Is AI taking over the RAN?

A gradual path to AI-driven efficiency and automation

SENZA FILI

NETSCOUT

Go back to the top

Roadmap

Report

Introduction
Making AI invisible
Doing the same, differently
The path to AI RAN
Layering AI from the top
The multiple flavors of AI
Does it all add up? Monetization and ROI

Dig deeper

The multiple roles of AI in the RAN
Lessons from Open RAN
AI amplification: The price of a dynamic RAN
Going AI native
Do we need and can we afford AI?
Does AI RAN need GPUs?
What makes a RAN an AI RAN?

For the impatient

Takeaways

Conversations

NETSCOUT: AI in the RAN: A data-first path to full automation. With Karsten Gaenger, Principal Product Line Manager for the RAN at NETSCOUT

IMDEA Networks: From AI in the RAN to AI for the RAN: Making innovation safe. With Nirlay Kundu, Head of Technology Standards at IMDEA Networks

Monica Paolini
Principal
Senza Fili

Karsten Gaenger
Principal Product Line Manager for the RAN
NETSCOUT

Nirlay Kundu
Head of Technology Standards
IMDEA Networks

Watch the videos of the conversations

AI is everywhere and nowhere in the RAN. It is pervasive and yet elusive. We know that eventually AI will permeate the entire RAN, but we are far from this. Wide commercial deployments are still a rarity, with AI technology selectively used for specific tasks. AI can bring huge cost savings, and higher performance and power efficiency mostly through automation, but it is not yet clear what the integration path will be, and everybody in the industry (not just operators) struggles to find the right strategy. How aggressive should the transition to AI RAN be? How should we treat the legacy infrastructure? How should we change operations, organizational structure and culture to integrate AI in the RAN and, more generally, in mobile networks? How disruptive will the introduction of AI be? What's the best way to prepare for AI and its attending challenges and risks? How should we change our approach to security, sovereignty and privacy? Will AI deliver the boost in revenues operators have been unable to get with 4G and 5G, and are not too confident that 6G will deliver?

AI will have a transformative impact on the RAN, and yet the function of the RAN will not change. The challenges will be the same that we have today, but they will be amplified. Security, for example, has always been a major concern, but AI increases the risk of security breaches. On a positive side, it also protects networks more effectively from security threats.

AI will increase complexity, but it also enables more streamlined operations and higher automation. Operators will be able to launch new services and benefit from greater efficiency, but it is unclear how fast the transition can (or should) be.

There are still many open questions that have to be addressed to chart a successful path to AI-native, fully automated RAN. There is no prescriptive roadmap that works for everyone. Each ecosystem player will have to find its own answers based on its specific context and requirements.

As always, I followed announcements and attended analysts events in preparation for this report, but the main source for what I present here is countless informal conversations I had with those actively working on AI in the RAN—from startups to tier-one vendors, operators worldwide, researchers and experts—which were informed by my graduate research in and teaching AI and cognitive science, at a time when AI was a mostly academic endeavor and which gave me some practical appreciation of the complexity and power of AI. The result is a report that presents a somewhat tangential perspective on AI RAN, centered on how to develop an AI RAN strategy, how to identify and address the unique challenges and opportunities, and how to get to a ROI-positive AI-RAN.

The multiple roles of AI in the RAN

Let's be honest: it can be confusing to keep control over the multiple incarnations of AI in the RAN, or elsewhere. As we are still in the early phase of AI RAN, when we still have to fully understand its costs and benefits and how we should deploy it, even the definition of these concepts is often a matter of debate.

Yet they all point to the multiple dimensions in which AI can get into the RAN, the choices and tradeoffs that are available, and the type and functionality of AI models that are best suited for specific goals.

