ERICSSON

Lead with AI in autonomous networks

The role of AI and rApps in the shift to autonomous networks.

ericsson.com/RANahead

Introduction

The journey to autonomous networks

"Managing networks often accounts for 60% to 75% of providers' total expenses – AI can make a significant reduction."

Mobile networks have become the backbone of human society. They are the technology that fuels today's economy, and a major driver of productivity and technological innovation.

Managing and operating these networks is expensive, often accounting for 60% to 75% of providers' total expenses. The combination of generational advancement, innovative services with ultra stringent requirements, hybrid network setups, spectrum scarcity, densification, and massive MIMO demands a paradigm shift in how these networks are operated. Soon, traditional optimization processes will be insufficient for handling the scale and complexity of vital mobile networks.

Service providers also face pressure to identify new opportunities, constantly exploring new business models to identify new revenue streams and drive top-line growth. Purpose-built, high-quality, differentiated and guaranteed connectivity will be key in the most impactful business models and use cases.

Artificial intelligence (AI) is critical in driving this transformation, and Ericsson's AI initiatives are defining how the technology can create value in telecoms. Centered on the non-real-time RAN intelligent controller within the service management and orchestration (SMO) and AI-powered rApps, Ericsson's AI vision will permanently transform how networks are built, operated and monetized.

Let's find out how.

Cognition in the network

AI for RAN automation

AI-enabled RAN optimization represents a foundational pillar of future mobile networks. By modernizing data pipelines, embracing Open RAN architectures, and strategically deploying machine learning models and closed-loop automation, service providers can significantly improve performance while keeping OPEX at bay.

AI-driven RAN optimization is no longer an innovation experiment; it is a strategic necessity for staying competitive in performance, efficiency, and cost control.

An autonomous network is operated by autonomous entities. At higher levels of autonomy, humans are not in the loop, and automation processes can no longer rely on their intelligent evaluation and decision-making. Instead, the software entities taking over must make decisions based on situational awareness as well as knowledge and deep understanding of business requirements and value.

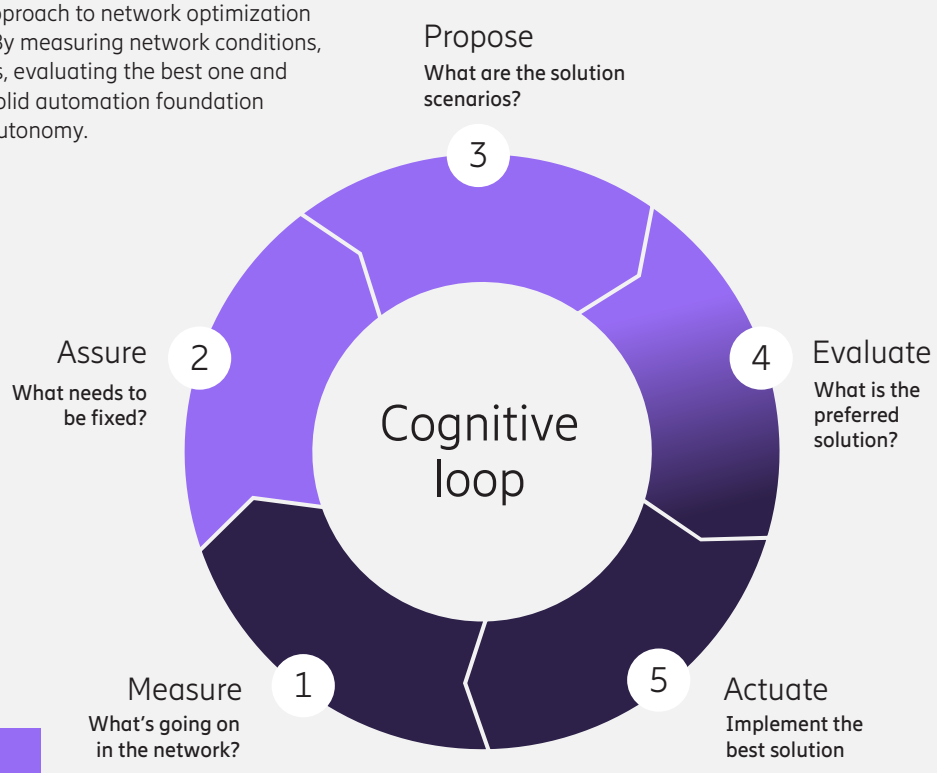
AI provides cognitive capabilities essential for autonomous network operations. AI techniques enable autonomous entities to develop situational awareness by analyzing network telemetry, traffic patterns, and performance metrics in real time. Through machine learning, these entities can recognize patterns, predict future states, and optimize network behavior based on both historical data and current conditions.

AI also facilitates autonomous decision-making by processing complex, multi-dimensional inputs, and determining appropriate actions without human intervention. Natural language processing and reasoning capabilities allow AI systems to interpret business requirements and translate them into operational objectives. As autonomous networks evolve, AI acts as the intelligence layer that bridges the gap between high-level business goals and low-level network operations, empowering networks to self-configure, self-optimize, and self-heal in response to changing conditions.

Making cognition a reality

The framework behind many of Ericsson's solutions (including rApps) is the cognitive loop—a five-phase approach to network optimization that other AI techniques build upon. By measuring network conditions, identifying issues, proposing solutions, evaluating the best one and implementing it, the framework is a solid automation foundation and takes the system to true level 4 autonomy.

Figure 1.
The cognitive loop framework



Realizing the cognitive loop, the network doesn't just think—it thinks ahead, acts decisively, and learns continuously.

Here is how the five-phase cognitive loop works:

- **The measure phase** uses specialized measurement agents to ingest multi-dimensional telemetry from OSS/BSS systems. These agents process over 50 million metrics per second through real-time feature engineering and temporal aggregation.
- **Assure agents then leverage advanced AI** and ensemble anomaly detection algorithms—combining isolation forests, autoencoders, and transformer-based sequence models—to identify performance degradations with over 95% precision. In parallel, causal root cause analyses are carried out using directed acyclic graphs to pinpoint fault propagation paths. Explainable AI shows the rational behind the decisions taken by these AI models.
- **In the propose phase**, Advanced AI and GenAI-powered optimization agents generate multiple solution scenarios. Each proposal undergoes multi-objective scoring based on KPIs, cost, risk, and SLA impact.
- **The evaluate phase** employs policies, Network Digital Twins (e.g. World Models) and hierarchical decision-making agents to resolve inter-solution conflicts. These agents automatically select. optimal actions based on learned network operator preferences and policy constraints encoded in neural networks.
- **Finally**, actuate agents orchestrate remediation through intent-based APIs and implement configurations via EIAP, while maintaining rollback capabilities.

