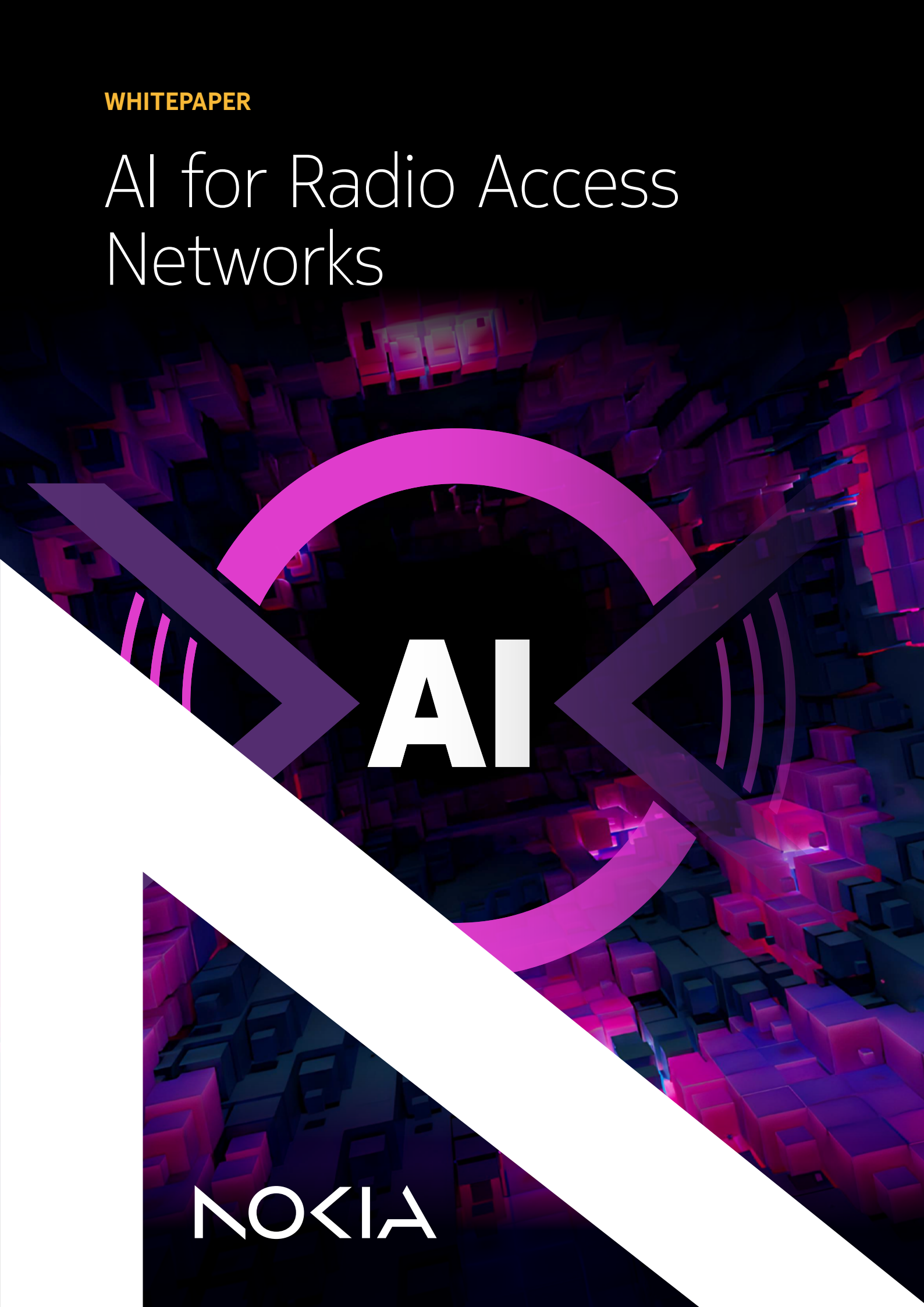


WHITEPAPER

AI for Radio Access Networks

The background of the cover is a dark, abstract 3D landscape composed of numerous small, rectangular blocks or cubes. These blocks are illuminated with a mix of deep purple and vibrant magenta light, creating a sense of depth and digital complexity. In the center of the image, the letters "AI" are prominently displayed in a large, bold, white sans-serif font. Behind the "AI" text is a large, semi-transparent magenta circle. Two large, dark purple arrows originate from the left and right sides of the circle, pointing towards the "AI" text. Each arrow is accompanied by several concentric magenta arcs, suggesting a signal or data flow. A thick, solid white diagonal line cuts across the lower-left portion of the image, starting from the bottom left corner and extending towards the center.

AI

NOKIA



Nokia's commitment to the AI-RAN vision

Artificial Intelligence (AI) along with related computational power and Radio Access Network (RAN) technologies are being adopted at a rapid pace.

These three technologies will increasingly complement one other in delivering significant improvements in operational and resource efficiency, mobile network performance and end-user experience. They will also support new exciting use cases.

Nokia is working very closely with the industry ecosystem to shape the technology evolution path towards AI-RAN. We share the vision of the AI-RAN Alliance to drive RAN performance and capability enhancements with AI.

Nokia was one of the founding members of the AI-RAN alliance and its commitment extends to representation on the board. Nokia chairs one of the key working groups that define the requirements, use cases, design architecture and solutions. We also have significant participation in other working groups and lead the AI-RAN marketing steering committee.

In addition, Nokia has publicly announced collaboration with some of the world's leading communications service providers (CSP) and other technology leaders to advance AI-RAN.

This work is supported by Nokia's research arm, Nokia Bell Labs, which has an unparalleled record in driving next-generation technology innovation.

In this white paper, we explore how and where AI will be increasingly used to enhance RAN efficiency and performance. This is not limited to a 6G vision, but also captures how AI is increasingly being used in today's 4G and 5G networks.

The evolution towards AI-RAN

Every cellular generation has been framed by the intended subscriber services, the available frequency spectrum and the affordable computational complexity of the decade.

The new frequency spectrum for 5G-Advanced and 6G will provide bandwidth for emerging immersive and extended reality (XR) services for consumers and enterprises.

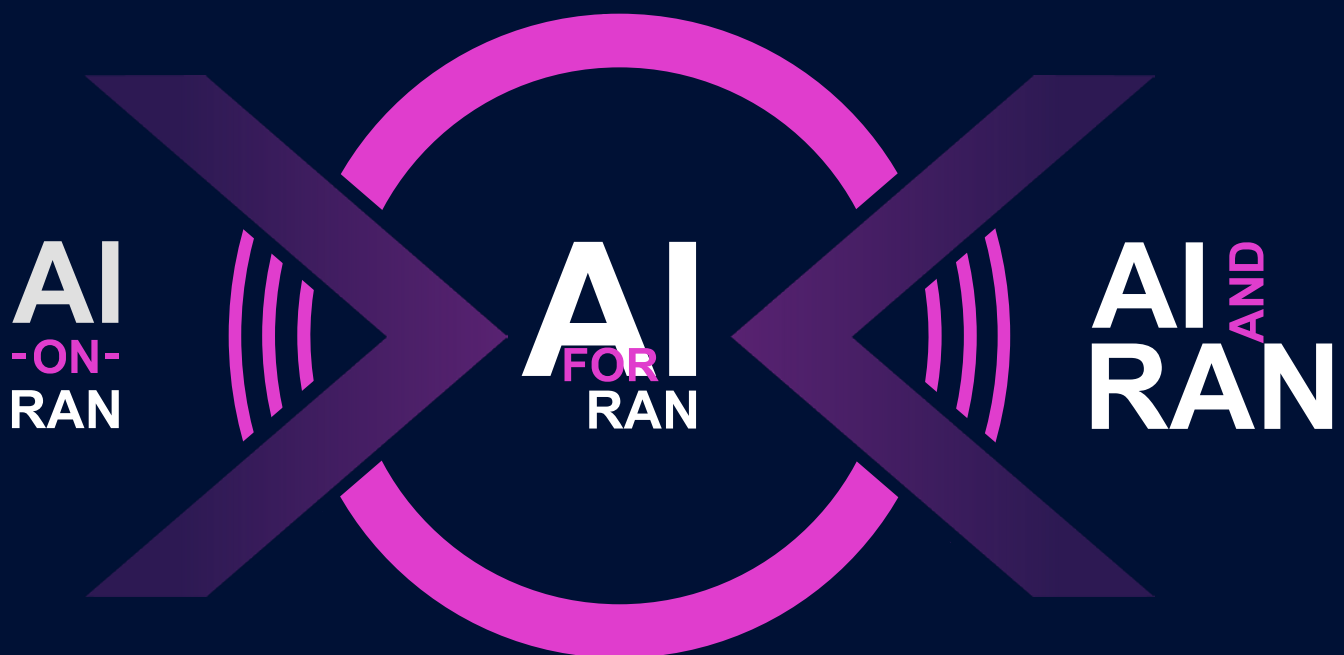
The progress of System-on-Chips (SoC), central processing units (CPU) and graphics processing units (GPU) will allow advanced RAN software, including AI, to get the most out of the given spectrum, cell sites and energy budget.