AI-native	AI-first	AI-powered	AI-driven
AI-for-RAN	AI-on-RAN	RAN-with-AI	AI-on-RAN
AI-managed	AIOps	AI-ready	On-site AI
Distributed AI	Edge AI	Agentic AI	Gen AI

Making AI invisible

New technologies generate hype and, in the case of AI, it is tempting to think that the hype is justified, and more justified than the hype surrounding 5G. Yet, hype does not help to understand the potential of the technology or guide us in deploying it successfully. This is the case in AI, where the mix of fear and awe that it inspires adds an emotional complexity to the assessment of its benefits and risks.

As a result, AI captures most of the limelight, and the AI RAN discussion is about AI more than it is about the RAN. We should not lose sight of the fact that we want to use AI to improve the RAN, rather than finding a way to use AI in the RAN. AI is a powerful and essential tool to enable RAN transformation, and we need to maximize its benefits. Instead, the dominant narrative appears to be predicated on a FOMO and centered on how to maximize the role of AI and speed up its adoption.

A more nuanced and selective perspective is slowly emerging among some vendors and operators, and it will be crucial to the success of AI in the RAN and, more importantly, to the evolution towards an autonomous, more efficient RAN.

AI is not always necessary or worth the cost or effort. We should not use AI for tasks for which we have an effective deterministic (e.g., a well-tested if-then set of rules) model. It would add unnecessary cost and complexity, and no benefit. For other tasks, AI may not have reached sufficient maturity to add any advantage and instead may increase risk.

At the same time, introducing AI is typically not a Boolean decision, but a gradual integration into a complex RAN environment with multiple legacy elements and an operational and organizational structure that may not be ready to adapt to full-fledged AI models.

Reducing the focus on AI and redirecting the attention to identifying the best evolution path for the RAN and, subsequently, the role that AI can play in that, will help ensure a smooth and successful transition to AI RAN.

This is not to deny that AI will eventually permeate all elements of the RAN and the rest of the network, and it will have a massive impact on how we deploy and run networks. On the contrary, its pervasiveness means that AI will not be a differentiating factor. Every vendor and every operator will continue to expand the use of AI. What will set them apart is the path they choose and the ability to successfully integrate AI into their products and networks.

Lessons from Open RAN

A rigid AI-centric view of the evolution of the RAN may also have a polarizing impact that limits innovation and effective deployments.

Open RAN offers a valuable lesson. Operators have created Open RAN to encourage multi-vendor deployments, and continue to support openness in the RAN. Understandably, some vendors have resisted the transition to a more open RAN, as this may reduce their end-to-end control over the RAN and revenues.

Despite this, there is an undeniable trend towards more openness in the RAN, and almost all vendors have come to accept it and develop products that work in multi-vendor environments, even though they prefer to control the end-to-end network as much as they can, and in most cases they still do.

Yet market estimates and forecasts for Open RAN are invariably disappointing, with commercial deployments mostly confined to greenfield operators or to small portions of the network of brownfield operators.

What many perceive as a lack of success or failure of Open RAN is the result of a too rigid and demanding definition of what Open RAN is, which does not fit into what operators need and want to deploy.

Operators have been gradually increasing openness in their networks, driven by their needs, financial and organizational resources, and commercial availability. In most cases, these changes do not neatly fit into the Open RAN definition, but they are gradually, organically and sustainably driving RAN evolution.

We should avoid to replicate this approach with AI RAN. If we force feed AI into the RAN and define a path for AI adoption that is not sufficiently flexible to meet operators' requirements and resources, we may end up with disappointment over what we may perceive a limited adoption of AI RAN. AI is a technology with a huge transformative potential, but it could backfire if operators adopt it to meet industry pressures rather than their own RAN transformation strategy.

Doing the same, differently

How is the AI RAN going to be different from today's RAN? The functionality will remain the same, with some added features that today are too computationally intensive to implement or are not yet available (e.g., those that will be available with 6G). Yet, the most promising potential of AI comes not from doing the same things more efficiently, but from doing them in fundamentally different ways. This is an aspect that is often glossed over, but that carries new challenges and risks that make the transition to AI a very delicate process that each operator has to manage carefully, avoiding herd thinking and finding their own path.