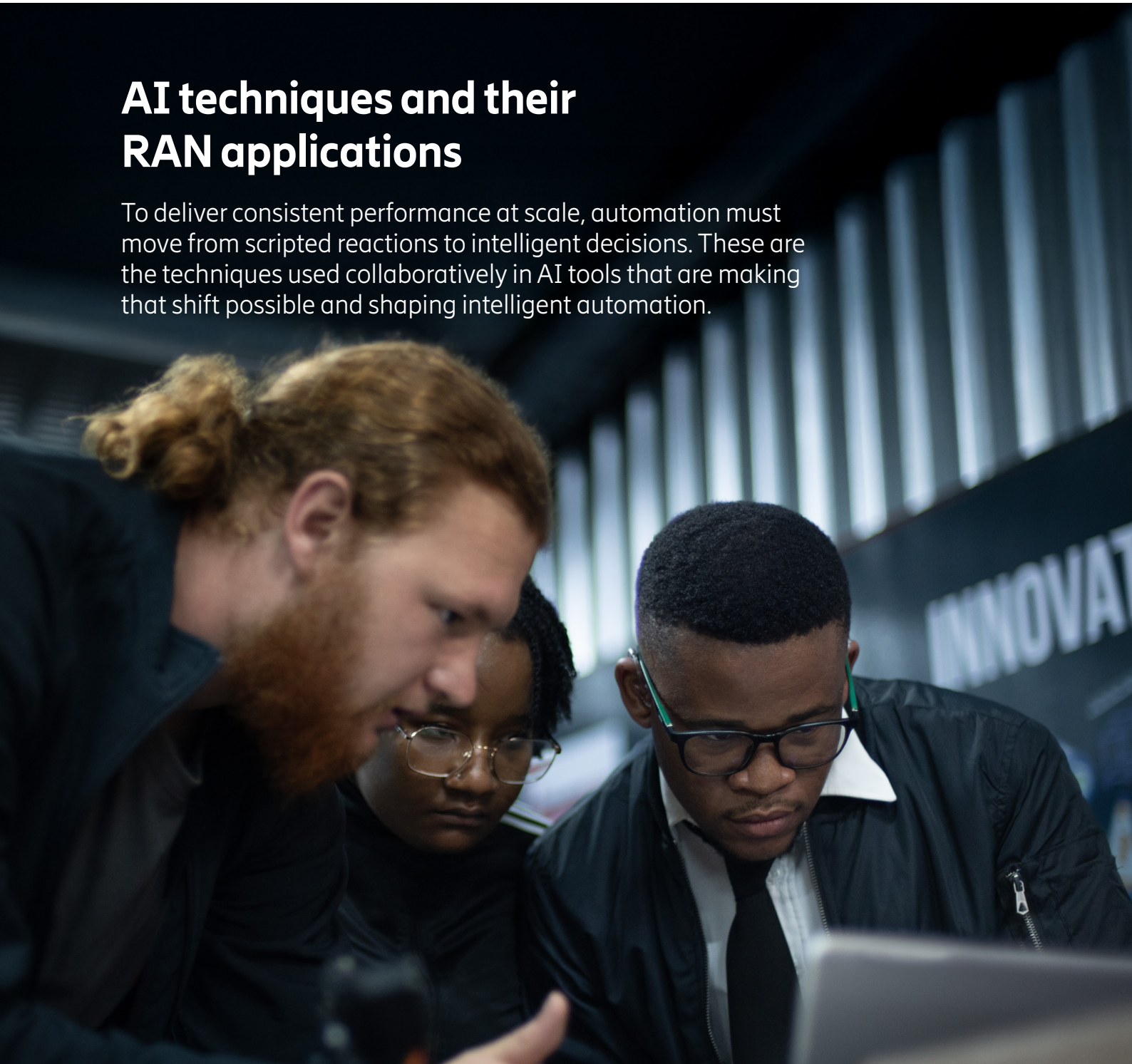
This TM Forum-aligned cognitive architecture transforms reactive network operations into proactive, self-healing systems—delivering 85% faster incident resolution and 40% reduction in mean time to repair, while eliminating 90% of manual interventions.

The shift to intent-based network operations represents a fundamental transformation in how mobile networks are managed. Rather than manually tuning individual parameters or reacting to alarms, service providers define high-level objectives—such as coverage, capacity, energy efficiency, or latency targets—and rely on AI-driven automation to translate these intents into actionable network decisions. This approach enables the network to continuously optimize itself across cells, layers, and domains, all while staying aligned with business goals and service-level agreements.

By focusing on intent rather than individual metrics, service providers gain greater agility, consistency, and resilience. This shift enables networks to adapt dynamically to changing traffic patterns, interference conditions, and user demands, while freeing engineering teams to focus on strategy and innovation rather than routine operations.

AI techniques and their RAN applications

To deliver consistent performance at scale, automation must move from scripted reactions to intelligent decisions. These are the techniques used collaboratively in AI tools that are making that shift possible and shaping intelligent automation.



Reinforcement learning

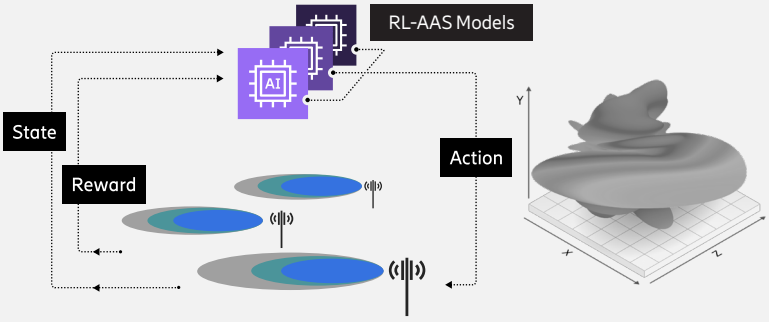
Reinforcement learning (RL) enables agents to interact with their environment and learn policies that maximize long-term accumulated rewards. Off-policy learning methods are especially valuable in dynamic and partially observable RAN environments due to their flexibility and sample efficiency. RL empowers agents, such as Ericsson's RET and AAS cell shaping rApps, to adapt to changing network conditions by leveraging ongoing feedback from the proposed actions.

By utilizing reinforcement learning, these agents can automate complex decision-making processes, reducing the time required for RF optimization by up to 75%. This streamlines operational workflows and leads to significant improvements in network performance, including enhanced coverage, capacity, and resource allocation.

RL also addresses the challenge of optimizing combinations of actions that surpass human capabilities, as the complexity and volume of possible adjustments in large-scale radio networks cannot be manually managed. As a result, RL-driven solutions enable service providers to achieve superior network outcomes that would be unattainable through traditional, human-driven approaches.

Figure 3.

Expanding capabilities with AI



Multi-objective reinforcement learning

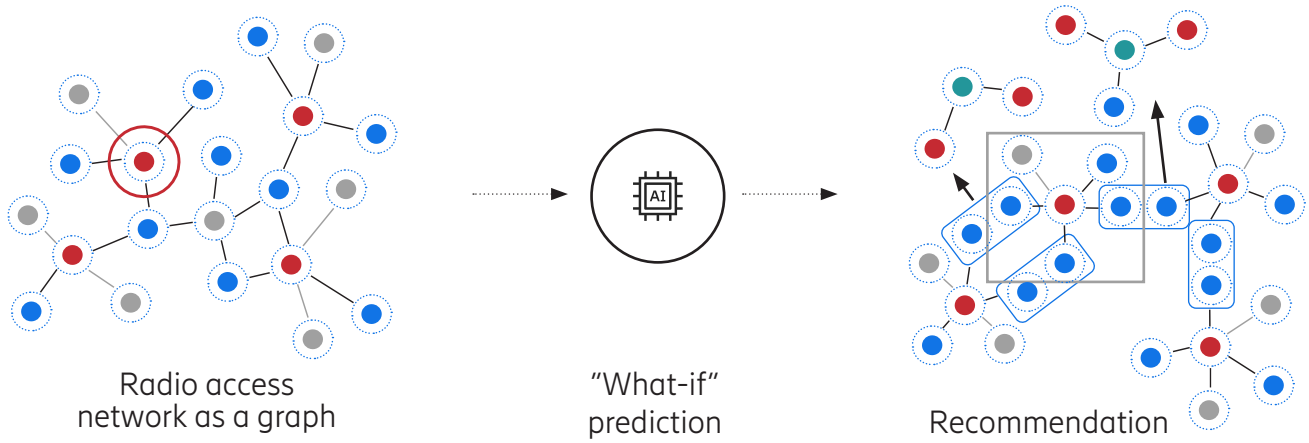
Develop reinforcement learning algorithms with reward models trained on millions of samples to strike the right balance between conflicting objectives, based on configuration changes and performance indicators.