As the energy efficiency of tensor-optimized computing evolves, solving more complex computational tasks with heavier AI workloads becomes economically viable.

The role of AI in radio access networks is often described using three categories, which the AI-RAN Alliance has adopted:

- **AI for RAN**, which means using AI to enhance the performance of RAN in areas such as spectral and energy efficiency.
- **AI on RAN**, which means that applications can take advantage of AI within the RAN, leveraging capabilities such as Slicing and open APIs.
- **AI and RAN**, which means that the RAN and AI workloads share computing infrastructure.

This paper focuses on AI for RAN.



AI for RAN

AI developments during the last decade

Three factors have particularly contributed to the emergence and tremendous growth of AI in many domains:

- Big data: collected from the web, social media, and sensors.
- Tensor-efficient computing: GPUs enabling efficient computing of vectors, matrices and multi-dimensional data structures (tensors) for complex calculations.
- Cloud-based services: scalable access to data and storage, AI models and computing for training and inference.

The business model of hyperscale companies has provided a platform for massive university and company research and development. This, in turn, has led to an ever-growing toolbox of open-source and commercial tools, which enable other industries to adopt AI without reinventing the enabling software.

Figure 1 presents a timeline of the most significant developments in the dynamic AI landscape in the past decade.

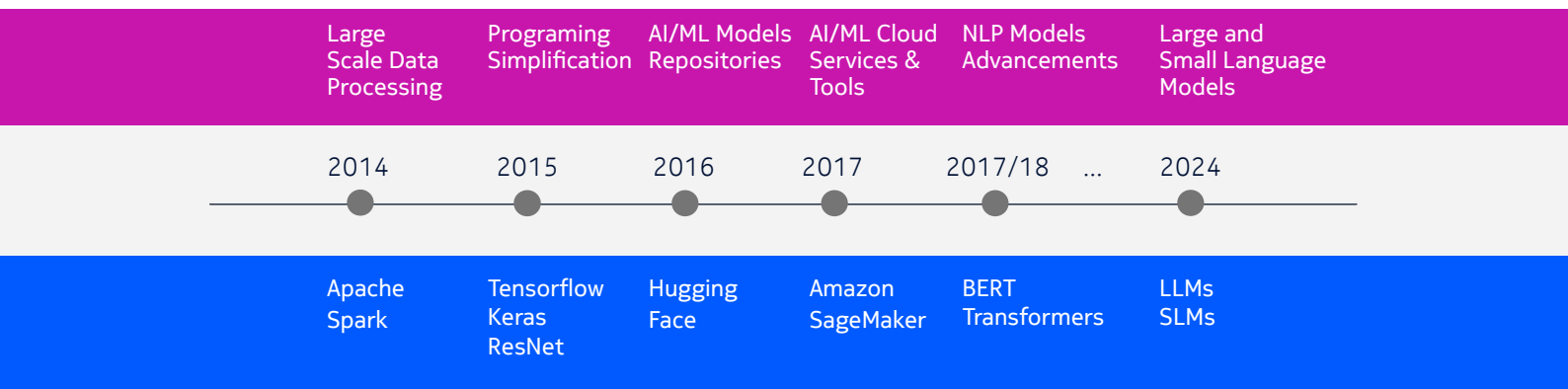


Figure 1. Timeline of AI developments

In particular, the introduction of large language models (LLM) represents a significant leap in generative AI’s capabilities. LLMs have huge computing requirements and high energy consumption. These factors will continue to drive further development of AI computing efficiency.

It will benefit other industries by improving efficiency, performance and customer experience.

RAN developments during the last decade

In the past decade, we have observed several key developments in the evolution of RAN.

There has been a massive increase in data volumes processed in radio access networks driven primarily by the surge in video consumption and the increasing smartphone penetration. Mobile broadband connectivity has become the norm, and several new use cases, such as Fixed Wireless Access (FWA) and IoT connectivity, are becoming increasingly widespread.

The increase in available frequency bands has accelerated the introduction of 5G, which marks a significant leap in performance and efficiency compared to 4G. Advancements such as Massive MIMO have enabled spectral efficiency gains.

The number of parameters used to optimize network performance and efficiency has significantly increased. We have also begun to see the opening of base station computing capabilities through Cloud RAN technologies.

There is a renewed industry focus on operational efficiency and energy efficiency based on the global increase in the cost of energy powering these networks and the ambitious environmental targets of companies and society.

Matching AI techniques with RAN tasks

Internally, RAN runs on structured data, which allows for efficient conventional and AI algorithms.

When we have a good understanding of a process, both the conventional algorithm and the initial AI model are typically good. The higher the unpredictability and complexity, the more likely it is that AI - or Machine Learning (ML) to be more precise - will outperform conventional algorithms after sufficient training and fine-tuning.

Table 1 provides some examples of how typical RAN computing tasks could be solved with fundamental ML techniques.

In practice, these techniques are often combined, for example, unsupervised learning can be implemented using a deep neural network with convolutional layers. These tasks range from real-time to offline execution requirements.

RAN computing tasks	ML technique	Description
Traffic prediction, channel quality estimation, signal quality predictions	Supervised Learning	Uses labeled data to make predictions and classifications
Anomaly detection and clustering of user behaviors	Unsupervised Learning	Finds patterns in unlabeled data
Dynamic radio resource allocation, self-organizing networks, power control optimization	Reinforcement Learning	Interacts with an environment and learns through trial and error
Enhanced signal processing	Deep Learning	Employs multi-layer neural networks for complex pattern recognition
Traffic forecasting, predicting user mobility and traffic patterns.	Recurrent Neural Networks (RNN)	Processes sequential data and learns from past information
Channel equalization and interference cancellation	Convolutional Neural Networks (CNN)	Analyzes data with grid-like topology (images, signal matrices)
Chatbot agents for network operation	Generative AI	Creates new data instances similar to training data patterns