AI models harness a compute power that is unmatched by the human workforce, and, by itself, improves network efficiency. Humans have strict short-term memory limitations (5-8 items at a time) and that sharply limits their ability to optimize systems with a large number of interacting parameters. They are also unable to operate continuously in real time, with no sleep or time off. Hiring more staff would not be sufficient, nor financially sustainable, to remedy these limitations. Not only does AI improve the efficiency of individual RAN elements, but it can optimize performance across network elements and sites continuously, based on real-time traffic load and resource availability. When interacting with humans, AI also provides limitless patience, availability, and willingness to adapt to the needs and capabilities of its interlocutors.

Impressive as this is, this is a brute-force extension of today's RAN: more of the same. The more disruptive innovation from AI comes from performing the same tasks differently from humans. Our experience and expertise make up for our limitations in number crunching and keep the networks running. Telecom AI models do not try to duplicate our learning processes—or indeed our "intelligence". At present, it would be more difficult to do so (we still have a limited grasp of how human intelligence works) and not desirable. Having an AI model with the same short-term memory limitations as humans would not take us to AI RAN that is much better than our existing RAN.

Instead, AI learning and inferencing use techniques developed for models that can digest massive amounts of data and produce different outcomes from humans. They can show us alternative approaches and solutions that we were not able to see before, and that may urge us to reevaluate how to deploy and run networks.

They may also be flawed, and have undesirable performance, reliability and security impacts. To protect the RAN, we need to adopt operational and organizational models that allow us to manage appropriately the behavior of AI models.

AI amplification: The price of a dynamic RAN

AI models have an inexhaustible stamina in absorbing data, finding problems and solutions, devising new tricks to increase efficiency, protecting networks from security threats, and assisting customers and staff. The increased activity creates an amplification problem.

AI and automation will make networks more dynamic and capable of adapting to requirements and resources in real time. But to do so, AI models will trigger many more network adjustments than we are currently used to, and this may increase the number of erroneous interventions even though the rate of error is lower than for human-driven ones.

Without AI-driven automation, operators do not have the capability to aggressively optimize the RAN and instead use overprovisioning to ensure reliability and good performance.

Without the capillary network monitoring that AI can add, operators may not be aware of issues that may not be crippling but nevertheless reduce the performance.

Optimizing the RAN in real time is one of the biggest benefits that AI brings, but it comes with increased complexity and, at least initially, as the models and humans supervising them are still learning, with a higher potential for suboptimal or wrong decisions.

It is imperative to proceed with caution to fine tune the AI models and keep the humans in the loop until they reach sufficient trust in the models.

This is a two-way process. Humans have to check the output of the models, but the models have to become transparent and let the humans understand how they operate and how they get to their output. In turn, this allows humans to take corrective action and, in the process, to gain a better insight and higher trust in the models. AIOps is gaining traction to address these issues by reframing how operations work in an AI-driven RAN.

The path to AI RAN

AI will make the RAN more efficient and less reliant on direct human intervention, improve the economics and user experience, mostly by enabling automation. At the same time, an automated RAN will be more complex and vulnerable, adding new security and reliability requirements. As we noted, it achieves this by injecting more compute capabilities and by fundamentally changing RAN operations.

The introduction of AI in the RAN and the final transition to a AI-native, fully automated RAN have to carefully managed to protect the RAN from the vulnerabilities and amplification issues AI introduces and to ensure that AI delivers the desired outcomes in cost-effective ways. It is not a one-time upgrade; it is a protracted process that requires investment and new hardware and software, but also an organizational and cultural reorganization, with three partly overlapping components:

- **Getting ready**: AI uses historical, current and real-time network data and documentation that is available but has not been extensively used in the past. We need to ensure that this data is reliable, accurate and accessible to AI models. We also have to feed the data at the right granularity level or give the models the ability to zoom to the appropriate level for a given task. Too much detail is likely to limit the ability to generalize. Insufficient data may prevent the model from learning or inferencing correctly.