Graph neural networks

Graph neural networks (GNNs) are effective for modeling the complex interconnections within radio networks, where cells influence each other continuously. GNNs capture geometric and relational information, enabling models to generalize across diverse network topologies and perform "what-if" scenario predictions, such as those used in digital twin world models. Ericsson's Uplink Interference Optimizer rApp demonstrates this methodology, employing a gradient-based optimizer that interacts iteratively with the GNN model within the rApp to identify globally optimal uplink power parameter configurations. Advances in spatial-temporal GNNs further enhance traffic prediction by analyzing both spatial and temporal patterns. These advancements improve the ability to anticipate congestion, optimize resources, and enable proactive management in radio access networks.

Figure 4.

The modeling process of graph neural networks



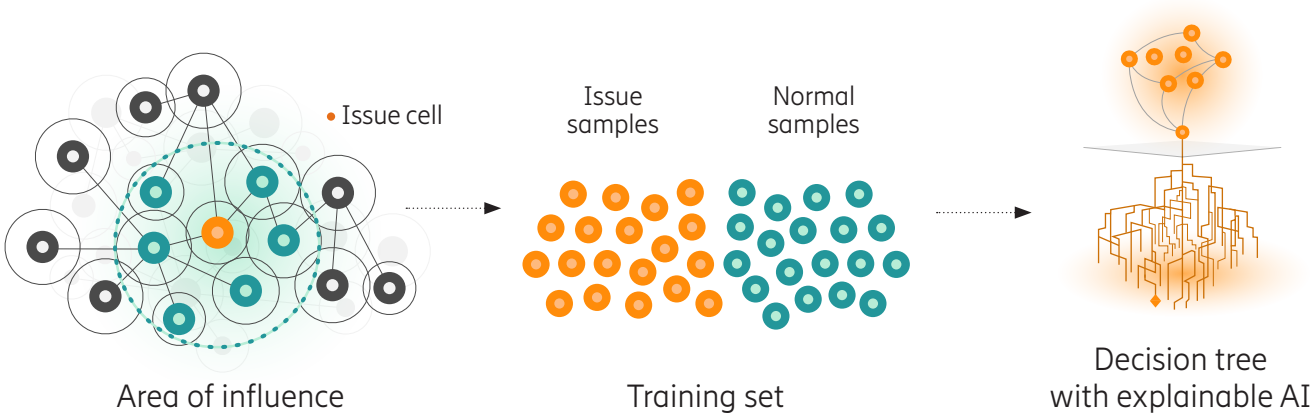
Deep probabilistic learning

Deep probabilistic learning models uncertainty by predicting probability distributions over outcomes rather than single-point estimates. This enables the quantification of prediction confidence, thereby enhancing explainability and transparency in machine learning systems.

Probabilistic outputs are essential for risk-based decision making, allowing systems to weigh the potential outcomes of different actions and allocate resources more effectively in uncertain environments. Such capabilities are especially useful in network optimization and self-healing tasks, where understanding the likelihood of different scenarios supports more robust operational strategies.

Figure 5.

The deep probabilistic learning process



World model

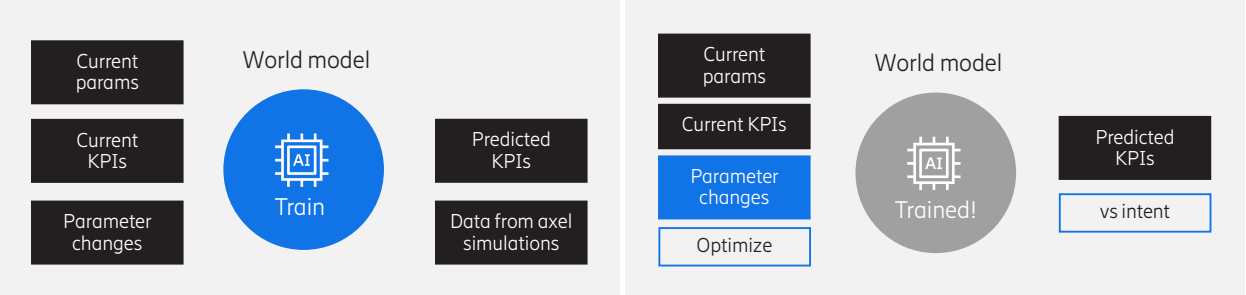
In this context, a world model is a special type of digital twin characterized by two main features. First, it is built using neural networks (NNs), which are inherently fast and also differentiable. Second, it does not carry out any bottom-up estimation of KPIs based on costly scenario characterizations. Instead, it uses current network KPIs, topology and configuration data as input to estimate how specific KPIs will change when certain configuration parameters are modified.

These traits facilitate native anchoring to the baseline situation and make world models highly suitable for "what-if" explorations, which allows optimization algorithms at the SMO level to experiment with different options before making decisions.

Beyond manual or brute-force exploration of configuration parameters, the fact that the world models are built with NNs allows for the use of standard backpropagation routines. These routines, available in any ML framework and applicable to differentiable systems, make it possible to find optimal configurations on a per-cell basis just by connectings the optimization loss function to the fulfilment degree of the intent(s). The NN parameters representing the candidate optimization settings for each cell are configured as tunable by the ML framework. In the concrete case of cellular networks, architecting the world model. As a GNN has been proven to be highly effective in capturing the interaction between neighboring nodes in a native manner.

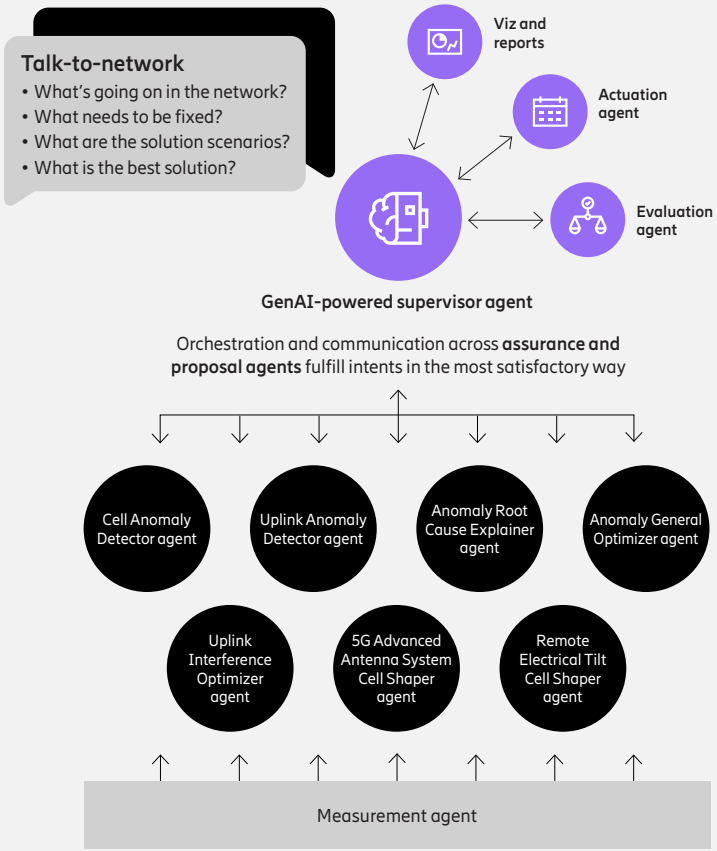
Figure 6.