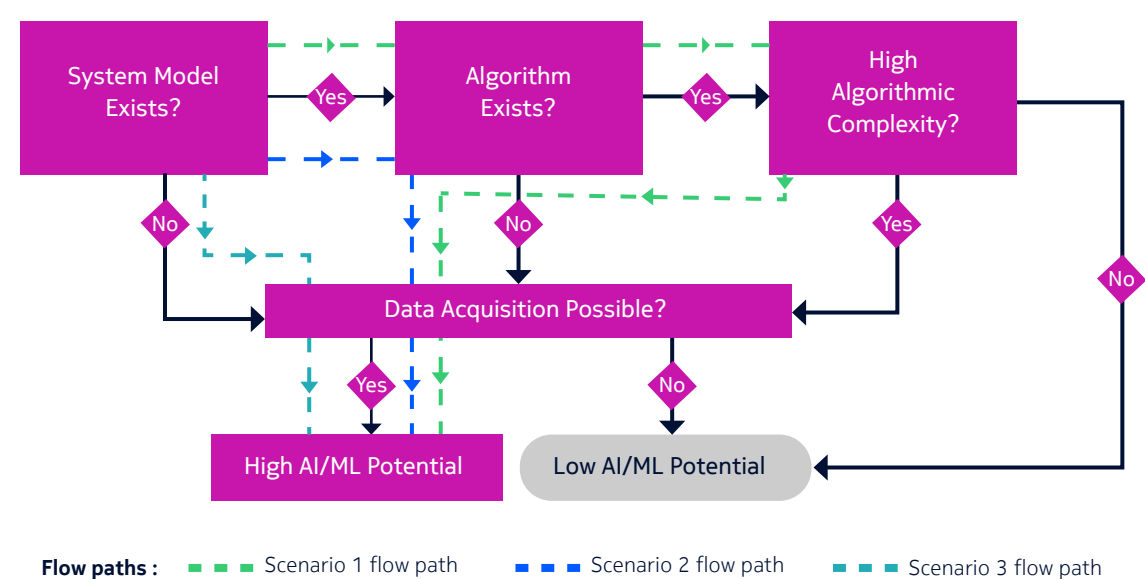
Table 1. Typical RAN computing tasks and related fundamental ML techniques

AI/ML outperforming conventional algorithms

Researchers at Nokia Bell Labs [1] developed a decision tree framework to evaluate the viability of AI/ML for a given system problem. Figure 2 illustrates this framework.

The framework considers four different questions:

1. Is there a mathematical model that describes the system?
2. Does an algorithm that can solve mathematical models exist?
3. Does the algorithm's complexity permit practical implementation?
4. Is there sufficient data available to train a robust ML model, or is data acquisition possible to obtain the required number of data pairs?



High AI/ML Potential

Scenario	System Model	Algorithm	Complexity	Data
1	•	•	•	•
2	•	x		•
3	x			•

Low AI/ML Potential

Scenario	System Model	Algorithm	Complexity	Data
4	•	•	•	x
5	•	•		
6	x			x
7	•	x		x

Figure 2. Decision tree framework for AI/ML feasibility evaluation

The decision framework presents three scenarios in which the use of AI has high potential. In all three scenarios, sufficient quality data must be available to build an AI model and algorithm.

In the first scenario, the system model and conventional algorithmic specification theoretically enable the use of a conventional algorithm, but the compute cost (or speed) of implementing it is prohibitive. Hence, the AI model has the potential to enhance or replace the conventional methods.

In the other two scenarios, either a conventional algorithm or a sufficiently accurate system model has not been identified. However, enough data can be obtained to build an algorithm and train a robust ML model. From IT and cloud, we know the operational efficiency mantra “centralize what you can, distribute what you can only do there.” In a radio access network, this translates into: Do what you can on network management and orchestration level, do in the base station what you can only do there.

Centralized approach or AI inside the base station

Real-time tasks need to be in the base station, regardless of whether the algorithm is conventional or AI/ML-based. In today’s networks, non-real-time tasks are typically handled in centralized network management and self-organizing networks (SON) systems.

Following a similar approach, centralized AI systems will help in operational efficiency, while AI inside the base station adds the capability to improve performance and efficiency at the cell level, which enables enhancements for spectral efficiency and user experience.

An advantage of AI execution in the base station is that the data does not need to be exposed further. This can be essential for meeting regulatory requirements and also for speeding up implementation in the network.

An advantage of the centralized approach is that the AI system has access to data from the entire network instead of just cell-level data. This approach also facilitates the use of commercial-off-the-shelf (COTS) data center equipment.

Figure 3 illustrates the different options for running AI/ML in the network.

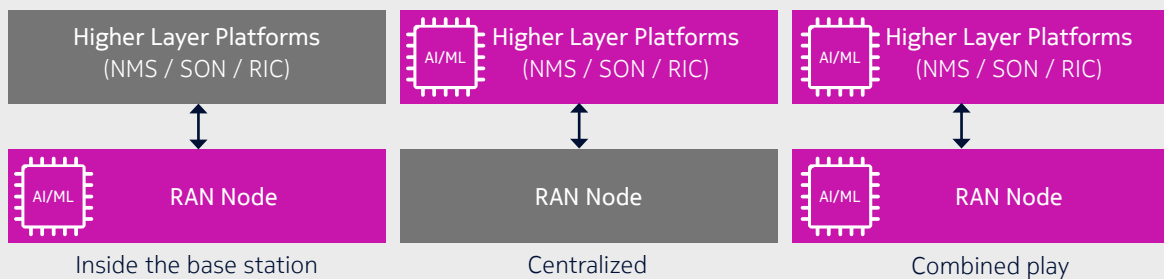


Figure 3. Implementation options for AI/ML in RAN

Centralized implementation of AI/ML

In this option, AI/ML is implemented in self-organizing networks (SON), RAN intelligent controller (RIC) and advanced network management system (NMS), which can handle additional intelligence and AI/ML algorithms with minimal impact on the existing architecture.

For example, in Cloud RAN, the virtualized centralized unit (vCU) is an element that could be centralized, where this makes business sense.

AI/ML inside the base station

In this option, AI/ML is implemented directly in baseband, or in the case of Cloud RAN, in the virtualized distributed unit (vDU). In some cases, AI/ML can be implemented in the Radio Unit.

The challenge is to hit the sweet spot between AI computing performance and the overall total cost of ownership (TCO) of a base station, including power consumption and hardware affordability. Progress in tensor-efficient computing will help. However, the renewal cycles of RAN and processors need to be considered as well.

Combined approach

In a combined approach, centralized AI/ML and AI/ML inside the base station will cooperate, with the centralized system taking the role of the orchestrator.

AI inside the base station

The 3GPP's functional framework for AI-enabled RAN intelligence is captured in 3GPP TR 37.817 [2] and illustrated in Figure 4. This framework encompasses several key functions such as data collection, model inference and training.

The feedback loop is integral, providing essential information for deriving training and inference data.

This setup ensures a dynamic and responsive AI/ML implementation in RAN, continuously improving model performance and network efficacy.

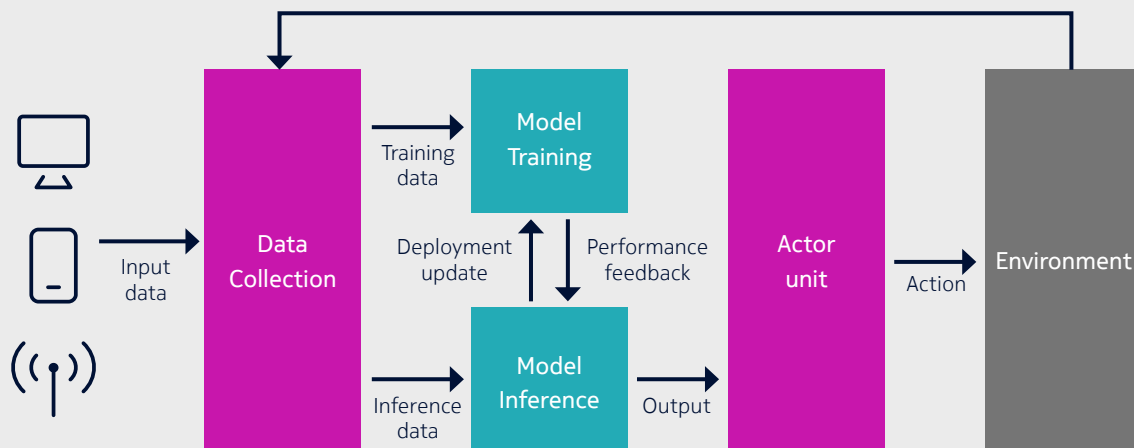


Figure 4. A generic framework for RAN intelligence proposed by 3GPP

Learning from general AI trends, we also need to consider splitting training into basic training, for example, on simulated (or AI-generated) data, and one or repeated fine-tuning activities.