- **Gradual integration**: AI has to be smoothly integrated into the existing RAN, and initially the adoption has to target specific and limited functions where AI can bring a provable short-term benefit and can be added to the legacy infrastructure. AI deployment can then expand as the operator becomes comfortable with an increasing role of AI and higher automation. This is a phase of intensive exploration and learning:

    - Selection of AI models and commercially available that are best suited the operator's needs
    - Integration of AI tools within the RAN and with already deployed AI tools
    - Training, testing and refining AI models
    - Organizational change to adapt to AI-driven automation

- **Pulling it all together**: The initial piecemeal addition of AI functionality throughout the RAN may lead to unexpected conflicts or undesirable interactions. Suggested actions from AI agents with different remits may create conflicts that orchestration

Going AI native

The endgame is to get to an AI-native RAN, where AI permeates all functions natively, starting from Layer 1. This requires a long-term, massive transformation of the RAN, and we may need to wait for 6G to get to that stage for commercial availability. Assuming a deployment timeline comparable to 5G, we should expect several years before reaching large commercial deployments.

We cannot wait for 6G and AI-native RAN to get AI in the RAN. Yet, the AI-native RAN is increasingly presented as an urgent strategic priority, rather than a long-term goal. This risks shifting the efforts of AI adoption in the RAN away from what is available and cost-effective today.

A more realistic approach that yields measurable benefits in the short term and can be integrated in the existing RAN is to layer AI on top of the existing infrastructure, initially focusing on low-hanging-fruit use cases such as power efficiency, management of radio resources, or fault management, where AI can be bolted onto the legacy RAN and no rip-and-replace is required. This keeps the cost of the addition of AI low and reduces risks and complexity.

The transition to AI-native RAN will be gradual and driven more by operator's life-cycle requirements than commercial availability. Similarly to virtualization, operators will shift to AI-native elements gradually, depending on their need to replace existing infrastructure or install new one, and their readiness to move to an AI-ready solution.

has to resolve. Maximizing the efficiency of the overall RAN or network or meeting specific performance targets requires more than the optimization of individual elements or functions. For instance, minimizing power consumption may have a negative impact on user experience, and tradeoffs are necessary to resolve issues like these. Similarly, optimizing the performance at one cell site may reduce the overall performance in the surrounding area, and multi-site optimization can lead to greater RAN efficiency. These are not not new, AI-specific challenges, but they become more consequential in an automated RAN that is inherently more dynamic, and they have to be addressed to unleash the full potential of AI.

Layering AI from the top

AI may be everywhere in the RAN, as I stated at the beginning, but true AI-native RAN is more than a decade away. How do we reconcile that? There are two dimensions to consider.

From bolt-on to native AI. For each RAN element, AI can be bolted onto what is already there. This requires limited investment and effort, carries lower risk, and enables operators to explore how AI can improve over a RAN mostly run by humans. (AI and ML models have been used for quite some time already, so AI is not new to vendors or operators, but the extent of its potential has expanded.) Many vendor solutions promise to do so, on the correct assumption that most operators are reluctant to replace their infrastructure to add AI capabilities. The downside is that this approach limits AI capabilities and benefits, and is likely to be less energy efficient because it is not optimized for AI.

When operators are ready for a lifecycle refresh or network extension, they may switch to a solution with a tighter, more efficient integration of AI, and eventually to an AI-native solution. At this point, operators will be ready to integrate the solution at a deeper level, with the additional investment and operational change required.

From AIOps and energy management to PHY/MAC AI-optimized silicon. Use cases such as AIOps and energy management are among the first ones to be deployed because they are relatively easy to integrate in the RAN today (i.e., no need to wait for 6G) can unleash the compute power of AI on massive data sets that are available, and address pressing operators needs to cut costs and increase network efficiency in the short term. At the same time, these use cases give staff exposure to how AI is changing not only the RAN, but also their jobs and the organizational structure.

Do we need and can we afford AI?