Training a world model



Training a world model (left). Optimizing configuration parameters on a per-cell basis with a world model (right)

Figure 7.
The central processes behind Agentic AI rApps



Agentic and generative AI

Ericsson's Agentic AI rApps represent a paradigm shift in network intelligence, combining large language model (LLM)-based reasoning with multi-agent orchestration to achieve level 4 autonomous networks.

Our agentic architecture deploys specialized AI agents and dynamic tool orchestration capabilities. These agents autonomously plan and execute complex network operations across the planning, optimization, and healing domains. To achieve this, they rely on cutting edge AI capabilities, such as transformer-based context understanding and GNNs for topology reasoning.

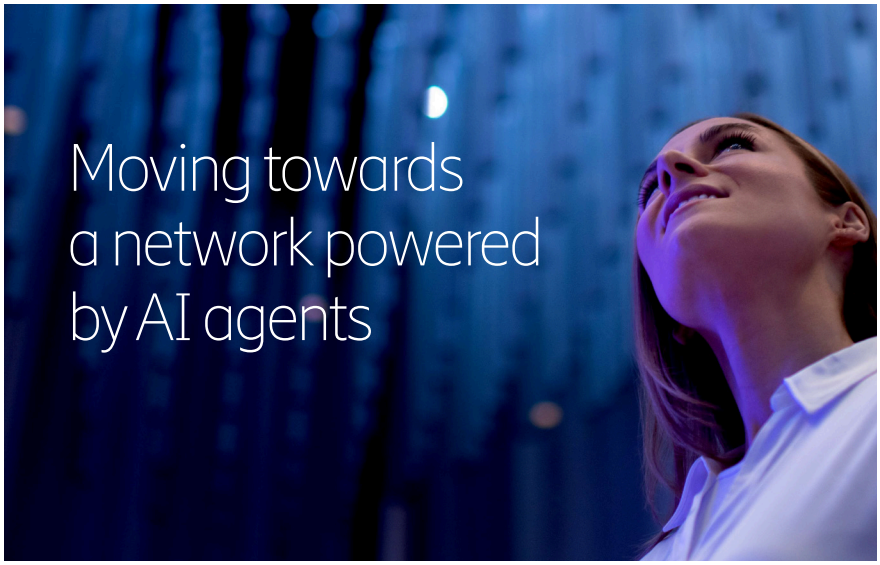
The generative AI layer enables natural language intent interpretation, translating business objectives into executable network strategies through telecom domain-specific LLMs trained on telecom-specific corpora. These models perform semantic decomposition of high-level intents, generate optimal action sequences, and synthesize human-readable explanations for autonomous decisions.

In addition, our "Talk to your Network" —a Gen-AI system for human-network interactions—allows human operatives to simplify their interactions with underlying network systems.

Simultaneously, our agentic AI framework implements goal-oriented behavior, where distributed agents negotiate with each other under the orchestration of the supervisor agent, carrying out resource allocation, prediction of network failures with high accuracy, and self-healing workflows without human intervention.

What distinguishes our solution is its cognitive closed-loop architecture: agents continuously observe network telemetry through streaming analytics pipelines, reason about causal relationships using probabilistic graphical models, and adapt policies via online learning.

This fusion of generative AI's linguistic intelligence with agentic autonomy delivers true level 4 autonomy—networks that don't just react, but anticipate, strategize, and evolve, reducing operational complexity by 70% while improving KPI performance by 35%.



Architecting the future

AI platforms for RAN automation

These AI-powered automation systems are hugely powerful, but they still need a solid foundation that supplies data and serves as an execution environment. That's what the Ericsson Intelligent Automation Platform (EIAP) aims to provide.

The O-RAN Alliance is a global community of mobile operators, vendors, and research and academic institutions with the mission to reshape radio access networks to be more intelligent, open, virtualized, and fully interoperable. By fostering collaboration among these stakeholders, the O-RAN Alliance aims to develop a more flexible, software-driven, and AI-enabled RAN architecture that accelerates innovation, reduces costs, and enhances network performance.

Within this architectural framework, the non-real-time RIC serves as the host for higher-level RAN intelligence through rApps. rApps are modular software applications that manage and optimize RAN functions in a non-real-time manner. They consume telemetry, apply analytics and policies, and recommend or trigger actions through standardized interfaces.

This design is powerful, but in practice, AI-driven automation can only deliver consistent value at scale when the underlying platform can industrialize data management, exposure, governance, execution, and closed-loop control. Building on this foundation, the rApp ecosystem gives developers access to shared APIs, documentation, test resources, and community support, and provides a pathway to onboard and list their rApps in a common directory. This helps CSPs explore and adopt AI-driven capabilities with lower integration effort, and offers developers increased visibility and an opportunity to monetize their innovations through a marketplace-style presence.

The Ericsson Intelligent Automation Platform (EIAP) brings together several SMO capabilities through several products such as the Ericsson Intelligent Controller (EIC), Ericsson Network Manager (ENM) and Ericsson Orchestrator Lifecycle Manager (EO-LM). Together, they provide four foundational capabilities that make AI both feasible and scalable:

1. Industrialized data management and exposure
AI quality is bound by data quality. EIAP centralizes data capture for RAN technologies and can ingest data from other domains. It then exposes this data through standardized services, ensuring rApps don't rebuild pipelines repeatedly.

2. Insights and analytics as reusable platform services
The platform provides higher-level insights, such as KPIs. Its ability to enrich information through other sources is one of the services that consistently feed AI/ML models across multiple rApps, increasing reuse and reducing duplicated logic and computation.

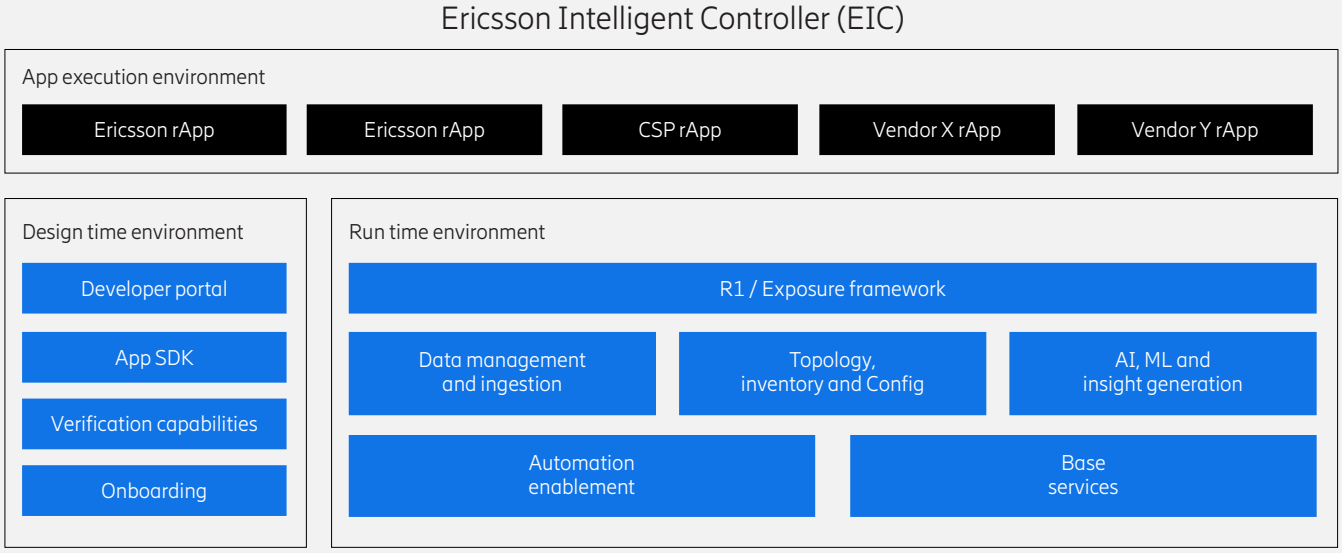
3. AI/ML lifecycle management
Beyond supporting model execution and monitoring, the platform is evolving toward full lifecycle management of AI/ML models, so these models and agents in future can be onboarded, served, tracked, and operated as first-class assets rather than "embedded code."

