Opening the Black Box of Zero-Touch 6G: Sequential XAI in the Control Loop

Mohamed Readh Fentazi
Communication Systems Department
EURECOM
Sophia Antipolis, France
mohamed-readh.fentazi@eurecom.fr

Adlen Ksentini
Communication Systems Department
EURECOM
Sophia Antipolis, France
ksentini.adlen@eurecom.fr

***Abstract*—Zero-touch Service Management (ZSM) delegates the control of 6G virtualised resources to AI/ML pipelines, increasingly built on time-series models whose opacity undermines operator trust and conflicts with regulatory requirements. Existing XAI-for-networking proposals largely (i) reuse tabular methods unsuited to networking time series and (ii) treat explanations as post-hoc artefacts for humans rather than as first-class signals inside the MAPE-K loop. This PhD aims to close this two-fold gap by designing XAI native to time-series predictors on networking data and by feeding their outputs — with Machine Reasoning (MR) over a knowledge base — into the Plan stage of MAPE-K. We present the problem context, a structured state of the art on MAPE-K-based ZSM and on XAI for sequential models, three research questions, an initial methodology, an evaluation plan grounded on injected ground truth, and a three-year roadmap.**

***Index Terms*—Explainable AI, Zero-Touch Service Management, MAPE-K, 6G, time-series, Machine Reasoning.**

I. Introduction and Problem Context

The shift towards 6G is tightly coupled with *Zero-touch Service Management* (ZSM), standardised by ETSI [1], which realises end-to-end automation through closed-loop control inspired by the IBM MAPE-K (Monitor–Analyze–Plan–Execute–Knowledge) blueprint [2]. Because 6G traffic is inherently temporal — bursty traffic, flash crowds, slice lifecycle events — time-series models have become the standard tool for KPI prediction and proactive control [3]. The state of the art consistently reports recurrent architectures (RNNs, LSTMs) as the dominant family in deployed network-management pipelines and on public 5G datasets [4]–[6], while attention-based and state-space models (Transformers, Mamba) are emerging in the research literature; the methodology proposed here applies to time-series predictors as a class, and we will use LSTMs as the primary instantiation for compatibility with existing testbeds before porting to attention-based predictors in Year 3.

Two corner cases illustrate why explainability cannot be deferred to a dashboard. *(i)* An auto-scaling decision triggered by an LSTM forecast may violate an enterprise SLA: the operator must justify the action to the customer and, under the EU AI Act [7], to a regulator, but a black-box prediction provides no admissible rationale. *(ii)* A sequence anomaly detector flags a base station; without a causal explanation linking the alert to specific input sub-sequences, the operator cannot decide between hardware failure, configuration drift or benign traffic anomaly. Surveys [8]–[10] consistently argue that explainability is a technical, operational and regulatory requirement.

This work is supported by the European Union's Horizon Program under the EXACT-6G project (Grant No. 101120297).

Thesis statement. *Sequential XAI attributions, when made faithful to time-series-predictor dynamics on networking data and consumed by a Machine-Reasoning layer at the Plan stage of MAPE-K, yield ZSM control decisions that are simultaneously more accurate, more stable, and logically auditable than both prediction-only and post-hoc-explained baselines.* An early high-level formulation appeared in [11]; the present paper adds the formal research questions (RQ1–RQ3), the gap analysis of Fig. 1, the three-contribution methodology (C1–C3), and the evaluation plan.

II. State of the Art

A. XAI methods and their limits for sequential models

XAI methods are commonly categorised as global vs. local, model-agnostic vs. model-specific [12]. SHAP [13], permutation importance [14], LIME [15], attention weights and Layer-wise Relevance Propagation (LRP) are the canonical tools. For sequential models, however, gradient- and feature-based attributions conflate the time and feature domains [16]; SHAP's kernel explainer scales exponentially with sequence length [17]; attention weights are unreliable as faithful importance measures [18]; and recent work shows that different post-hoc explainers can disagree on the same input [19], undermining their use as control signals.

Sequence-tailored frameworks such as TimeSHAP [17] and TSR [16] were validated on healthcare/finance, not on networking time series, which combine heavy categorical features (slice, cell, QCI), multi-scale temporality (from μs packet-level to minute-scale orchestration) and strong spatial diversity. AIChronoLens [4], [5] partly addresses this by linking legacy XAI to temporal properties of mobile-network sequences via a Gramian Angular Field representation, but it remains *post-hoc* and *human-facing*: explanations are produced for offline

auditing, not as machine-readable signals consumed by a controller.