The AI train has left the station. I do not know of any RAN vendor that is staying out of AI. The depth of AI integration and the benefits of AI vary across solutions and vendors, but AI has an increasing role that goes beyond the marketing pitch that was most prominent one in past years. Equally, operators are all committed to deploying AI, even though their readiness and speed of adoption vary greatly. Surveys capture the mood around the adoption of AI, but give little insight into how AI is being deployed and the challenges and benefits it brings.

But do we need AI, or are we following an unavoidable hype trend? AI is a hugely transformative technology and it would be foolish to ignore it. Most importantly, it is not an option; it is a necessity to manage the complexity of today's and future networks and to optimize them to get the economics right and extract all the benefits of new telecom technologies. Deploying the latest beam or power management solutions without AI greatly limits their efficiency and with it the soundness of the capex investment. It is inconceivable to get to L5 fully autonomous networks without a major AI contribution, but we need AI well before we get there. We need it today to progress to L3 or L4. The open question is not whether AI is for real, but how to best integrate it.

A frequent objection to AI is that it is a shiny object that operators cannot afford in the current challenging environment. This is the case for a GPU-dominated AI RAN, but that's not the plan for operators in the short-to-mid term.

Depending on how AI is introduced, the capex investment may be limited, but AI requires an internal operational effort for integration that may be demanding of financial and human resources. More importantly, AI makes sense to deploy only in cases with demonstrable benefits, and that clears the ROI hurdle.

Gradually operators will move to other use cases, such as beam, spectrum and traffic management, or sensing, and eventually may decide to go native all the way with neural processing units with AI-optimized chipsets that will replace today's ASICS, DPSs and FPGAs. It is too early to know exactly how expensive and effective these AI-native solutions will be and how long before they will become available and affordable for mass deployment, but it is difficult to imagine a 6G air interface untouched by AI.

The multiple flavors of AI

AI fundamentally differs from traditional deterministic models because it does not need instructions or established expertise. It probabilistically extracts patterns from massive data sets to generate the desired output. Because it does not follow the if-then algorithms of deterministic models, it is much more powerful in finding novel or more accurate answers, but it is also more prone to mistakes. This is an acceptable tradeoff, but one we have to get used to, by buildign guardrails to reduce the likelihood of mistakes and biases, and learning to spot them.

Nevertheless, even the best AI model adds complexity, cost and risks over a deterministic model, and we should keep AI to perform tasks for which we do not have effective deterministic algorithms. This is the case for most tasks in the RAN and in many cases, an AI model may show us that the deterministic model we have used in the past could be improved by moving to AI.

The same data-driven, probabilistic approach of AI, however, works over a large variety of model architectures, with different capabilities and suited for different tasks. We need this diversity and flexibility to harness AI to help us across use cases.

Initially, LLMs were often presented as the next-generation technology, replacing traditional AI/ML models. Today, many treat agentic AI as the new frontier that will solve all the problems where AI/ML and LLMs models failed.

AI and ML models that use historical and real-time network data to optimize and automate the network will retain a central role in the AI RAN. They are operationally efficient but not transparent, and often leave humans wondering how they got to their results and if they can trust them.

This is where genAI and telco small language models (SLMs) come in. Unlike the hallucination-prone LLMs we use for our less mission-critical daily work and personal

Does AI RAN need GPUs?

A contentious issue is whether AI RAN needs GPUs. The initial vision of an AI RAN that required AI-optimized GPUs at the edge to support the higher compute requirements of AI inferencing for the air interface has been challenged on multiple fronts.

In a layered, gradual integration of AI in the RAN, the initial compute needs are contained and will subsequently grow in line with the expansion of the role of AI in the RAN. GPUs are not a requirement at this stage. Over the next few years, when only limited deployments of AI should be expected, GPU adoption will not be necessary, and this is good news for the ROI, because GPUs are an expensive addition to the RAN.