4. Closed-loop "point of command" for actuation
EIAP/EIC centralizes closed-loop activity toward the network, providing controlled actuation and conflict management to ensure that multiple rApps and automation loops can operate without causing destabilizing outcomes.

With these foundations, EIAP becomes a central location for consolidating automation from multiple siloed automation tools and landscapes that may currently exist at the communications service provider premises. Structured data, along with catalog and subscription mechanisms, as well as the ability to tag and enrich available data with external data sources, makes it an excellent foundation for building AI and ML rApps today and agentic rApps in the future.

Figure 8.

EIC components



EIAP supports an open rApp ecosystem, enabling onboarding and lifecycle management for rApps from Ericsson and ISVs and CSPs while maintaining the security posture and guardrails required in operator environments. In other words, the ecosystem expands as the platform reduces the cost and risk of building automation through common services, consistent exposure, and controlled execution.

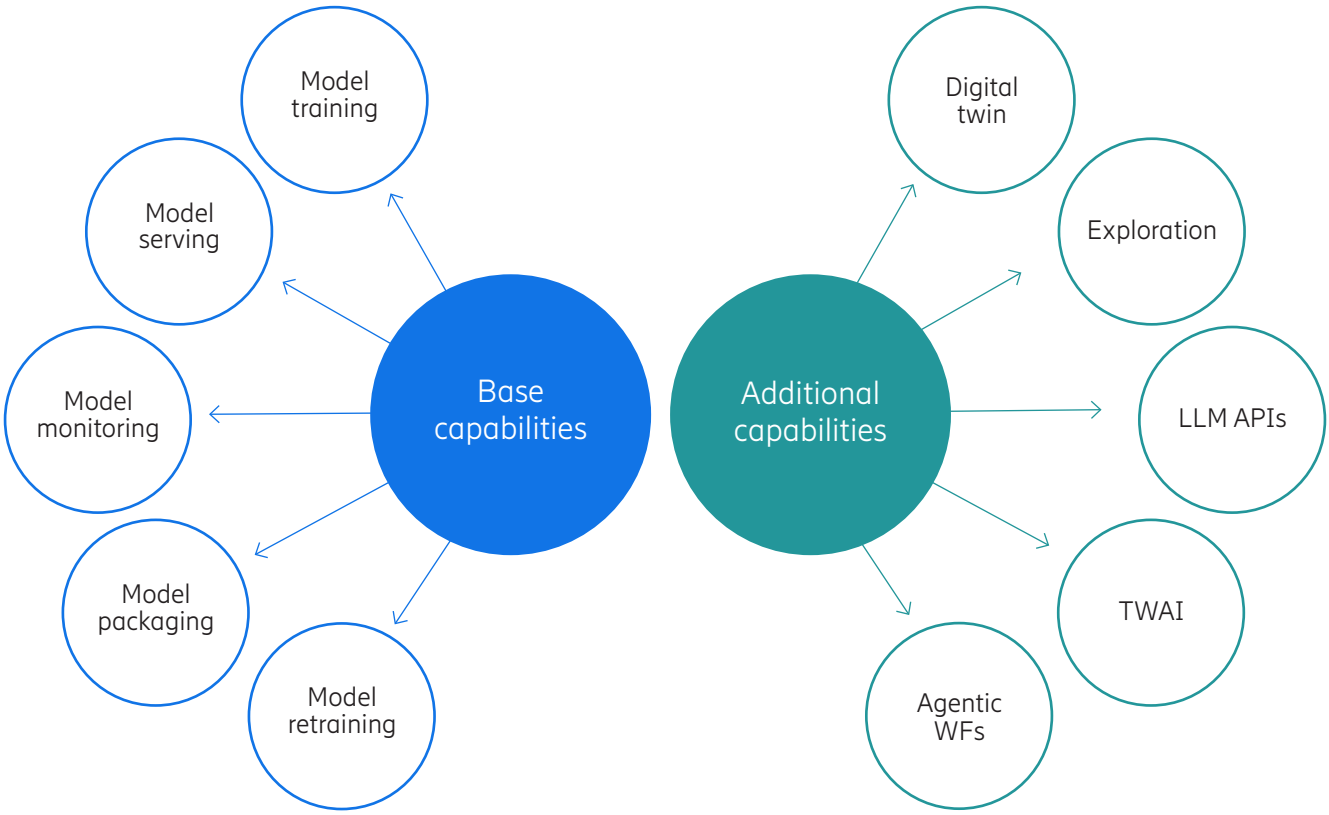
Artificial intelligence and machine learning are at the core of rApp innovation. By leveraging these capabilities, rApps can adapt to changing network conditions, anticipate traffic patterns, and propose or implement optimizations that improve performance and efficiency across the RAN.

The O-RAN Alliance defines key APIs for the entire ML lifecycle—including model registration, discovery, training, deployment, and inference—ensuring that rApps remain portable across different Service Management and Orchestration (SMO) platforms and ML providers. These APIs are exposed through the R1 interface and are documented in the Software Development Kit (SDK) as base capabilities when available.

Additional APIs under consideration are the capability registration and capability discovery APIs. They enable any ML capability provider to offer their services to rApps by extending the platform and accessing them over R1, interface, thereby increasing rApp interoperability and reuse.

Figure 9.

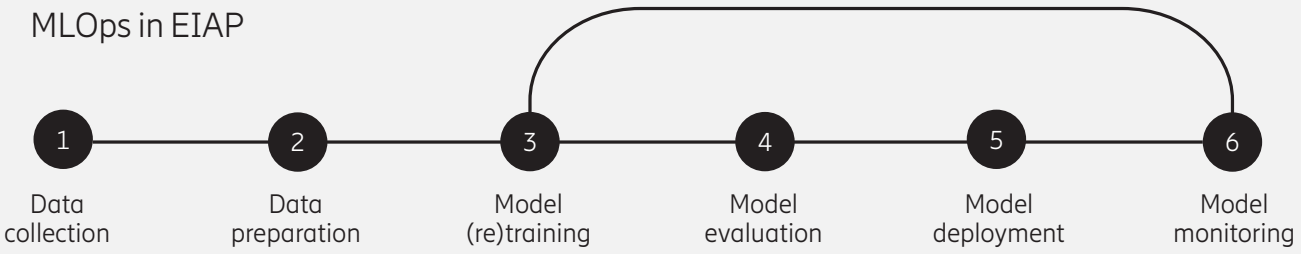
Types of ML capabilities



To run AI in production, service providers need repeatability: models must be versioned, deployed safely, continuously monitored, and retrained under governance. EIAP's direction toward lifecycle management and catalog-based reuse supports this operational discipline by enabling multiple rApps to share models and metadata while keeping operational control in the hands of the service provider.