B. MAPE-K-based closed-loop automation and XAI in ZSM

The ETSI ZSM reference architecture [1] explicitly inherits the MAPE-K blueprint [2]. The H2020 AI@EDGE NSAP [6] embodies a closed-loop architecture with a Data Collection Pipeline (Monitor), Artificial Intelligence Functions for prediction and anomaly detection (Analyze) — including LSTM-based time-series anomaly detection [6] — and an Intelligent Orchestrator (Plan/Execute). Closely related doctoral work [20] addresses multi-criteria, data-driven resource orchestration for MEC within similar closed-loop architectures. A recent 6G zero-touch security survey [10] confirms MAPE-K-equivalent loops as the de facto target.

Across these works, three integration mechanisms recur between XAI and the closed loop: *(i) reward shaping*, where SHAP- or symbolic-derived signals enter the reward of a DRL agent [21]; *(ii) constraint incorporation*, where explainability metrics act as constraints in federated optimisation [22], [23]; and *(iii) steering*, where an explainer modifies a DRL agent's actions at runtime, as in EXPLORA for the Open RAN [24]. Complementary lines couple anomaly detection with LLMs to produce natural-language rationales for human operators [25], and recent surveys cover explainable DRL for AI-RAN and causal/RL learning for 6G stakeholders [9], [26]–[28].

Two limitations therefore systematically emerge. *First*, most XAI-for-networking works apply off-the-shelf XAI without adapting it to sequential, mixed-feature networking data [14], [19]. *Second*, when XAI is integrated into the loop, it operates over tabular feature vectors or DRL state vectors — not over the temporal attribution maps of a time-series predictor — and either guides a black-box DRL agent or is routed to a human via an LLM. The fundamental question of *why a time-series predictor produced a given temporal output, expressed in a form a Machine-Reasoning planner can consume*, remains open.

C. The gap

Figure 1 summarises this positioning. **No work simultaneously (i) tailors XAI to the sequential and mixed-feature nature of networking data and (ii) feeds the resulting sequential attributions into a MR-based Plan stage of a MAPE-K loop.**

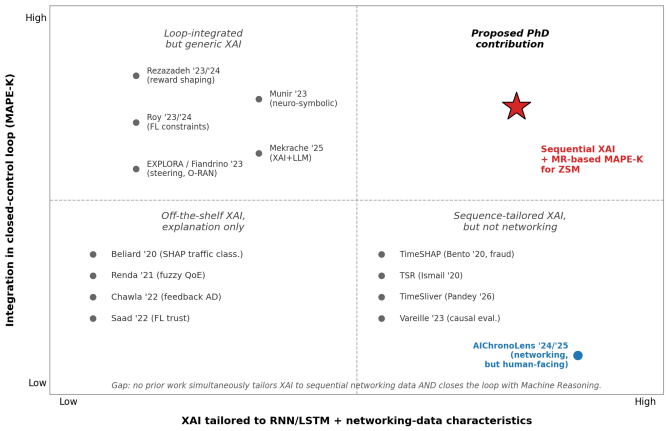

Fig. 1. Positioning along two axes: tailoring to networking time series (horizontal) and integration into a MAPE-K-equivalent loop (vertical). The proposed PhD targets the upper-right quadrant.

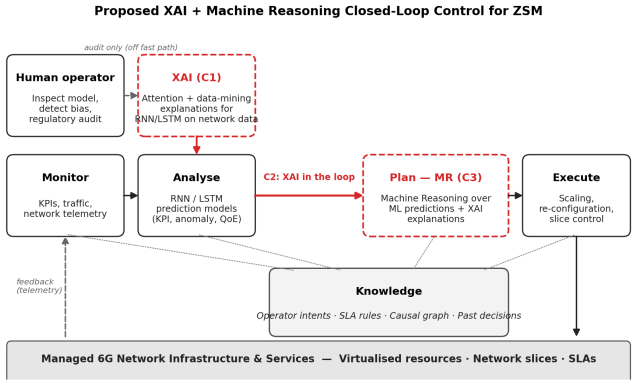

Fig. 2. Proposed architecture mapped onto MAPE-K. C1 (Analyze): sequential XAI for time-series predictors on networking data. C2: routing of attributions from Analyze to Plan. C3 (Plan): Machine Reasoning over predictions, attributions and a knowledge base.

III. Research Questions

RQ1. How can XAI methods natively capture the temporal structure of time-series predictors on networking data, and produce attributions whose faithfulness is provably higher than that of independence-assuming post-hoc methods, measured against injected ground-truth shapelets and causal proxies?