Repeated fine-tuning might be needed to adapt to major changes in traffic patterns or the addition of a frequency layer. If continuous fine-tuning is needed, reinforcement learning might be the better option.

Embedded AI/ML use cases in RAN

In practice, specialized AI/ML units, often referred to as AI/ML accelerators, can be implemented in various ways to optimize the performance-to-cost ratio.

Conventional RAN design is framed by TCO minimization, including reducing energy consumption and space requirements.

Increasing the number and complexity of AI for RAN workloads within the base station over time will result in growing AI computing and memory needs.

In the conventional approach, these resources are added in evolutionary upgrade steps, which ensures the use of the latest technology and avoids over-dimensioning AI computing and related costs at any particular period of time.

The alternative approach is to deliberately accept to have spare AI computing capacity. Using general-purpose GPUs, the spare AI computing power could be used for workloads beyond RAN.

The advantage of the second approach is its readiness for multi-purpose AI workloads in the edge cloud. This is compelling when the location and backhaul bandwidth of the base station site or Cloud RAN data center meet the demand of other AI workloads and use cases beyond RAN.

Considering the end-to-end delay budget for immersive AI-enhanced services, an XR-mass market could become the catalyst for the far edge cloud and the second approach.

Embedding AI/ML within the base station can serve a multitude of use cases, spanning across lower layers of the RAN architecture. These use cases are prevalent in the L1 physical layer, which addresses signal processing and modulation tasks, and the L2, which handles various data link functions such as beam selection, link adaptations and intelligent scheduling.

In addition, embedded RAN use cases extend to the L3 network layer, facilitating various functions such as multi-layer management or power optimization.

While the list of practical AI/ML use cases is extensive, Table 2 highlights some key examples.

Use case	Layers	Description
Multi-layer management and traffic steering	L3	Employs reinforcement learning to manage and optimize data flow across multi-layer networks, aiming to maximize user throughput and improve overall user experience through intelligent carrier assignment decisions. These decisions lead to more efficient resource allocation, improved load balancing, better interference management, and enhanced quality of service (QoS).
LLM-based generative agent of Idle Mode Load Balancing (IMLB)	L3	Driven by customized intent(s), the IMLB generative agent takes its current RAN environment and past experiences as inputs to generate frequency offsets for IMLB as outputs. It leverages pre-trained LLM to deepen its understanding of the network environment and experiences. Finally, it retrieves the most relevant subset of overall considerations to infer proper actions for load balancing.
Generative AI-based carrier aggregation for the optimal radio resource assignment	L3	Frames radio resource assignment as data modeling of signal trajectories, akin to processing “sentences” in a wireless network. By training a transformer model, the goal is to manage the complexities of wireless networks, including diverse devices, service quality requirements, channel conditions, and traffic demands. It aims to optimize radio resource assignment, enhancing both user experience and resource efficiency.
Link adaptation	L2	Leverages reinforcement learning to adapt downlink modulation and coding schemes (MCS) for individual user equipment (UE). It aims to boost overall user throughput by determining the optimal channel quality indicator (CQI).
RRM parameter optimization	L2	Uses reinforcement learning to learn the best action to optimize radio resource management (RRM) parameters considering feedback of UE behavior, RAN configuration and potential interference.

Table 2. Examples of embedded AI/ML use cases

As mentioned, ML can be of particular benefit for tasks where conventional heuristic methods would require enormous computing power for execution in real-time or where AI-enabled prediction complements conventional measurement-based approaches.

Next, we will explain in a little more detail one use case for each of these scenarios

Use case: Intelligent Massive MIMO Scheduler

Massive MIMO (mMIMO) is central to 5G, playing a critical role with beamforming. However, efficiently selecting optimal beams for users is challenging due to hardware and power constraints.

In multi-user settings, beams must be paired spatially to avoid interference, but the combinatorial complexity, such as scheduling 4 out of 32 beams with over 30,000 configurations, makes it computationally expensive.

AI/ML-powered mMIMO schedulers, such as Deep Q-Networks (DQN), solve this by estimating future rewards for actions without evaluating every combination.

This significantly improves performance and approaches theoretical capacity with practical computational complexity.

Use case: AI/ML-based carrier aggregation

Carrier aggregation (CA) tackles bandwidth limitations in cellular networks by combining multiple carriers into a wider aggregated channel.

The process involves a Primary Cell (PCell) that acts as the anchor for user equipment and Secondary Cells (SCell) that provide additional bandwidth.

Traditional carrier aggregation relies on signal strength and network load metrics, employing inter-frequency measurements and mobility trigger events for SCell selection.

In contrast, AI/ML-driven approaches utilize neural network based spectral efficiency predictions to optimize SCell selection, thereby accelerating the setup time.

3GPP Standardization

Many AI/ML algorithms inside the base station do not require standardization. Like most conventional algorithms, they are implementation-specific.

However, considering interworking between AI implementations in different entities, such as the user equipment and the base station, standardization will be mandatory, similar to the standardization of network interfaces.

The 3GPP RAN AI/ML standardization efforts encompass a broad spectrum of activities spanning multiple working groups, from services and system aspects (SA WG) to radio access network (RAN WG).

3GPP Release 17 carried out an initial study on AI-enabled RAN, outlining fundamental principles and a functional architecture. This study also explored diverse use cases such as network energy savings, load balancing and mobility optimization ^[3].

In Release 18, 3GPP furthered its efforts by exploring AI/ML to enhance the 5G New Radio (NR) air interface. The study's scope included creating a general AI/ML framework and investigating specific use cases, such as channel state information (CSI) feedback, beam management and positioning. Table 3 describes these use cases.

3GPP use case	Motivation
CSI feedback	Reduce feedback overhead and improve accuracy.
Beam management	Reduce overhead and latency associated with the beam measurement and reporting.
Positioning	Improve the positioning accuracy, reduce overhead, and improve positioning integrity.

Table 3. Use cases studied in 3GPP Release 18

For all these use cases, AI/ML is utilized independently either on the device side, on the network side, or collaboratively on both sides of the communication system. The process of delivering an AI/ML model from one entity to another over the wireless network is referred to as model transfer.

In Release 18, 3GPP SA-WG1 also started a study to identify use cases and service and performance requirements for model transfers. The study categorized AI/ML operations into three types:

- Operation splitting, which aims to optimize privacy and latency by partitioning AI/ML tasks between UE and network endpoints.

- Model/data distribution, which enables dynamic model delivery to devices.
- Distributed/federated learning, which aggregates training results of local UEs to create a global model.

In Release 19, we can expect further AI/ML use cases to be addressed in the air interface, RAN and system architecture. It will also address outstanding issues identified during Release 18.

In Release 20, 6G will be studied with AI/ML as an integral part of the system.

Towards AI-native 6G and beyond

The term AI-native defines a set of principles for mobile network design where AI/ML is integrated from the outset as a fundamental component, rather than being an add-on to an existing architecture.

In this approach, ML becomes crucial to the system, with many system capabilities relying on ML inference for optimal performance.

Moreover, a foundational, standardized framework for implementing AI/ML is established, encompassing the complete lifecycle of AI/ML components. This integrates AI/ML as an intrinsic and essential part of the system.

6G is envisioned to be fully AI-native with AI/ML enablers at all network layers. Various research efforts are already underway toward developing an AI-native air interface (AI-AI) that enables optimized communication schemes for various radio environments and applications [\[16,17\]](#).

