrated in diffractive optical network (DON) control, where Lyapunov-constrained RL achieves provably stable recovery from disturbances in less than 2.1 seconds, while TL reduces training requirements by over 70% [47]. A complementary strategy is TrustOPT, where a Lipschitz-bounded trusted configuration space lets online optimization explore QoT control actions on live traffic with a worst-case safety guarantee, validated in a field trial [48]. These results motivate the adaptation of Lyapunov-constrained RL to telecom optical network orchestration, where stability guarantees are equally critical.

V. Challenges and Future Directions

While the trust engineering mechanisms described represent significant advances, several fundamental challenges must be addressed before autonomous optical network operation can become a practical reality.

Real-world open data scarcity. This is a long-lasting challenge that remains critical, as the majority of current works rely on synthetic data. Openly released experimental datasets such as [49] are encouraging first steps, yet they remain isolated and cover only a fraction of production operating conditions. Systematic benchmarks on live network telemetry, accompanied by standardized dataset formats, labeling conventions, and privacy-preserving sharing mechanisms, are essential for validating trust-relevant metrics under realistic conditions [4], [18].

Scalability of formal verification. Current formal verification methods are computationally prohibitive for real-time application in large-scale network state spaces. While incremental techniques such as verified safe reinforcement learning (VSRL) offer order-of-magnitude speedups [38], verifying safety properties across the full state space of a continental-scale optical network remains an open problem [39].

Deployment barriers. Bridging the simulation-to-real gap remains a persistent challenge. The structural similarity gap of 0.53 observed for Lyapunov-constrained DON controllers [47] and the widespread reliance on simulators for training underscore the need for tighter integration between simulation environments and physical testbeds. Heterogeneous multi-vendor equipment and the lack of interoperable telemetry datasets and governance frameworks further impede unified AI orchestration [4], [32].

Continuous adaptation. Optical networks evolve continuously, and static models inevitably degrade over time. Online and continual learning paradigms, equipped with formal drift detection and automated retraining triggers, are essential for maintaining model validity throughout the network lifecycle [15], [50], [51]. Without such mechanisms, the trust established during initial deployment erodes as the network diverges from the conditions under which the model was validated.

Digital twins as trust sandboxes. High-fidelity virtual representations of optical infrastructure can enable risk-free validation of autonomous actions [52]. Early results are promising. NDT-assisted failure localization achieves 100% accuracy in testbeds [22], and modular digital twins reach $1000\times$ speedup over split-step simulation [53]. However, scaling to network-wide fidelity with continuous telemetry synchronization remains an open challenge [51], and there is a lack of open solutions for widespread adoption and research.

Agentic AI and LLM-based orchestration. large language models (LLMs) initially enabled chatbot-like functionalities [54] and are now enabling agentic systems that show promise for autonomous network management [55], [56]. However, inference latency remains prohibitive for real-time control, and LLM hallucination poses a fundamental trust barrier that must be addressed before these architectures can participate in closed-loop operation.

VI. Conclusion

This paper has argued that the transition from AI-assisted monitoring to AI-driven autonomous control in optical networks requires a paradigm shift from optimizing empirical accuracy to engineering verifiable trust. The three pillars of trust engineering, namely explainability, uncertainty quantification, and formal verification, must be systematically integrated into the AI/ML lifecycle, from model design through deployment and continuous operation. We call upon the community to adopt trust-relevant metrics such as conformal coverage, violation rates, and calibration error alongside traditional predictive metrics as standard reporting practice. Direct validation of these pillars through interviews and joint studies with network operators is an essential next step. The convergence of safe RL, CP, and emerging technologies such as digital twins and agentic AI can potentially enable the transition to higher levels of optical network autonomy within the next decade.

References

[1] E. Astaiza Hoyos, H. F. Bermúdez-Orozco, and J. A. Aldana-Gutierrez, "Towards 6G: A review of optical transport challenges for intelligent and autonomous communications," *Computation*, vol. 13, no. 12, p. 286, Dec. 2025.

[2] D. Wang and M. Zhang, "Artificial intelligence in optical communications: From machine learning to deep learning," *Frontiers in Communications and Networks*, vol. 2, Mar. 2021.

[3] R. Webb and D. Bushaus, "Autonomous networks: In search of best practice," TM Forum, Tech. Rep., Dec. 2024.

[4] F. N. Khan, "Non-technological barriers: The last frontier towards AI-powered intelligent optical networks," *Nature Communications*, vol. 15, no. 1, p. 5995, Jul. 2024.

[5] C. Natalino and P. Monti, "Optical network automation, programmability, and AI: The path to 6G readiness," in *IEEE International Conference on Machine Learning for Communication and Networking (ICMLCN)*, May 2025.