RQ2. Under which conditions does feeding XAI attributions into the Plan stage of a MAPE-K loop improve decision quality (SLA violations, OPEX, decision latency) versus prediction-only and post-hoc-explained DRL baselines, and what formal properties (faithfulness, stability, discretisability) must the attributions satisfy?

RQ3. What representation of XAI outputs and causal knowledge allows a Machine-Reasoning layer to produce decisions that are simultaneously (a) empirically competitive with reward-shaping and steering DRL orchestrators [21], [24] and (b) logically auditable against operator intents and SLA rules?

IV. Early Ideas and Methodology

The PhD comprises three coupled contributions, summarised in Fig. 2 as an extension of the MAPE-K loop [2].

C1 — Sequential XAI for networking time-series predictors. We will combine *attention-guided* scoring with *sequential pattern mining* (PrefixSpan / frequent-episode mining) over categorical sub-sequences. Attention is used as a prior [18], then validated against patterns co-occurring with the predicted outcome, yielding attributions that are both attention-guided and empirically grounded. This addresses two limitations of the closest competitors: TimeSHAP [17] lacks native

categorical-feature support and incurs exponential SHAP cost; AIChronoLens [4], [5] is post-hoc and human-facing, and does not output the discrete, machine-consumable attributions a MR planner requires.

C2 — XAI in closed-loop control. Attributions are routed into the Plan stage rather than to a dashboard. They (a) augment the state of the decision module, weighting predictions by causal support and gating actions whose explanation is unstable, and (b) expose drift to operators for offline correction, preserving auditability [7] without putting humans on the fast path.

C3 — XAI + Machine Reasoning. The Plan stage is implemented as a Machine Reasoning module [29] consuming (i) time-series-predictor outputs, (ii) sequential XAI attributions and (iii) a knowledge base of operator intents, SLA thresholds and causal relations learned from telemetry. MR produces root-cause diagnoses and rule-grounded actions. Unlike reward-shaping [21] and steering [24] approaches, the DRL agent is replaced — not augmented — by a logically inspectable reasoner.

Evaluation methodology

Faithfulness of C1. Attributions will be evaluated against three criteria: (a) faithfulness via insertion/deletion curves on held-out traces [30]; (b) stability under input perturbation [17]; and (c) causal validity, via precision/recall against ground-truth causes following [31]. To sidestep the lack of ground-truth in operator data, we will inject controlled *shapelets* into synthetic traces generated with open mobile-traffic simulators seeded by public 5G datasets, and verify their recovery.

Closed-loop performance of C2/C3. We will compare four configurations on the same testbed: (B0) prediction-only RL; (B1) post-hoc XAI on (B0) consumed by a human; (B2) reward-shaping DRL with sequential XAI signals [21]; (B3) the proposed C2+C3. Metrics: SLA violation rate, OPEX proxy, faithfulness, decision latency. To address generalisation, the evaluation will span at least *two distinct ZSM tasks* — slice resource allocation and anomaly detection — on an OpenAirInterface-based testbed.

From 5G data to 6G claims. Public datasets are predominantly 5G/B5G, a constraint shared with most XAI-for-networking studies. The methodology is architecture-agnostic: it does not assume a specific air interface, the MR layer is decoupled from the radio stack, and the knowledge base uses telemetry abstractions present in both 5G and 6G AI-native designs [28]. Year 3 will add a Transformer port and a sensitivity analysis on synthetic 6G-like traces.

V. Research Roadmap

Year 1 (current). Literature review; problem formulation; first prototype of sequential XAI for an LSTM KPI predictor with shapelet-based faithfulness evaluation (RQ1).

Year 2. Integration into a MAPE-K loop on an OpenAirInterface testbed; comparison against B0–B2 on slice allocation (RQ2); first MR module and causal-graph learning from telemetry (RQ3).

Year 3. Extension to a second task (anomaly detection); Transformer port; end-to-end evaluation; open-source release; thesis writing.

References

[1] ETSI GS ZSM 002, "Zero-touch network and Service Management; Reference Architecture," ETSI, 2019.

[2] IBM, "An architectural blueprint for autonomic computing," IBM White Paper, 2005.

[3] M. S. Munir et al., "Neuro-Symbolic Explainable AI Twin for Zero-Touch IoE in Wireless Networks," *IEEE IoT J.*, 2023.

[4] P. Fernández Pérez et al., "AIChronoLens: AI/ML Explainability for Time Series Forecasting in Mobile Networks," *IEEE TMC*, 2025.

