Agentic AI transforms telco network operations

WHITEPAPER

ust.com



UST

Abstract

Telecommunications networks are scaling faster than traditional operations can manage. Traffic surges, services diversify, and reliability expectations rise—yet most operators still rely on manual, ticket-driven workflows that slow response and inflate costs. The result is an operations model optimized for visibility, not safe, timely action.

This paper introduces a deployable blueprint for Agentic AI—software agents that observe, reason, and act safely across multi-vendor estates. Unlike dashboards or static automation, Agentic AI delivers service-level agreement (SLO)-gated, closed-loop operations where every action is explainable, reversible, and policy-bound.

The architecture integrates industry standards, combining 3GPP NWDAF analytics, ETSI ZSM/ENI closed-loop patterns, and O-RAN RIC control. Governance and safety are first-class, including risk tiers, guardrails, staged approvals, auto-revert on SLO regression, and continuous observability via AgentOps.

A 0–360-day adoption roadmap begins in read-only "shadow mode," progresses through proposal and approval stages, and culminates in low-risk autonomy. Key metrics include MTTR, alert-to-ticket conversion, error budget burn, energy efficiency, and compliance.

Early results are promising. Programs report order-of-magnitude reductions in alert noise, 15-30% faster recovery, double-digit RAN energy savings, and predictive prevention of critical incidents. Agentic AI enables operators to transition from dashboard-centric monitoring to intent-driven execution, enhancing resilience, reducing OPEX, and laying the groundwork for autonomous networks.

Table of contents

- Executive summary 4
- Industry context and stakes 4
- The hidden Constraint: Action, not visibility 5
- Why the current path fails 5
- What the status quo costs 6
- SLO Gated, explainable closed loops 6
- UST's Architecture and operating model 8
- Illustrative opportunities and case studies 8
- Governance, safety and observability 9
- Conclusion 10
- References/bibliography 11
- List of abbreviations 12

Executive summary

U.S. Tier-1 operators are showing cloud-native cores, virtualized RANs, and Open RAN deployments scaling rapidly. AT&T aims to achieve roughly 70% of traffic being open-capable by 2026. Verizon manages tens of thousands of vRAN sites with AI-assisted energy optimization, and T-Mobile advances AI-driven RAN innovation alongside fixed wireless expansion.

Despite this progress, operational challenges remain. Outages, ticket backlogs, and energy inefficiencies strain teams, highlighting a familiar truth: networks report data, but operators cannot always act on it safely or efficiently. Dashboards and manual workflows lag behind dynamic deployments, frequent feature releases, and heterogeneous vendor equipment.

Agentic AI embeds intelligent, policy-bound agents into workflows. They observe the network, reason within defined service-level objectives, and act through approved interfaces, automatically reverting if performance risks arise. Early results show dramatic reductions in alert noise, faster incident resolution, measurable energy savings, and predictive prevention of critical events.

By moving beyond dashboards to intent-driven execution, Agentic AI transforms network operations centers (NOCs) from reactive monitoring centers into proactive, resilient, and cost-efficient engines of network reliability.

Industry landscape and business pain points

Modern telco networks are increasingly complex. Cloud-native cores, software-defined RANs, and diverse vendor equipment increase the potential for failure points. Human-only runbooks and manual triage cannot keep pace with dynamic deployments, frequent feature releases, or vendor diversity.

Automation capabilities are maturing:

- NWDAF (3GPP) quantifies conditions, forecasts risk, and informs 5G Core operations
- ETSI ZSM provides closed-loop automation patterns across domains
- O-RAN intelligent controller (RIC) separates non-real-time policy (A1) from near-real-time control (E2), enabling fast, localized decisions.

Combined with modern observability (logs, metrics, traces, telemetry via OpenConfig/gNMI), operators can apply controlled autonomy. Agents propose changes with evidence, execute through approved interfaces, and revert automatically if SLOs regress—the outcome: safe, explainable, SLO-gated closed-loop operations.

The stakes are high: every incident that escalates to a late-night bridge impacts customer experience, operational costs, and brand reputation. Networks must detect issues and act on them safely and efficiently.

The hidden constraint: Action, not visibility

Most NOCs are awash in data: KPIs, logs, traces, fault traps, and change events stream into sophisticated dashboards. Yet seeing the problem is not the bottleneck—deciding what to do and executing it safely is.