Figure 10.

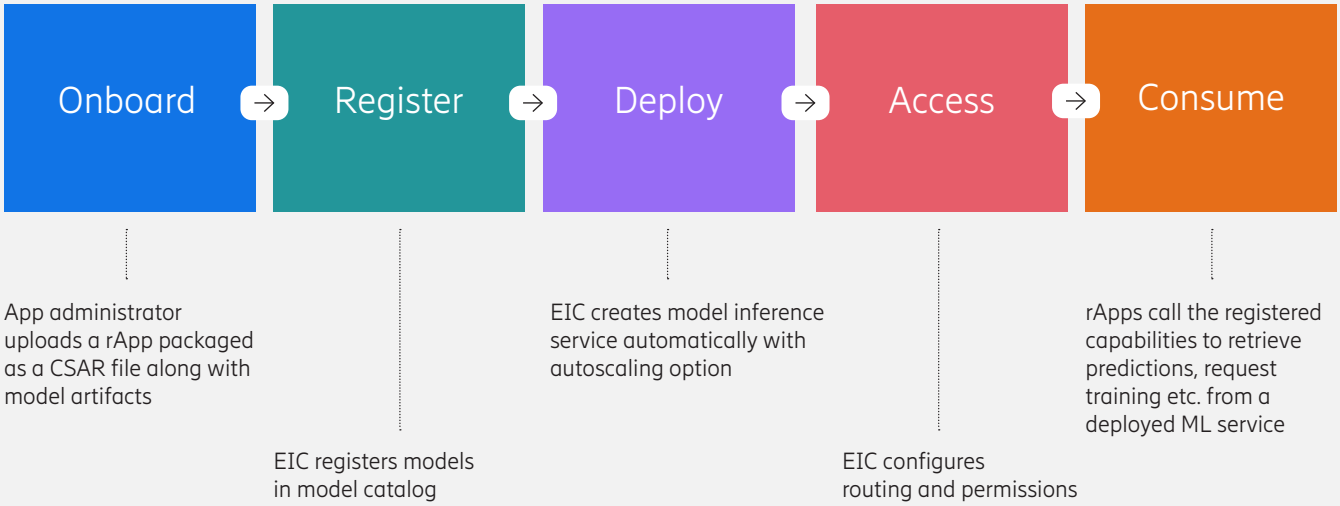
MLOps in EIAP



In EIAP, MLOps processes mirror the way rApps are handled in the platform. Models are onboarded, versioned and made available through a shared catalog for reuse across multiple rApps. Looking ahead, the platform roadmap introduces built-in support for model retraining and future capabilities such as performance monitoring and drift detection, allowing ML assets to remain accurate and up-to-date over time.

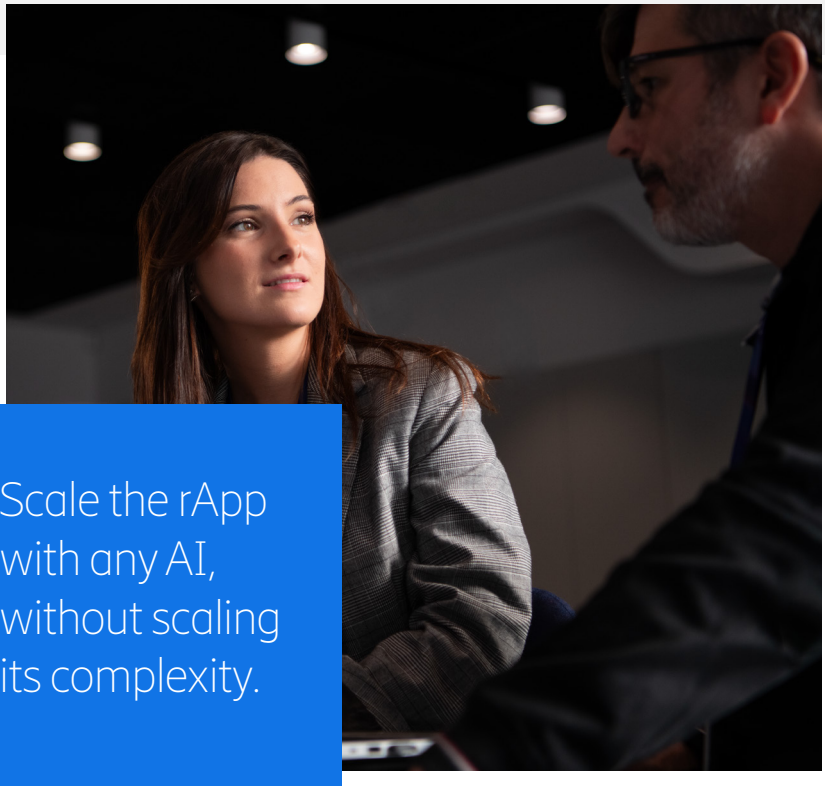
Figure 11.

Platform does the heavylifting to support ML Lifecycle



As we evolve toward rApp agents and agentic operations, the platform will extend the MLOps framework to host and serve LLMs on-premise, similar to how it currently hosts and serves traditional ML models. Also, the platform will need to host agents and agentic rApps, serving as the crucial backbone for building fully autonomous networks.

With EIAP/EIC providing the runtime foundation for AI-driven and future agentic automation, the next step is to ensure these capabilities are trusted by implementing the right guardrails, explainability, governance, and operational controls. Read on to learn how trust in automation is built using rApps.



Scale the rApp with any AI, without scaling its complexity.

Trustworthiness

Building trustworthy agentic AI for autonomous networks

The development of Agentic AI rApps for network planning, optimization, and healing requires an unwavering commitment to trustworthiness as the foundation of level 4 autonomous networks.

Recognizing that trust cannot be assumed in AI-driven systems, a zero-trust architecture is adopted alongside explainable AI (XAI) principles to ensure that every autonomous decision can be traced, understood, and validated by human network engineers.

Agentic systems incorporate comprehensive guardrails and safety shields that act as fail-safe mechanisms. These mechanisms continuously monitor AI actions against predefined boundaries to prevent unintended network behaviors, configuration drift, or service degradation.

These protective layers are essential for maintaining network integrity while enabling AI agents to operate with meaningful autonomy.

Central to our approach is the implementation of an automation maturity ladder, which recognizes that different operational contexts require different levels of autonomy. Our configurable autonomy framework—referred to as the "Agency Slider"—allows service providers to dynamically adjust the degree of AI independence based on network conditions, regulatory requirements, and organizational risk tolerance. It includes built-in maturity fallback mechanisms that automatically reduce autonomy levels when anomalies are detected or confidence thresholds are not met.