[5] C. Fiandrino et al., "AIChronoLens: Advancing Explainability for Time Series AI Forecasting in Mobile Networks," in *IEEE INFOCOM*, 2024 (Best Paper).

[6] B. Ahlgren et al., "D3.2 Final Report on Systems and Methods for AI@EDGE Platform Automation," AI@EDGE H2020 Project, Public Deliverable, Sep. 2023.

[7] W. Guo, "Explainable AI for 6G: Improving Trust between Human and Machine," *IEEE Commun. Mag.*, 2020.

[8] S. Wang et al., "Explainable AI for 6G Use Cases: Technical Aspects and Research Challenges," *IEEE OJ-COMS*, 2024.

[9] H. Sun et al., "Advancing 6G: Survey for Explainable AI on Communications and Network Slicing," *IEEE OJ-COMS*, vol. 6, pp. 1372–1412, 2025.

[10] L. Yang et al., "Towards Zero Touch Networks: Cross-Layer Automated Security Solutions for 6G," *IEEE TCOM*, 2025.

[11] M. R. Fentazi and A. Ksentini, "Explainable AI for Autonomous Management in 5G/6G Networks: Towards Trustworthy Zero-Touch Operations," in *IEEE NFV-SDN*, Athens, Nov. 2025.

[12] T. Zhang et al., "Interpreting AI for Networking: Where We Are and Where We Are Going," *IEEE Commun. Mag.*, 2022.

[13] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *NeurIPS*, 2017.

[14] A. Altmann et al., "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, pp. 1340–1347, 2010.

[15] M. T. Ribeiro et al., "Why Should I Trust You? Explaining the Predictions of Any Classifier," in *ACM SIGKDD*, 2016.

[16] A. A. Ismail et al., "Benchmarking Deep Learning Interpretability in Time Series Predictions," in *NeurIPS*, 2020.

[17] J. Bento et al., "TimeSHAP: Explaining Recurrent Models through Sequence Perturbations," in *ACM SIGKDD*, 2021.

[18] Y. Zhu, "Applications of Attention Mechanisms in Explainable Machine Learning," *Acad. J. Comput. & Inf. Sci.*, 2025.

[19] K. Dietz et al., "Irreconcilable Differences? Investigating Consensus of Post-hoc XAI for ML-NIDS via Decomposition," in *CNSM*, 2025.

[20] N.-E.-H. Yellas, "Multi-criteria Optimization for Resource Allocation in Multi-access Edge Computing," Ph.D. dissertation, HESAM Univ./Cnam, Dec. 2023. HAL: tel-04766847.

[21] F. Rezazadeh, H. Chergui, and J. Mangues-Bafalluy, "Explanation-Guided Deep Reinforcement Learning for Trustworthy 6G RAN Slicing," in *IEEE ICC Workshops*, 2023.

[22] S. Saad, B. Brik, and A. Ksentini, "A Trust and Explainable Federated Deep Learning Framework in Zero Touch B5G Networks," in *IEEE GLOBECOM*, 2022.

[23] S. Roy, H. Chergui, and C. Verikoukis, "Toward Bridging the FL Performance-Explainability Tradeoff: A Trustworthy 6G RAN Slicing Use-Case," *IEEE TVT*, 2024.

[24] C. Fiandrino et al., "EXPLORA: AI/ML EXPLainability for the Open RAN," *PACMNET*, vol. 1, no. CoNEXT3, 2023.

[25] A. Mekrache et al., "On Combining XAI and LLMs for Trustworthy Zero-Touch Network and Service Management in 6G," *IEEE Commun. Mag.*, 2025.

[26] M. Arana-Catania et al., "Explainable Reinforcement and Causal Learning for Improving Trust to 6G Stakeholders," *IEEE OJ-COMS*, 2025.

[27] N. Wehner et al., "A Tutorial on Data-driven QoE Modeling with Explainable AI," *IEEE Commun. Surveys & Tut.*, 2025.

[28] S. Shafaei et al., "Towards AI in 6G: Concepts, Techniques, and Standards," *IEEE Access*, 2025.

[29] K. Čyras et al., "Machine Reasoning Explainability," *arXiv:2009.00418*, 2020.

[30] D. Solís-Martín, J. Galán-Páez, and J. Borrego-Díaz, "On the Soundness of XAI in Prognostics and Health Management," *Information*, 2023.

[31] E. Vareille et al., "Evaluating Explanation Methods of Multivariate Time Series Classification through Causal Lenses," in *IEEE DSAA*, 2023.