Decision-making fragments across tools and domains; execution is slowed by brittle runbooks, ad-hoc approvals, and inconsistently rehearsed rollbacks. Error budgets rarely gate action in real time, leaving risk tolerance implicit. Predictably, alert storms become ticket storms, mean time to repair (MTTR) stretches across cross-team handoffs, and change windows remain tense.

Execution becomes even more challenging due to fragmented vendor ecosystems and isolated operational processes. While NWDAF analytics, ZSM/ENI orchestration, and RIC policies operate across different network layers, converting their insights into unified, automated actions often demands manual effort. Operators are left navigating multiple interfaces, resolving conflicting recommendations, and maintaining compliance—all under the constant pressure to reduce downtime and maintain service continuity.

Safety and reliability are rarely baked into workflows. Even small misconfigurations or delayed actions can propagate across the network, impacting quality of experience (QoE) and compliance. Auto-revert procedures, SLO gating, and staged approvals are inconsistently applied, leaving teams to balance speed against risk with incomplete information.

In short, the constraint is not a lack of visibility—operators see the network—but the inability to act decisively, safely, and consistently at scale. Networks generate data far faster than traditional operations can convert it into meaningful, reliable action. This gap underpins cascading outages, slow incident resolution, and suboptimal resource utilization.

Why the current path fails

Modern NOCs have unprecedented visibility but remain reactive. Threshold-based rules trigger cascading alarms during topology shifts or traffic surges, resulting in near-duplicate alerts that require manual correlation. Without real-time SLO and error-budget enforcement, every page feels urgent. Fragmented tools force “swivel-chair” operations; even simple mitigations await ad-hoc approvals. Rarely rehearsed rollbacks heighten perceived risk and slow cloud-native releases.

Static RAN policies underutilize energy and capacity: off-peak power settings waste kWh, and peak settings constrain QoE. Compliance tasks—such as evidence packs, least-privilege verification, and incident reconstruction—remain largely manual, diverting senior engineers from preventive work to paperwork.

What the status quo costs

When network incidents rely on late-night bridges and ad-hoc approvals, the human toll is immediate and unmistakable. Senior engineers experience burnout, and tribal knowledge is concentrated in a few individuals; as a result, hiring and retention suffer as top talent seeks a healthier work-life balance. In response, organizations adopt defensive change management: broad change freezes during peak periods, cautious rollback practices that slow recovery, and dormant feature flags.

The implicit message to the organization is clear: we do not trust our own ability to act safely and swiftly. This mistrust is amplified when customers—or social media—report service degradations before the NOC has correlated them, eroding confidence in both tools and operations.

The business impact is equally tangible. Delays in decision-making extend MTTR, increase SLA risk, and inflate third-party support costs as duplicated alerts create cascading ticket queues. Manual, overly cautious runbooks slow capacity recovery and compromise QoE. Static RAN policies leave energy and capacity on the table, while compliance tasks divert engineers from preventive work and strategic innovation.

Ultimately, the inability to act safely slows service rollout, prolongs integration timelines, and dampens leadership confidence in modernization initiatives. Every new capability adds as much uncertainty as it adds value, creating a cycle where operational risk compounds alongside opportunity costs.

SLO-gated, explainable closed loops

The challenges outlined—fragmented insight, slow and risky execution, and operational inefficiencies—highlight a critical need: a framework that converts visibility into safe, measurable, and automated action. The following approach demonstrates how Agentic AI addresses these constraints without compromising safety or governance.

At the core of agentic AI is a reference architecture following the cycle of **Sense -> Understand -> Reason -> Act -> Explain**. Each phase converts raw data into safe, measurable action:

- Sense: The network continuously emits time-aligned telemetry — 3GPP performance measurements, end-to-end KPIs, logs/metrics via OpenTelemetry, device state via OpenConfig/gNMI, and RIC KPM data. This information is normalized into a standard service and topology model, enabling event correlation across RAN, transport, and core.

Understand: Unsupervised anomaly detection, change correlation, and causal/topology analysis identify issues. NWDAF quantifies slice- and user-level impact, promoting only anomalies that threaten SLOs.