The transition to an AI-native air interface in 6G is anticipated to occur in phases, beginning with the replacement of certain receiver blocks. It will then extend to ML handling multiple processing blocks and culminate in using AI/ML to design parts of the physical and media access control (MAC) layers.

This marks a significant shift in communication system design and standardization.

Figure 5 illustrates the transition from traditional architecture towards the AI-AI architecture.

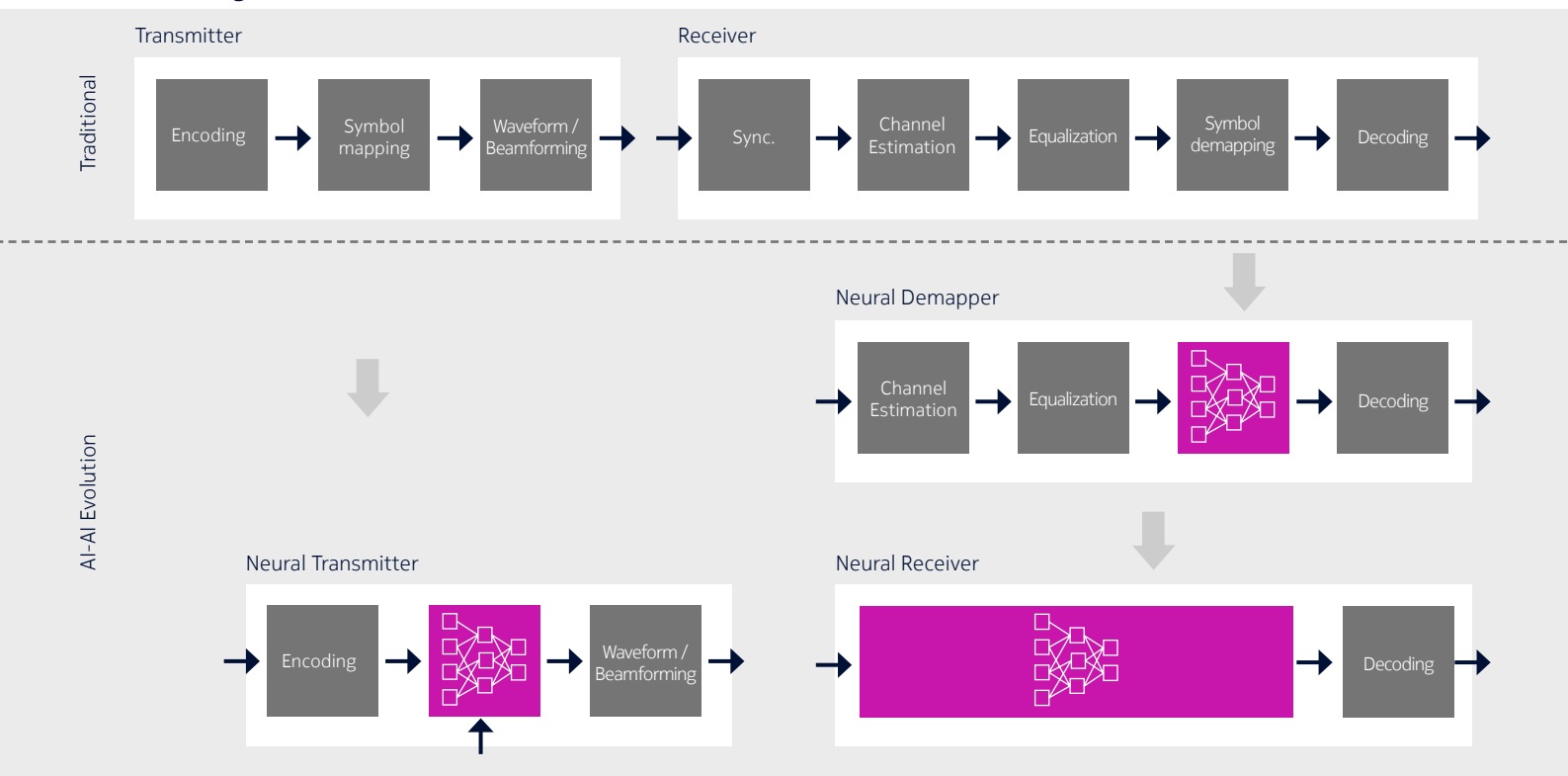


Figure 5. Towards the AI-AI architecture

Research initiatives across these phases have yielded encouraging outcomes. For instance, using a convolutional neural network as a mapper at the receiver end has shown a 2dB bit error rate (BER) performance improvement compared to a non-ML receiver ^[4].

Similarly, data-aided joint de-mapping, channel estimation, and equalization have achieved performance close to perfect channel state information (CSI).

Likewise, the geometric shaping of the constellation at the transmitter, jointly optimized with the neural receiver, enables pilotless transmissions without performance loss. This eliminates the control overhead and need for demodulation reference signals.

Centralized AI for RAN

Centralized AI for RAN can be broadly categorized into two distinct approaches.

The first approach involves offline applications, where AI/ML models augment processes, such as RAN planning and design but without much direct interaction with the RAN nodes.

The second focuses on the online integration of AI/ML models with RAN nodes for activities such as network management, optimization and ran intelligent control.

Figure 6 illustrates some key use cases related to centralized AI for RAN.