[6] E. Seve, J. Pesic, and Y. Pointurier, "Associating machine-learning and analytical models for quality of transmission estimation: Combining the best of both worlds," *Journal of Optical Communications and Networking*, vol. 13, no. 6, pp. C21–C30, Jun. 2021.

[7] A. D'Amico, S. Straullu, A. Nespola, I. Khan, E. London, E. Virgillito, S. Piciaccia, A. Tanzi, G. Galimberti, and V. Curri, "Using machine learning in an open optical line system controller," *Journal of Optical Communications and Networking*, vol. 12, no. 6, pp. C1–C11, Jun. 2020.

[8] Y. Pointurier, "Machine learning techniques for quality of transmission estimation in optical networks," *Journal of Optical Communications and Networking*, vol. 13, no. 4, pp. B60–B71, Apr. 2021.

[9] R. Ayassi, A. Triki, N. Crespi, R. Minerva, and M. Laye, "Survey on the use of machine learning for quality of transmission estimation in optical transport networks," *Journal of Lightwave Technology*, vol. 40, no. 17, pp. 5803–5815, Sep. 2022.

[10] J. Mata, I. de Miguel, R. J. Durán, J. C. Aguado, N. Merayo, L. Ruiz, P. Fernández, R. M. Lorenzo, E. J. Abril, and I. Tomkos, "Supervised machine learning techniques for quality of transmission assessment in optical networks," in *International Conference on Transparent Optical Networks (ICTON)*, Jul. 2018, p. We.B3.5.

[11] I. Khan, M. Bilal, and V. Curri, "Assessment of cross-train machine learning techniques for QoT-estimation in agnostic optical networks," *OSA Continuum*, vol. 3, no. 10, pp. 2690–2706, Oct. 2020.

[12] J. Yu, W. Mo, Y.-K. Huang, E. Ip, and D. C. Kilper, "Model transfer of QoT prediction in optical networks based on artificial neural networks," *Journal of Optical Communications and Networking*, vol. 11, no. 10, pp. C48–C57, Oct. 2019.

[13] C. Rottondi, R. di Marino, M. Nava, A. Giusti, and A. Bianco, "On the benefits of domain adaptation techniques for quality of transmission estimation in optical networks," *Journal of Optical Communications and Networking*, vol. 13, no. 1, pp. A34–A43, Jan. 2021.

[14] P. Paudyal, S. Shen, S. Yan, and D. Simeonidou, "Toward deployments of ml applications in optical networks," *IEEE Photonics Technology Letters*, vol. 33, no. 11, pp. 537–540, Jun. 2021.

[15] P. Lechowicz, C. Natalino, and P. Monti, "QoT estimation with margin-driven transfer learning in time-varying optical networks," in *Optical Fiber Communications Conference and Exhibition (OFC)*, Mar. 2025, p. M1J.5.

[16] F. Musumeci and M. Tornatore, "Failure management in optical networks with ML: A tutorial on applications, challenges, and pitfalls [invited]," *Journal of Optical Communications and Networking*, vol. 17, no. 8, pp. C144–C155, Aug. 2025.

[17] S. Cruzes, "Failure management overview in optical networks," *IEEE Access*, vol. 12, pp. 169 170–169 193, 2024.

[18] S. Ghosh and A. Adhya, "Soft failure detection and identification in optical networks using cascaded deep learning model," *Computer Networks*, vol. 262, p. 111159, May 2025.

[19] K. Abdelli, J. Y. Cho, F. Azendorf, H. Griesser, C. Tropschug, and S. Pachnicke, "Machine-learning-based anomaly detection in optical fiber monitoring," *Journal of Optical Communications and Networking*, vol. 14, no. 5, pp. 365–375, May 2022.

[20] D. Rafique, T. Szyrkowiec, H. Grießer, A. Autenrieth, and J.-P. Elbers, "Cognitive assurance architecture for optical network fault management," *Journal of Lightwave Technology*, vol. 36, no. 7, pp. 1443–1450, Apr. 2018.

[21] V. Karunakaran, R. Romero Reyes, B. Shariati, J. Karl Fischer, A. Autenrieth, and T. Bauschert, "Dynamic network-aware soft failure localization using machine learning in optical networks," *Journal of Optical Communications and Networking*, vol. 17, no. 11, pp. 1032–1046, Nov. 2025.

[22] K. S. Mayer, R. P. Pinto, J. A. Soares, D. S. Arantes, C. E. Rothenberg, V. Cavalcante, L. L. Santos, F. D. Moraes, and D. A. A. Mello, "Demonstration of ml-assisted soft-failure localization based on network digital twins," *Journal of Lightwave Technology*, vol. 40, no. 14, pp. 4514–4520, Jul. 2022.