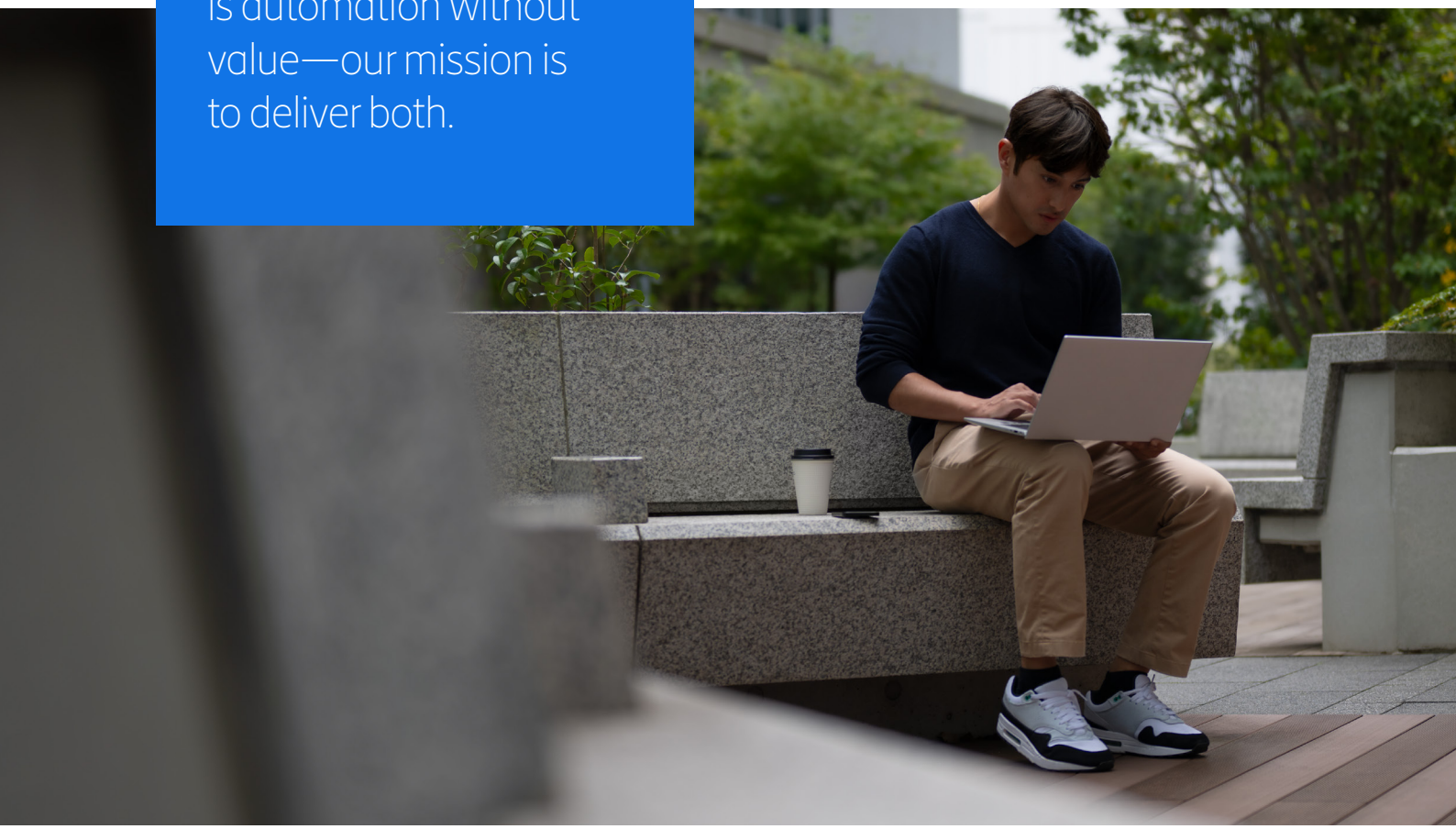
This approach ensures that organizations can progressively advance their autonomous capabilities while maintaining operational control and safety.

The interoperability of our Agentic AI systems is verified through rigorous authentication and authorization protocols, ensuring that AI agents securely interact across environments and platforms while maintaining strict identity management and access controls. This verification framework is critical for enabling agent-to-agent collaboration within complex network ecosystems, where multiple AI systems must coordinate actions without compromising security boundaries.

Foundational to all these capabilities is robust data governance and management, which establishes clear lineage tracking, quality assurance, and compliance frameworks for the data that fuel our AI decisions. We implement comprehensive policies governing data collection, retention, processing, and sharing, to ensure that our agentic systems operate on trusted, high-quality datasets while respecting privacy regulations and organizational policies.

This holistic governance approach not only enhances the reliability of AI-driven insights but also provides the auditability and accountability necessary for production deployment in mission-critical telecommunications infrastructure. Ultimately, it enables our customers to confidently progress toward truly autonomous networks.

Autonomy without trust is automation without value—our mission is to deliver both.



Adapting to the new paradigm

A new era of security

The latest generation of AI primarily interacts through unstructured natural language, general-purpose interfaces, and prompts. This introduces a new set of security challenges in the detection and policing of fraudulent behavior.

A general challenge lies in the presence of more or less concealed additional requests within an otherwise legitimate interaction. This is often referred to as prompt injection, a technique designed to make the receiving AI agent react and behave in a particular way.

How can we verify that two AI entities, when requesting tasks and sharing information, remain within their assigned roles and boundaries? Furthermore, how do we verify that the requests made by one AI entity are understood as intended by another AI entity?

In classical network operation, the active components are implementations of standardized network functions. Standardization defines interfaces specific to these functions, which are also based on formally defined information models and detailed specifications of correct interpretation and behavior. Authorization based on access to APIs and verification with formal methods is feasible.

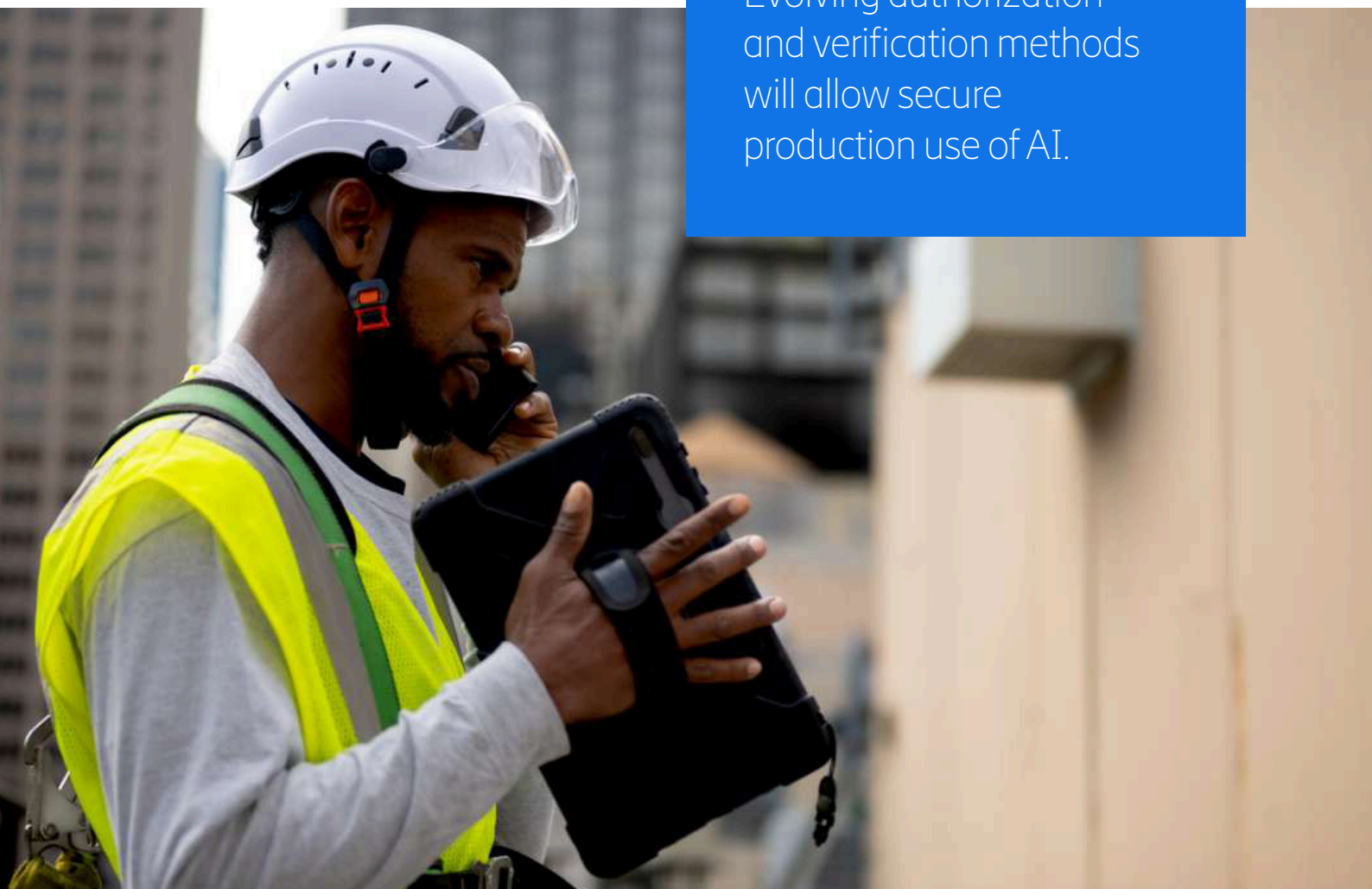
In contrast, new interfaces based on agent-to-agent (A2A) and multi-cloud platform (MCP) interactions are general-purpose interfaces that are not tied to a specific role. Furthermore, they facilitate unstructured, language-based interactions. The consequence is that authorization must become content- and topic-based rather than dependent on interface access. What tasks are requested and what data is shared must be actively controlled, as traditional formal or rules-based methods are not well-suited for managing unstructured interactions. In this sense, faulty and malicious interactions are a threat to the network, and they are equally hard to detect and handle.