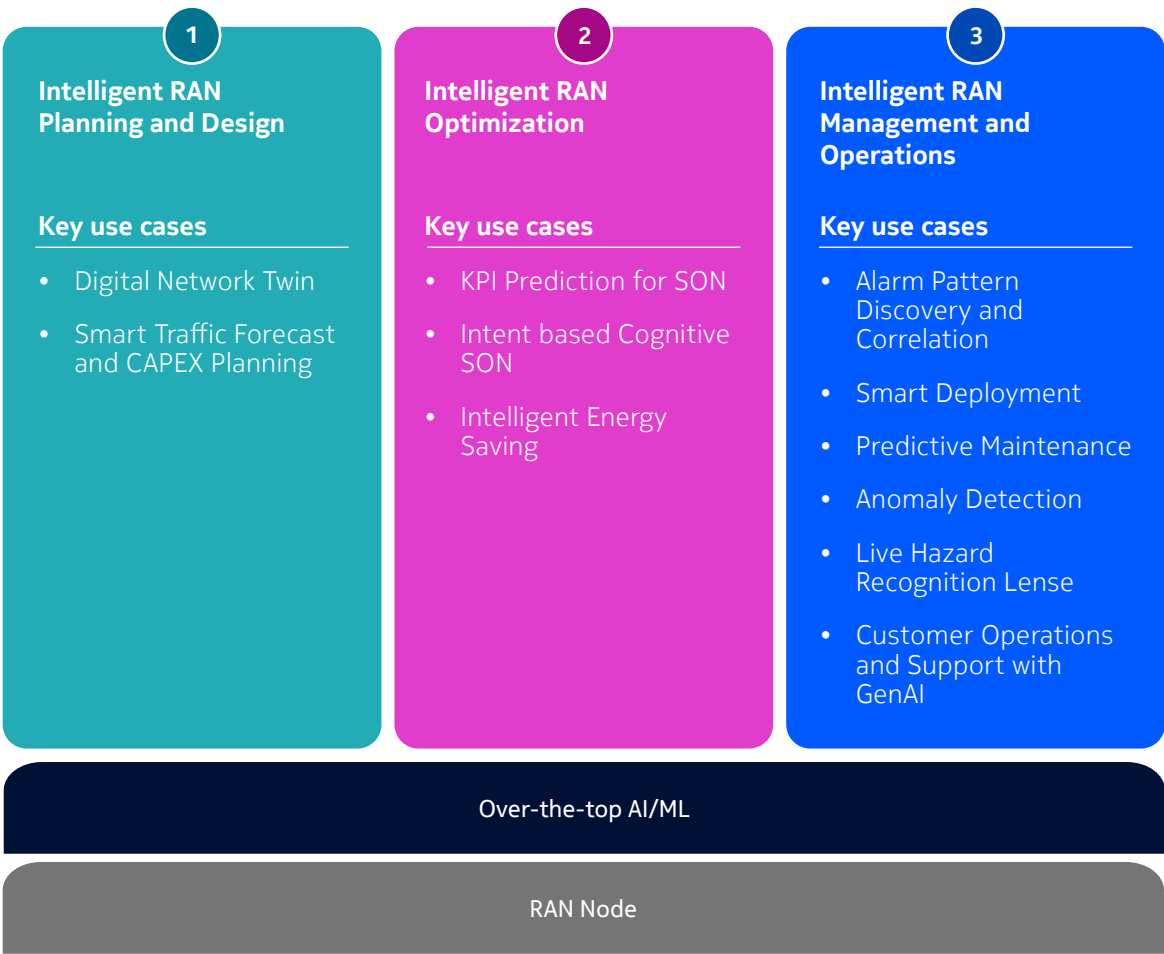


Figure 6. Centralized AI for RAN use cases

Intelligent Radio network planning, design and optimization

AI/ML-driven network techniques significantly enhance the efficiency of radio planning and design tasks, enabling more precise results while cutting down on time and resource costs. These techniques have also demonstrated promising results in RAN optimization.

Machine learning facilitates the automation of repetitive, routine network management activities, thereby improving overall efficiency and performance. These intelligent approaches can also significantly reduce base station energy consumption through advanced energy-saving features.

Table 4 showcases some key use cases in radio planning, design, deployment and optimization, illustrating their benefits and impact.

Domain	Use case	Use case brief description
Intelligent radio network planning and design	Digital Twin Networks (DTN):	Digital Twins are accurate, data-driven virtual replicas of real networks using data from actual networks or simulations. This helps operators model complex scenarios like congestion and link failures, improving network planning and decision-making ^[5] .
	Smart traffic forecast and CAPEX planning	AI/ML-based traffic prediction methods enhance forecasting accuracy by considering multiple factors like regional trends and network events. This allows for more precise capacity planning, reducing costs and improving resource allocation compared to traditional methods.
Network deployments	Live Hazard Detection Lense	AI/ML and AR systems monitor RAN sites in real-time to detect hazards, predict risks and provide immediate safety alerts, reducing accidents and enhancing site safety ^[7] .
Intelligent RAN optimization	KPI prediction for SON	Prediction algorithms forecast network performance KPIs and augment SON actions that improve network performance and efficiency without manual intervention.
	Intent-based Cognitive SON	The system contextualizes the network by categorizing cells based on their static and dynamic attributes. It then translates performance objectives into network configurations and policies thereby autonomously orchestrating SON modules without human intervention. This automation accelerates cell contextualization by 99% compared to manual methods. It also enables a 90% faster detection and resolution of network issues ^[6] .
	Intelligent energy savings	Predictive ML algorithms determine the period of cell switch off by predicting optimal periods of low load in the power-saving group. During peak traffic periods, these carriers are automatically reactivated thereby ensuring uninterrupted service. The AI/ML-powered energy-saving features result in 10-20% RAN energy savings.

Table 4. AI/ML-enabled use cases for radio network planning and design

Intelligent RAN management and operations

AI/ML is transforming network operations, moving operators closer to fully automated zero-touch executions. This shift reduces manual interventions and enables predictive and proactive operations.

Table 5 outlines some use cases in network operations where the application of AI/ML is delivering promising results.

Domain	Use case	Use case brief description
Intelligent RAN Management and Operation	Alarm pattern discovery and correlation	AI/ML techniques like FPGrowth accurately correlate alarm patterns, enabling quick root-cause analysis and reducing alarms by 70%, improving network monitoring efficiency ^[8] .
	Smart deployment and management	AI/ML frameworks, using algorithms like Random Forest, automate management information base (MIB) trap mapping and network element integration, streamlining multi-supplier network management and reducing complexity.
	Predictive maintenance	Predictive algorithms identify network faults with 85% accuracy, while unsupervised learning detects hidden issues, optimizing network performance and preventing service disruptions.
	Anomaly detection	AI/ML uses unsupervised and graph-based algorithms to identify hidden performance issues in networks early, provide recommendations for configuration changes and prevent major disruptions ^[9] .
	Customer operations and support with generative AI	GenAI agents trained in telecom data enhance customer support by offering contextualized communication, advanced troubleshooting and handling complex inquiries efficiently ^[10,11] .

Table 5. AI/ML-enabled use cases for RAN management and operations

TM Forum's road to autonomous networks

TM Forum defines an autonomous network as a system of networks and software platforms that are capable of sensing its environment and adapting its behavior accordingly with little or no human input.

TM Forum's autonomous network framework defines a technical architecture that supports developers in designing and implementing autonomous solutions for network operations and optimization. It defines six levels of automation, with level 0 being fully manual and level 5 fully autonomous ^[12].

Some advanced Cognitive SON solutions can reach TM Forum's level 4 already today ^[6]. Levels 4 and 5 represent highly autonomous operations and cannot be achieved without AI/ML.

AI-RAN

In the previous chapters, this paper has focused on how AI could be used to increase the efficiency and performance of RAN.

One of the challenges identified is the cost-benefit balance of providing tensor-optimized computing in the long term. While at some point every computing resource will be utilized, there might be transition times, during which significant resources could lay idle.

To overcome this investment and lifecycle challenge, an option is to share (lease) spare AI computing capacity in the RAN with other AI workloads on top of the RAN.

By doing so, the business case for investing in and operating additional AI computing resources in the RAN might become attractive even for initially relatively over-dimensioned AI computing.

The AI-RAN Alliance is a new collaborative initiative aimed at transforming RAN with the power of artificial intelligence. It comprises prominent industry stakeholders including operators, suppliers and academic and research institutions.

In contrast to standards bodies, which focus on developing interoperability specifications, the AI-RAN Alliance prioritizes creating practical implementation guidelines and benchmarking AI/ML algorithms for the RAN context.

The alliance has three working groups focusing on three key areas: AI for RAN, AI and RAN integration, and AI on RAN ^[13] as shown in Figure 7.

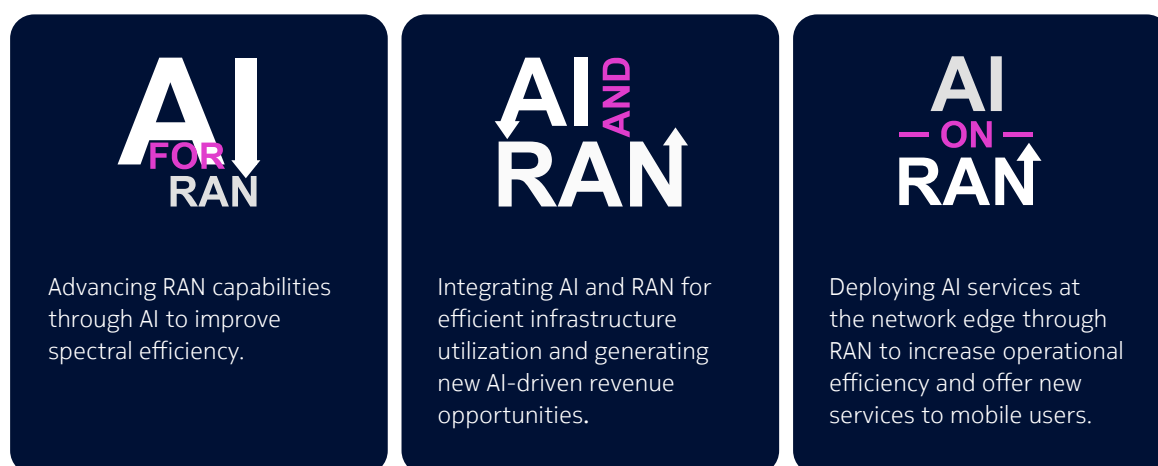


Figure 7. AI-RAN Alliance workgroups

In addition to participating in the work of the AI-RAN Alliance, Nokia is working closely together with leading mobile network operators and technology partners to pioneer the possibilities of AI-RAN ^[14, 15,16].