- **Reason:** Evidence is converted into an actionable plan. Options are evaluated against SLOs and error budgets, blast radius is scored, canaries selected, and rollback paths pre-computed.
- **Act:** Intents are executed via the Non-RT RIC for policy-level guidance and near-RT RIC for rapid control. SMO, OSS, and ITSM run standard procedures, with auto-revert mechanisms triggered if SLOs degrade.
- **Explain:** Every action is annotated with full provenance—standards, MOPs, tickets, and test results—ensuring auditability and generating learning signals to improve future cycles.

Safety and resilience are engineered into the system rather than assumed. Risk-tiered guardrails align with regulatory frameworks and formal governance processes, encompassing:

- Risk assessment, security controls, and ethical/compliance checks
- Real-time constraints through prompt-injection shields, PII/bias filters, and parameter validators
- Human-in-the-loop intervention for medium- and high-risk actions.

AI-driven guardrails, when properly implemented, have been shown to improve detection and recovery speeds by up to 90%, reducing the risk of unsafe actions and operational disruptions. Agents are treated as production software, monitored via AgentOps observability, which provides unified traces, metrics, logs, drift detection, resource accounting, and chargeback mechanisms. In mature deployments, this approach has demonstrated a reduction of up to 90% in downtime costs, as incidents are detected earlier, remediations are reversible by design, and unsafe actions are blocked before reaching the network.

UST's architecture and operating model

UST translates the concepts in this paper into a practical, deployable blueprint for production. The architecture is standards-aligned, vendor-neutral, and designed to scale safely, allowing modular adoption atop existing OSS/NMS, SMO/RIC, and ITSM investments without a disruptive cutover:

- SLO-first: Every decision and change is gated by explicit SLOs and error budgets
- Explainable by default: All proposals and actions carry full provenance and audit trails
- Reversible by design: Canaries, blast-radius limits, and auto-revert ensure safe execution
- Standard interfaces: NWDAF, ETSI ZSM/ENI patterns, O-RAN (A1/E2), OpenTelemetry, OpenConfig/gNMI
- Human-in-the-loop tiers: Autonomy scales with risk and evidence.

UAR SmartOps (Agentic AIOps for Telcos) and AgentOps Fabric

UST's execution-first foundation connects insights directly to safe, reversible action:

- UST SmartOps capabilities: Anomaly detection, storm correlation, evidence-cited RCA, predictive faulting, and SOP/runbook execution
- Action safety: Interventions are SLO-gated, canaried, and auto-reverted if SLOs regress; parameter and policy validators block unsafe changes
- Observability spine: A unified data plane normalizes logs, metrics, traces, and telemetry across RAN, transport, and core
- AgentOps controls: Unified decision tracing, pre-production shadow runs, drift detection, resource accounting, and chargeback for measurable autonomy.

Responsible AI governance for operations

Governance is embedded to scale autonomy safely:

- Formal decision path: Risk assessment, security/safety protocols, ethics review, compliance checks, and transparency reports for each agent and use case
- Runtime guardrails: Prompt-injection shields, PII/bias filters, output and parameter validators, policy boundaries per domain
- Human-in-the-loop tiers: Low-risk routines run automatically; medium risk requires approval; high risk demands dual control and tight rollback
- Auditability: All recommendations and actions reference standards, MOPs, and tickets, creating durable evidence for regulators and internal audit.

Data and integration accelerators

Accelerators streamline integration and improve decision quality by closing context gaps:

- UST MPulse: single-window ingestion and enrichment, including unstructured data, enabling safe incident correlation, approvals, and actuation
- R&S®SSV: automated FCC/CPUC evidence packs and reporting workflows to reduce compliance toil
- Touchstone: CASS/core synthetic benchmarking correlates probes with live load and change events to anticipate incidents and propose mitigations.

RAN/edge control, energy optimization, and 6.6 operating model

UST operationalizes safe edge control and outcome-aligned operations:

- RIC-aligned control: rApps/xApps dynamically adjust PRB, power, beam, and deep-sleep windows; Non-RT policies supervise near-RT actions; QoE is protected by SLO-gated auto-revert
- Cross-domain actuation: SMO/OSS/ITSM connectors execute runbooks across transport and core with canaries and blast-radius limits
- Centralized automation platform: platform SLOs, change catalogs, and service credits align incentives; delivery can be managed centrally or federated with the customer's NOC
- Proven impact: reference programs show significant MTTR reductions, energy savings, and high platform availability, translating architecture into measurable OPEX and resilience gains.