There are two directions for solving these challenges. The first involves the development of new methods, which may also be AI-based. New policing models can listen to the online conversation of other AI entities to detect and mitigate threats immediately. This can be combined with integration test schemes that, based on hypothetical interactions of the AI entities, can conclude correct interpretation and interactions with sufficient certainty. These new methods and processes are not yet fully developed and understood.

This leads to the second potential solution: the continued use of formal interactions based on limited vocabulary and semantics. Domain-specific languages are a subset of natural languages, but they are easier to verify using established methods. A good example of such models is standardized intent expression models. They are already quite versatile and expressive while still being formally defined.

This discussion shows that the introduction of AI is a gradual evolution. New capabilities will be introduced while the processes ensuring secure operation continue to evolve.

Evolving authorization and verification methods will allow secure production use of AI.



Conclusion

Advancing towards autonomy

AI is becoming the intelligence layer of mobile networks, much like the control plane enabled programmability in earlier generations. RAN automation, powered by advanced AI techniques is the cornerstone of this transformation. It enables networks that are not only faster and more efficient but also self-learning, self-optimizing, and self-evolving.

As mobile networks continue to expand in scale and societal importance, AI-driven RAN automation will define their ability to meet future demands. It is not merely an enabling technology—it is the fundamental role that will shape the future of mobile networking.

Each investment in data, learning, and closed-loop control compounds over time, accelerating operational efficiency while steadily raising the autonomy ceiling. Service providers who act early will enter the next generation of networks with systems that are already adaptive, resilient, and intelligent by design.

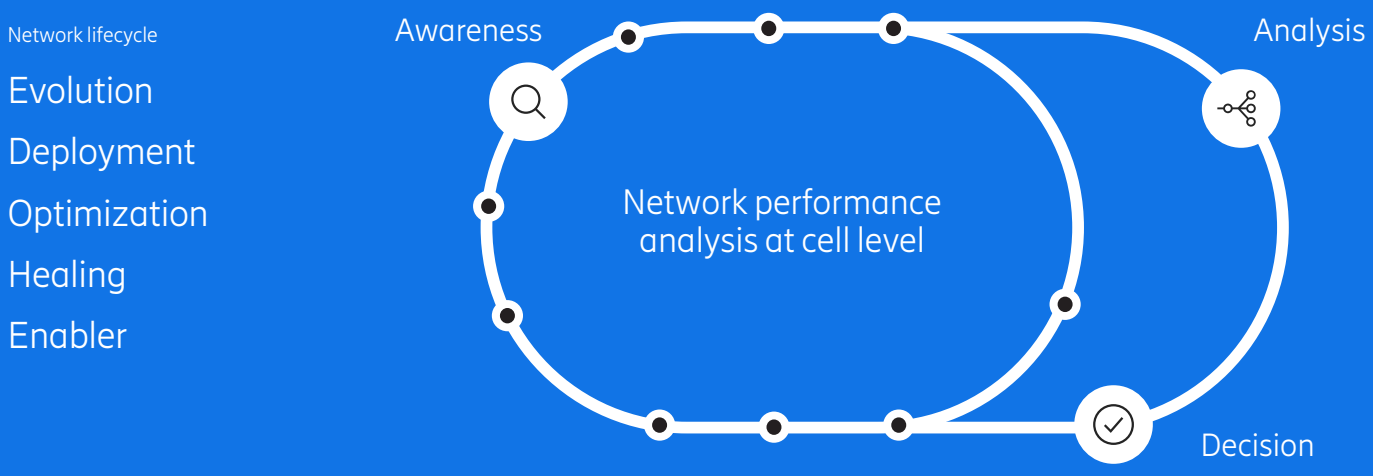
By embedding AI into RAN operations today, service providers build intelligence, data pipelines, and trust frameworks required to integrate future 6G capabilities smoothly. Such an approach transforms 6G from a disruptive step change into a natural evolution of increasingly autonomous mobile networks.

Ericsson is at the forefront of this transformation, providing the intelligence and tools that future mobile networks rely on. With its extensive portfolio of AI-native rApps, service providers are well-positioned to prepare for the next wave of technological breakthrough.

Optimize today's networks while laying the groundwork for intelligent, autonomous networks of the future by:

- **Investing in AI-driven RAN automation today**
  Deploy AI-enabled optimization, closed-loop control, and predictive analytics to improve network performance, energy efficiency, and operational scalability. Early adoption builds a foundation for future autonomous networks and 6G readiness.
- **Shifting to intent-based network operation**
  Transition from manual parameter tuning to high-level, intent-driven control. Define service objectives, and let AI translate them into actionable network decisions, ensuring consistent QoS and more agile operations.
- **Leveraging SMO and rApps for modular, scalable intelligence**
  Adopt a Service Management and Orchestration (SMO) framework and rApp ecosystem to embed AI-driven intelligence into the RAN. This approach enables modular, multi-vendor automation, rapid deployment of new optimization functions, and continuous learning across the network—accelerating innovation while maintaining operational flexibility and control.

Figure 12.
Continuous operation of deployed rApps in the network



Establish the intelligent layer of future mobile networks today

Our vision for tomorrow's reliable and highly performing networks is rooted in automation at every level. Adapting automation practices to meet the demands of tomorrow's networks is essential for providers that want to continue providing great connectivity to users.

About Ericsson

Ericsson's high-performing networks provide connectivity for billions of people every day. For nearly 150 years, we've been pioneers in creating technology for communication. We offer mobile communication and connectivity solutions for service providers and enterprises. Together with our customers and partners, we make the digital world of tomorrow a reality.

ericsson.com