[23] C. Zhang, D. Wang, L. Wang, J. Song, S. Liu, J. Li, L. Guan, Z. Liu, and M. Zhang, "Temporal data-driven failure prognostics using BiGRU for optical networks," *Journal of Optical Communications and Networking*, vol. 12, no. 8, pp. 277–287, Aug. 2020.

[24] S. Li, X. Lin, Y. Liu, G. Li, and J. Li, "OpticGAI: Generative AI-aided deep reinforcement learning for optical networks optimization," in *SIGCOMM Workshop on Hot Topics in Optical Technologies and Applications in Networking*. New York, NY, USA: ACM, Aug. 2024.

[25] Y. Teng, C. Natalino, F. Arpanaei, H. Li, A. Sánchez-Macián, P. Monti, S. Yan, and D. Simeonidou, "DRL-assisted QoT-aware service provisioning in multi-band elastic optical networks," *Journal of Lightwave Technology*, vol. 43, no. 19, pp. 9090–9101, Oct. 2025.

[26] R. Koch, S. Kühl, W. Schairer, B. Spinnler, and S. Pachnicke, "High-generalizability reinforcement learning based capacity optimization in wdm long-haul networks," *IEEE Photonics Technology Letters*, vol. 34, no. 17, pp. 891–894, Sep. 2022.

[27] M. Salani, C. Rottondi, L. Ceré, and M. Tornatore, "Dual-stage planning for elastic optical networks integrating machine-learning-assisted QoT estimation," *IEEE/ACM Transactions on Networking*, vol. 31, no. 3, pp. 1293–1307, Jun. 2023.

[28] N. Di Cicco, J. Talpini, M. Ibrahimi, M. Savi, and M. Tornatore, "Uncertainty-aware QoT forecasting in optical networks with Bayesian recurrent neural networks," in *IEEE International Conference on Communications (ICC)*. IEEE, May 2023, pp. 441–446.

[29] W. He, Z. Jiang, T. Xiao, Z. Xu, and Y. Li, "A survey on uncertainty quantification methods for deep learning," *ACM Comput. Surv.*, vol. 58, no. 7, pp. 179:1–179:35, Feb. 2026.

[30] "Architectural framework for machine learning in future networks including IMT-2020," ITU-T, Recommendation ITU-T Y.3172, Jun. 2019.

[31] C. Natalino, B. Shariati, P. Safari, J. K. Fischer, and P. Monti, "Analysis and mitigation of unwanted biases in ML-based QoT classification tasks," in *Optical Fiber Communications Conference and Exhibition (OFC)*, Mar. 2024, p. M1H.3.

[32] A. Mitrovska, B. Shariati, A. Jafari, P. Safari, J. Karl Fischer, and R. Freund, "Network data sharing: A governance framework for ensuring data sovereignty and privacy compliance," *Journal of Optical Communications and Networking*, vol. 17, no. 11, pp. 1019–1031, Nov. 2025.

[33] O. Ayoub, S. Troia, D. Andreoletti, A. Bianco, M. Tornatore, S. Giordano, and C. Rottondi, "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, Jan. 2023.

[34] Q. Wang, Z. Cai, and F. N. Khan, "Lifelong QoT prediction: An adaptation to real-world optical networks," *Journal of Optical Communications and Networking*, vol. 16, no. 11, pp. 1159–1169, Nov. 2024.

[35] D. Zibar, L. H. H. de Carvalho, M. Piels, A. Doberstein, J. Diniz, B. Nebendahl, C. Franciscangelis, J. Estaran, H. Haisch, N. G. Gonzalez, J. C. R. F. de Oliveira, and I. T. Monroy, "Application of machine learning techniques for amplitude and phase noise characterization," *Journal of Lightwave Technology*, vol. 33, no. 7, pp. 1333–1343, Apr. 2015.

[36] K. Rezaei, O. Ayoub, P. Monti, and C. Natalino, "Policy-driven conformal prediction for trustworthy QoT estimation," in *Optical Fiber Communications Conference and Exhibition (OFC)*, Mar. 2026, p. M4A.3.

[37] S. A. Khader, H. Yin, P. Falco, and D. Kragic, "Learning deep energy shaping policies for stability-guaranteed manipulation," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8583–8590, Oct. 2021.

[38] J. Wu, H. Zhang, and Y. Vorobeychik, "Verified safe reinforcement learning for neural network dynamic models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 117 762–117 783, Dec. 2024.

[39] D. Corsi, E. Marchesini, and A. Farinelli, "Formal verification of neural networks for safety-critical tasks in deep reinforcement learning," in *Conference on Uncertainty in Artificial Intelligence*. PMLR, Dec. 2021, pp. 333–343.