Governance, safety, and observability

Effective autonomy requires more than intelligent agents—it demands a framework that manages risk, makes actions transparent, and holds every intervention accountable. In mature telco environments, governance is embedded at every operational layer, enabling agentic AI to scale safely without introducing unforeseen risk.

Risk tiering and approval paths determine how autonomy is applied. Low-risk routines can run automatically, medium-risk actions require staged approvals with full evidence, and high-risk interventions follow dual-control measures with tighter rollback windows. Codifying these thresholds allows operators to expand autonomy without compromising network stability or QoE.

Safety is built directly into the agents. Runtime guardrails—prompt-injection shields, PII and bias filters, and output and parameter validators—keep interventions within policy boundaries. Auto-revert mechanisms undo actions that threaten SLOs, protecting performance and user experience.

Observability ensures accountability. Every agent decision, from telemetry ingestion to actuation and rollback, is traced and logged, creating a complete provenance trail for internal review and regulatory compliance. Metrics monitor safety, quality, reliability, and cost, while drift detection, offline replay, and A/B testing drive continuous improvement. Chargeback mechanisms make the costs and benefits of autonomy visible across teams, and a "kill switch" enables immediate shutdown if SLOs are breached. Post-incident reviews reconstruct decisions to validate safety and efficacy.

Together, these practices transform agentic AI from a theoretical concept into a reliable operating system. Operators transition from reactive dashboards to proactive, SLO-gated execution, characterized by transparency, measurable outcomes, and confidence under high-stakes conditions. Networks become truly managed, with resilience, compliance, and operational efficiency built in.

Conclusion

Agentic AI is not another dashboard; it is the operating system for telco action. By anchoring every intervention to explicit SLOs, enforcing guardrails and approvals, and providing full auditability, operators can shift from reactive visibility to intent-driven, safe, and reversible execution.

Key reasons to act now:

- Networks are already evolving: Cloud-native cores, RAN upgrades, and Open RAN adoption
- Delay compounds cost: Prolonged MTTR, wasted energy, and slow change cycles
- Governance and regulatory expectations are rising
- Early unification of telemetry and provenance creates defensible automation advantages
- Competition is already learning fast, turning cloud-native flexibility into operational leverage.

UST enables operators to accelerate adoption, reduce operational risk, and achieve tangible results. Programs have demonstrated significant reductions in alert noise, 15-30% improvements in MTTR, measurable RAN energy savings, and lower compliance overhead. By implementing standards-aligned, policy-bound, and fully observable agentic AI, carriers move beyond dashboards to actual intent-driven operations.

Unlock the full potential of your network and transform how your teams operate: explore how UST SmartOps for Telcos can help your organization achieve safer, faster, and more efficient telco operations at scale.

References

1. AI-RAN Alliance. (2024, February 26). Industry leaders in AI and wireless form the AI-RAN Alliance.
2. AI-RAN Alliance. (2025, February 27). AI-RAN Alliance celebrates its first anniversary at MWC 2025.
3. Analysis Mason. (2025, May) Open Network Index: 2025 Results.
4. AT&T. (2023, December). AT&T collaborates with Ericsson to accelerate Open RAN in the U.S.
5. Capgemini Research Institute. (n.d.). Industrialization of RAN energy saving for greener connectivity at scale.
6. Capgemini Research Institute. (2024). Networks with intelligence [Report].
7. DevOps Research and Assessment (DORA). (2022). Accelerate: State of DevOps report 2022 [Report].
8. Ericsson. (n.d.). Ericsson Mobility Report (latest edition hub).
9. European Union. (2024). Artificial Intelligence Act (Regulation (EU) 2024/1689). EUR-Lex.
10. Federal Communications Commission. (2025, June 24). Disaster Information Reporting System (DIRS).
11. Google Cloud. (n.d.). Alerting on your error-budget burn rate.
12. Google SRE. (n.d.). Alerting on SLOs.
13. Google SRE. (n.d.). Error-budget policy for service reliability.

14 Google SRE. (n.d.). Implementing SLOs.

15 Legal Information Institute. (n.d.). 47 C.F.R. § 4.18 — Mandatory DIRS reporting. Cornell Law School.

16 Mobile World Live. (2025, February). AI-RAN Alliance momentum grows.

17 NEC. (2023). Reducing energy consumption in mobile networks. NEC Technical Journal.

18 National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) [NIST AI 100-1].

19 National Institute of Standards and Technology (NIST). (n.d.). NIST AI RMF Playbook.

20 National Institute of Standards and Technology (NIST). (2024, July 26). AI Risk Management Framework: Generative AI profile (overview).