Responsible and realistic AI for RAN

As AI becomes integral to RAN, the industry faces the challenge of ensuring that these systems work seamlessly across technologies from different suppliers ^[18,19]. This requires not just innovative technical solutions, but also a shift in how we operationalize AI within networks ^[20,21].

Amid this rapid adoption, it's also crucial to embed responsible AI practices to ensure that our advancements are ethical and sustainable ^[22].

Figure 8 lists six fundamental principles that must underpin all future AI research and development. These principles should be adhered to from the inception of any new AI solution through its development, deployment and operational phases ^[23].

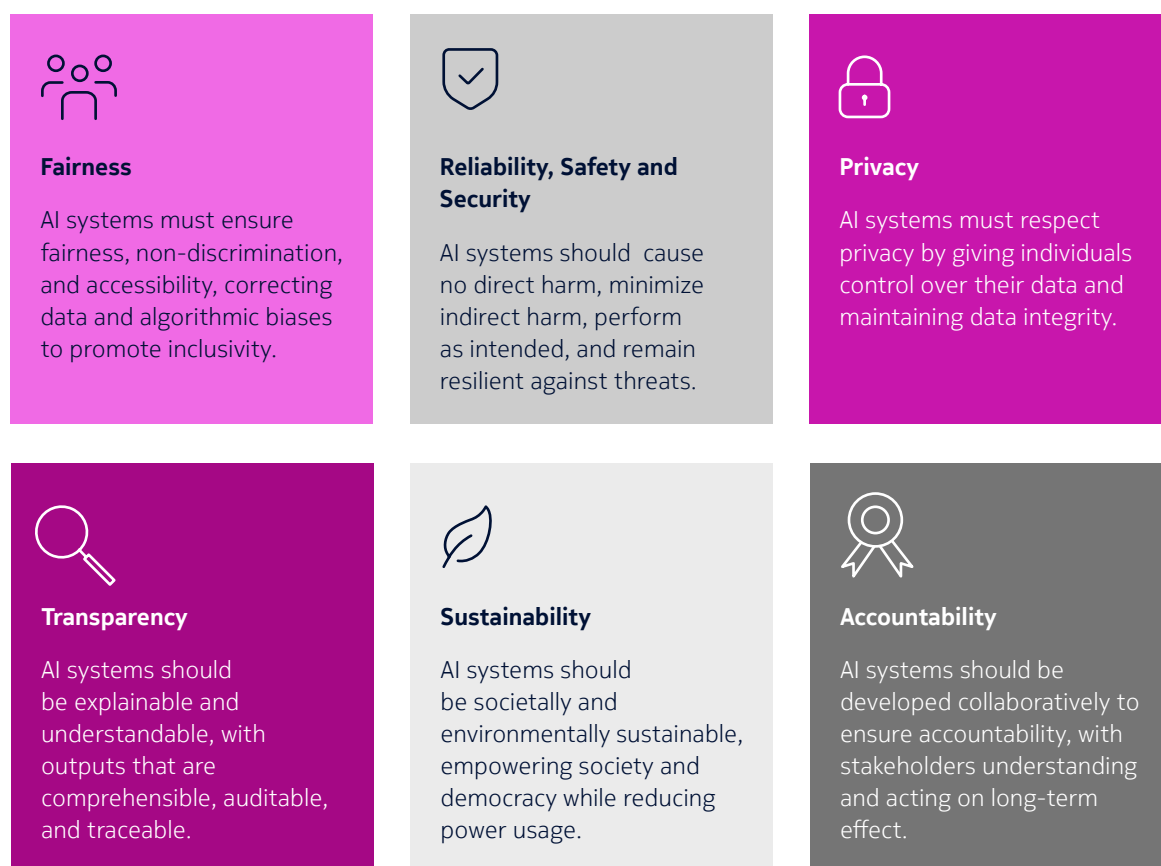


Figure 8. Six tenets of responsible AI

Conclusions

As in many other areas, AI will be an integral part of RAN. Together, the technological progress in AI models and algorithms, tensor-optimized computing and the path to 6G in RAN will unleash efficiencies and enable enhanced services.

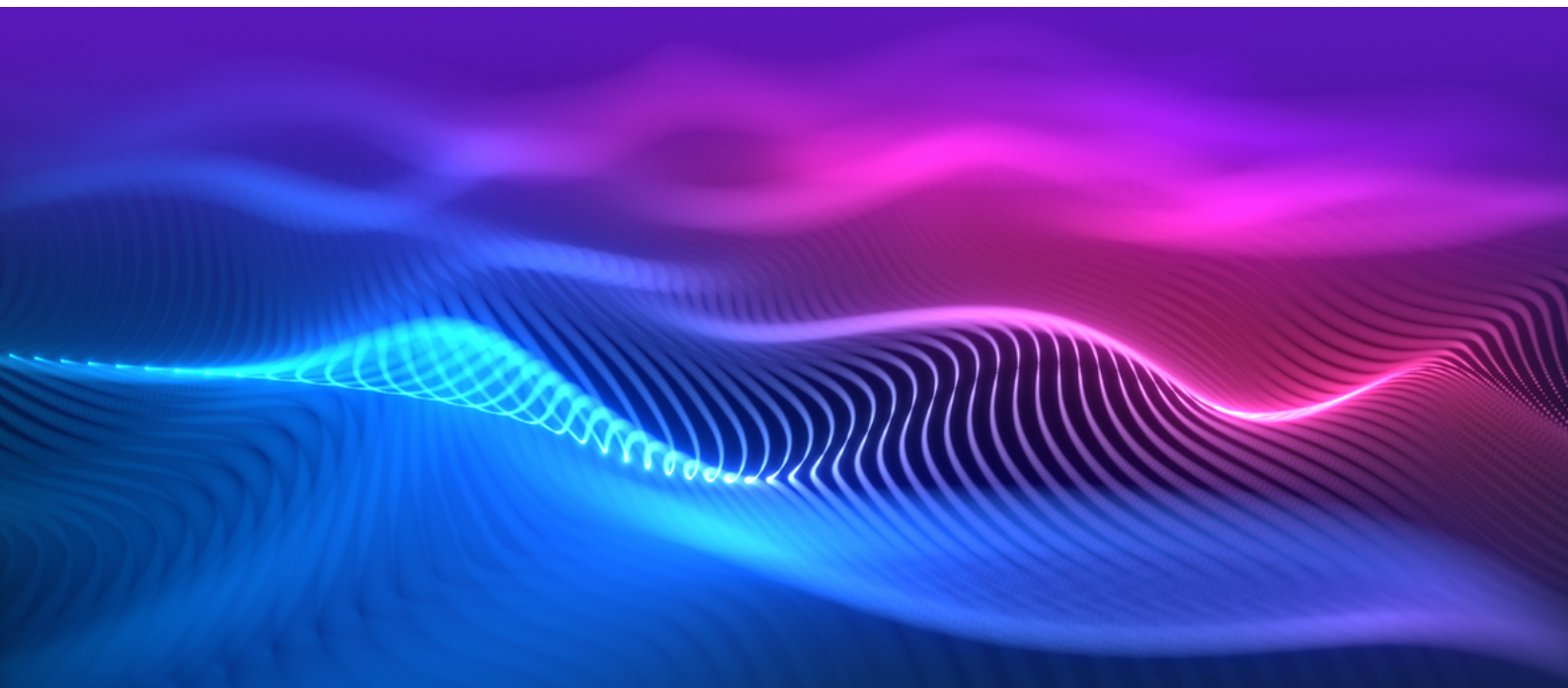
A widening range of AI technologies will be used for RAN, within the base station for real-time tasks and in centralized systems for non-real-time tasks.