[40] Y. Cheng, Z. Chen, Z. Hu, M. Xiang, Z. Yan, Y. Qin, and S. Fu, "Cognitive learning enabled agile optical network," *Journal of Optical Communications and Networking*, vol. 16, no. 11, pp. 1170–1178, Nov. 2024.

[41] 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)*, Mar. 2024, p. W4I.6.

[42] K. Rezaei, O. Ayoub, S. Troia, F. Lelli, P. Monti, and C. Natalino, "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)*, Oct. 2025.

[43] D. Azzimonti, C. Rottondi, and M. Tornatore, "Using active learning to decrease probes for QoT estimation in optical networks," in *Optical Fiber Communications Conference and Exhibition (OFC)*, Mar. 2019, p. Th1H.1.

[44] H. Maryam, T. Panayiotou, and G. Ellinas, "Learning quantile QoT models to address uncertainty over unseen lightpaths," *Computer Networks*, vol. 212, p. 108992, Jul. 2022.

[45] X. Fan, F. Ren, J. Zhang, Y. Zhang, J. Niu, and J. Wang, "Reliable optical performance monitor: The combination of parallel framework and skip connected generative adversarial network," *IEEE Access*, vol. 8, pp. 158 391–158 401, 2020.

[46] X. Fan, Y. Su, T. Dong, Y. Jie, Y. Zhang, F. Ren, J. Niu, J. Zhang, and J. Wang, "Enhancing the credibility of the optical performance monitor with adversarial training," *IEEE Access*, vol. 8, pp. 75 682–75 690, 2020.

[47] M. Xie, X. Zhang, J. Cai, and Y. Yue, "Verifiably stable nonlinear control with reinforcement-learned diffractive optical networks," *Optics Express*, vol. 34, no. 1, pp. 15–30, Jan. 2026.

[48] Q. Qiu, X. Liu, Y. Zhang, L. Yi, W. Hu, and Q. Zhuge, "TrustOPT: A trusted online optimization strategy for reliable autonomous optical networks with field-trial demonstration," *Journal of Optical Communications and Networking*, vol. 18, no. 1, pp. A44–A53, Jan. 2026.

[49] C. Santos, A. Moawad, B. Shariati, R. Emmerich, P. Safari, C. Schubert, and J. K. Fischer, "Experimental dataset for developing and testing ML models in optical communication systems," *Journal of Optical Communications and Networking*, vol. 16, no. 11, pp. G1–G10, Nov. 2024.

[50] L. Velasco, B. Shariati, F. Boitier, P. Layec, and M. Ruiz, "Learning life cycle to speed up autonomic optical transmission and networking adoption," *Journal of Optical Communications and Networking*, vol. 11, no. 5, pp. 226–237, May 2019.

[51] R. M. Morais, "On the suitability, requisites, and challenges of machine learning [invited]," *Journal of Optical Communications and Networking*, vol. 13, no. 1, pp. A1–A12, Jan. 2021.

[52] D. Wang, Y. Song, Y. Zhang, X. Jiang, J. Dong, F. N. Khan, T. Sasai, S. Huang, A. P. T. Lau, M. Tornatore, and M. Zhang, "Digital twin of optical networks: A review of recent advances and future trends," *Journal of Lightwave Technology*, vol. 42, no. 12, pp. 4233–4259, Jun. 2024.

[53] D. Sequeira, M. Ruiz, N. Costa, A. Napoli, J. Pedro, and L. Velasco, "OCATA: A deep-learning-based digital twin for the optical time domain," *Journal of Optical Communications and Networking*, vol. 15, no. 2, pp. 87–97, Feb. 2023.

[54] D. Adanza, Ll. Gifre, P. Alemany, C. Natalino, P. Monti, R. Muñoz, and R. Vilalta, "Leveraging generative AI for intent-based networking operations in network slices," *Computer Networks*, vol. 272, p. 111647, Nov. 2025.

[55] Y. Zhang, Q. Qiu, X. Liu, X. Yu, D. Fu, X. Liu, Z. Wang, H. Lin, Y. Chen, L. Yi, W. Hu, and Q. Zhuge, "AI agent for autonomous optical networks: Architectures, technologies, and prospects [Invited Tutorial]," *Journal of Optical Communications and Networking*, vol. 18, no. 2, pp. A159–A178, Feb. 2026.

[56] Y. Pang, M. Zhang, Y. Liu, X. Li, Y. Wang, Y. Huan, Z. Liu, J. Li, and D. Wang, "Large language model-based optical network log analysis using LLaMA2 with instruction tuning," *Journal of Optical Communications and Networking*, vol. 16, no. 11, pp. 1116–1132, Nov. 2024.