21 TM Forum. (2024). IG1443 — Autonomous Networks implementation guide (v7.0.0).

22 TM Forum. (n.d.). Autonomous Networks toolkit.

23 TM Forum. (2024). Autonomous Networks framework (IG1218F v2.0.0).

24 Verizon. (2022, September 12). Verizon deploys more than 8,000 vRAN cell sites; rapidly marches toward goal of 20,000+ by end-2025 [Press release].

List of abbreviations

5G SA — 5G Standalone

5GC — 5G Core

A/B — A/B testing (controlled experiment)

A1 — O-RAN Non-Real-Time policy/intents interface

AI — Artificial Intelligence

AIOps — AI for IT Operations

AI-RAN — AI for Radio Access Networks (industry alliance theme)

AMF — Access and Mobility Management Function (in 5GC)

AN — Autonomous Networks (TM Forum context)

Auto-revert — Automatic rollback on SLO regression (not a standard acronym, included for clarity)

BLER — Block Error Rate

CASS — Core synthetic benchmarking suite (program codename in provided materials)

CPU — Central Processing Unit

CPUC — California Public Utilities Commission

E2 — O-RAN near-Real-Time control interface

E2SM — E2 Service Model (O-RAN)

EBITDA — Earnings Before Interest, Taxes, Depreciation, and Amortization

ENI — Experiential Networked Intelligence (ETSI)

EU — European Union

FCC — Federal Communications Commission (U.S.)

FWA — Fixed Wireless Access

gNB / gNodeB — 5G base station

gNMI — gRPC Network Management Interface (OpenConfig)

GPU — Graphics Processing Unit

ID — Identifier

ITSM — IT Service Management

KMS — Key Management Service

KPI — Key Performance Indicator

KPM — Key Performance Measurement (O-RAN, e.g., E2SM KPM)

kWh — Kilowatt-hour

MOP — Method of Procedure

MTTD — Mean Time to Detect

MTTR — Mean Time to Repair/Restore

NMS — Network Management System

NOC — Network Operations Center

Non-RT — Non-Real-Time (RIC domain)

NWDAF — Network Data Analytics Function (3GPP)

Open RAN — Disaggregated/virtualized RAN architecture (industry concept)

OPEX — Operating Expenditure

O-RAN — O-RAN Alliance specifications (for Open RAN)

OSS — Operations Support Systems

P95 — 95th percentile

PDCP — Packet Data Convergence Protocol

PII — Personally Identifiable Information

PRB — Physical Resource Block

QoE — Quality of Experience

R&S®SSV — Regulatory Automation and Reporting (program codename in provided materials)

RAG — Retrieval-Augmented Generation

rApp — Non-RT RIC application

RBAC — Role-Based Access Control

RCA — Root Cause Analysis

RIC — RAN Intelligent Controller

RLC — Radio Link Control

RMF — Risk Management Framework (e.g., NIST AI RMF)

RRM — Radio Resource Management

SLA — Service Level Agreement

SLI — Service Level Indicator

SLO — Service Level Objective

SMO — Service Management and Orchestration (O-RAN/ORAN-SMO)

SON — Self-Organizing Network

SOP — Standard Operating Procedure

SRE — Site Reliability Engineering

TM Forum (TMF) — Industry body for telecom standards and best practices

TTL — Time to Live

VoNR — Voice over New Radio (5G)

vRAN — Virtualized RAN

WORM — Write Once, Read Many

xApp — near-RT RIC application

Together, we build for boundless impact

Since 1999, UST has worked side by side with the world's best companies to make a powerful impact through transformation. Powered by technology, inspired by people, and led by our purpose, we partner with our clients from design to operation. Our digital solutions, proprietary platforms, engineering, R&D, products, and innovation ecosystem turn core challenges into impactful, disruptive solutions. With deep industry knowledge and a future-ready mindset, we infuse expertise, innovation, and agility into our clients' organizations—delivering measurable value and positive lasting change for them, their customers, and communities around the world. Together, with 30,000+ employees in 30+ countries, we build for boundless impact—touching billions of lives in the process.

ust.com

© 2025 UST Global Inc.

Version 0101-20251029

U ▪
S T