Real-time AI has the potential to enhance user experience and air interface performance. AI in centralized systems will further automation and efficiency.

Nokia, together with our research arm Nokia Bell Labs, is committed to a responsible AI mindset and accountability. These are essential when developing solutions that leverage AI.

They are also essential for operators running cellular networks, which represent the critical infrastructure of modern societies.

We are leveraging our extensive RAN domain knowledge when working very closely with industry partners to shape the technology evolution path towards AI-RAN and AI-native 6G.



Glossary

AI	Artificial Intelligence
AI-AI	AI-native Air Interface
AR	Augmented Reality
BER	Bit Error Rate
BOGP	Bayesian-optimized Gaussian Process
CA	Carrier Aggregation
CNN	Convolutional Neural Networks
COTS	Commercial-Off-The-Shelf
CPU	Central Processing Unit
CSI	Channel State Information
DQN	Deep Q-Network
DTC	Decision Tree Classifier
FWA	Fixed Wireless Access
GPU	Graphics Processing Unit
IMLB	Idle Mode Load Balancing
IoT	Internet of Things
KNN	K-Nearest Neighbor
LLM	Large Language Model
MAC	Media Access Control
MAPPO	Multi-Agent Proximal Policy Optimization
MARL	Multi-agent Reinforcement Learning
MIB	Management Information Base
ML	Machine Learning
NLP	Natural Language Processing
PCell	Primary Cell
PPO	Proximal Policy Optimization
QOS	Quality of Service
RA	Random Access
OAM	Operations, Administration and Maintenance
NMS	Network Management System
RAN	Radio Access Networks
RIC	RAN Intelligent Controller
RRM	Radio Resource Management
RNN	Recurrent Neural Networks
SAC	Soft Actor Critic
SCell	Secondary Cell
SLM	Small Language Model
SoC	System-on-Chip
SON	Self-Organizing Network
TCO	Total Cost of Ownership
TTI	Transmission Time Interval
UE	User Equipment
vCU	Virtualized Centralized Unit
vDU	Virtualized Distributed Unit
WG	Working Group
XR	Extended Reality

References

1. J. Hoydis, F. A. Aoudia, "Recent Progress in End-to-End Learning for the Physical Layer", Nokia-IHES Workshop, 2019, <https://www.youtube.com/watch?v=EPLJzsxReH4>
2. 3GPP, "TR 37.817 - Study on enhancement for data collection for NR and ENDC", 2022 <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3817>
3. Xingqin Lin, "Artificial Intelligence in 3GPP 5G-Advanced: A Survey", IEEE CTN, Sep 2023, <https://www.comsoc.org/publications/ctn/artificial-intelligence-3gpp-5g-advanced-survey>
4. J. Hoydis, F. A. Aoudia, A. Valcarce and H. Viswanathan, "Toward a 6G AI-Native Air Interface," in IEEE Communications Magazine, vol. 59, no. 5, pp. 76-81, May 2021, <https://ieeexplore.ieee.org/document/9446676>
5. Nokia "Achieve optimal network performance with AI and digital twin based services", Feb 2024, <https://www.nokia.com/blog/achieve-optimal-network-performance-with-ai-and-digital-twin-based-services/>
6. Nokia, "MantaRay Cognitive SON: Powering AI-driven Autonomous RAN Operations", July 2024, <https://onestore.nokia.com/asset/213071>
7. Nokia, "Hazard Detection Lens", July 2024, <https://onestore.nokia.com/asset/214089>
8. Nokia, "Radio Access Network energy efficiency", March 2024, <https://onestore.nokia.com/asset/213911>
9. Nokia, "Intelligent RAN Operations Improve radio network quality with Machine Learning", August 2023, <https://onestore.nokia.com/asset/212331>
10. Nokia, "Generative AI implications for telco operations", May 2024, <https://onestore.nokia.com/asset/213859>
11. Nokia, "Nokia Digital Assistant," August 2024, <https://digital-assistant.nokia.com/>
12. TMForum, "Autonomous Networks: Empowering digital transformation – evolving from Level 2/3 towards Level 4 (IG1326)", Sep 2023, <https://www.tmforum.org/resources/how-to-guide/ig1326-autonomous-networks-empowering-digital-transformation-evolving-from-level-2-3-towards-level-4/>
13. AI-RAN Alliance, "AI-RAN Alliance Working Group", Aug 2024, <https://ai-ran.org/working-groups/>
14. T-Mobile, Press Release: "T-Mobile Announces Technology Partnership with NVIDIA, Ericsson and Nokia to Advance the Future of Mobile Networking with AI at the Center", Sept 2024, <https://www.t-mobile.com/news/business/t-mobile-launches-ai-ran-innovation-center-with-nvidia>
15. Nokia, Press Release: "Nokia and partners are driving the fusion of AI, Cloud and RAN", Sept 2024, <https://www.nokia.com/blog/nokia-and-partners-are-driving-the-fusion-of-ai-cloud-and-ran/>
16. Softbank, Press Release: "SoftBank Corp. and Nokia partner to research AI-RAN and 6G Network Technologies", Sept 2024, https://www.softbank.jp/en/corp/news/press/sbkk/2024/20240911_01/
17. Nokia, Press Release: "Nokia, SKT, NTT and DOCOMO team up to implement AI in the 6G air interface", Feb 2024, <https://www.nokia.com/about-us/news/releases/2024/02/22/nokia-skt-ntt-and-docomo-team-up-to-implement-ai-in-the-6g-air-interface/>
18. Oana-Elena Barbu, "AI shapes the radio architecture of future cellular devices", Jul 2024, <https://www.nokia.com/blog/ai-shapes-the-radio-architecture-of-future-cellular-devices/>
19. Nokia, Press Release: "Nokia and Qualcomm jointly research AI-interoperability technology that boosts wireless capacity and performance", Feb 2024, <https://www.nokia.com/about-us/news/releases/2024/02/20/nokia-and-qualcomm-jointly-research-ai-interoperability-technology-that-boosts-wireless-capacity-and-performance/>
20. Analysys Mason, "MLOps: streamlining machine learning for CSPs", June 2023, <https://www.analysismason.com/research/content/short-reports/mlops-short-report-rma14/>
21. Nokia, "Transforming CSPs with AIOps: Navigating silos and ensuring business value", Jan 2024, <https://onestore.nokia.com/asset/213719>
22. Anne Lee, "Responsible AI for Telecom", August 2023, <https://onestore.nokia.com/asset/212898>
23. Nokia Bell Labs, "What is Responsible AI", <https://www.bell-labs.com/research-innovation/ai-software-systems/responsible-ai/#What-is-Responsible-AI>

Nokia OYJ
Karakaari 7
02610 Espoo
Finland

Tel. +358 (0) 10 44 88 000

CID: 214372

nokia.com

NOKIA

At Nokia, we create technology that helps the world act together.

As a B2B technology innovation leader, we are pioneering networks that sense, think and act by leveraging our work across mobile, fixed and cloud networks. In addition, we create value with intellectual property and long-term research, led by the award-winning Nokia Bell Labs.

With truly open architectures that seamlessly integrate into any ecosystem, our high-performance networks create new opportunities for monetization and scale. Service providers, enterprises and partners worldwide trust Nokia to deliver secure, reliable and sustainable networks today – and work with us to create the digital services and applications of the future.

© 2024 Nokia