As AI-native elements are added to the RAN, GPUs may be required in some cell sites. This may change when we get a true AI-native RAN, with AI-driven PHY/MAC, but vendors have divergent views on this. Because we still do not know the specs and requirements of AI-native RAN and what type of hardware solutions will be available at deployment time, it is not possible to adjudicate the matter yet.

One further issue is that the high cost of the GPUs or alternative solutions may slow down or delay the adoption of AI-native RAN, in the same way that M-MIMO adoption was initially limited by the high hardware costs. AI-native RAN may offer better performance, but the cost may be too high to justify the investment.

At present, planning for GPUs in early AI RAN deployments is only going to make the business model unsustainable and slow down the integration of AI in the existing RAN.

queries, telco SLMs are more accurate and reliable and have lower compute and power requirements because they are sharply limited to network data that should have been verified and cleansed. These language models create the interfaces humans need to gain access to information about the network, to get visibility into AI/ML models, and, ultimately, to get the support they need to complete their tasks. This is especially valuable for junior staff who may not have the experience and expertise needed to execute a task.

Agentic AI is the latest flavor of AI, and it is likely to be followed by new ones. It is difficult to escape the anthropomorphizing urge to see agents as human replicas, arguing with each other and complaining about their orchestration agent boss. They have a well-defined scope of action and help humans or other AI models in specific tasks. Language models give us information; agents help us do things and find solutions to specific problems. They are highly complementary and will often be used in a collaborative environment.

AIOps is an example of how different AI models come together to transform the way we operate the RAN and gradually move towards autonomous networks. AI/ML models are the foundation for applications such as energy optimization, fault and anomaly detection, or traffic and resource management. Language models combined with RAG can assist humans in gathering and managing the output of these models. Agents can then build on this to take further action and execute tasks. On top of this, AIOps can use digital twins to explore what-if scenarios that project and assess the impact of changes in the existing network in a virtual environment, and to avoid costly mistakes.

Does it all add up? Monetization and ROI

At least half of the hype on AI RAN comes from the monetization opportunity it is expected to unleash—and that sounds ominously similar to the one presented for 5G and, before that, for 4G that failed to materialize.

ROI. The case for a positive ROI for AI in the RAN is easy to make, as long as operators choose wisely how, where, when and for what to integrate AI. AI can increase network efficiency and thus reduce costs, and this can justify the investment. A gradual adoption of AI is also a strategic move within a longer-term RAN transformation to autonomous networks that requires deep and ongoing organizational change, even though it may reduce the initial ROI attractiveness because of the upfront costs.

What makes a RAN an AI RAN?

There are many ways to answer this question, but it may be best to not try to answer it at all.

In the current market environment, where almost every vendor product has AI built in, it is difficult to decide what counts as AI RAN. A solution may be AI-ready, but the operator can decide whether to use the AI functionality and to what extent to do so. Does it mean that any RAN that includes an AI-ready element is an AI-RAN? Or do we require more than AI readiness? Is it sufficient if AI is used only for one use case? This creates a useless definition rabbit hole that does not benefit anybody, with the possible exception of some vendors.

Similarly, defining AI RAN as being AI native, pushes AI RAN into a long-term adoption timeline, and ignores the fact that AI is rapidly taking hold in the RAN along different depths and dimensions, through a layering approach.

A strict definition of AI RAN may introduce similar risks to the Open RAN overly strict definition. It creates polarization in the industry and adversarial narratives (i.e., AI RAN will never work versus AI RAN is everywhere) that fail to capture the transformative, but gradual contribution of the technology.

My suggestion is to avoid spending time defining what AI RAN is, and to be careful in assessing the market size and revenue opportunity for AI RAN. Dell'Oro forecast does exactly this. It predicts that AI RAN will account for about a third of RAN capex (and that's not in addition to the overall capex forecast) by 2029, but it uses a flexible definition of AI RAN that includes bolt-on, non-AI-native AI solutions. Over the next decade, AI RAN may capture most of the RAN market by this definition and, at that point, we may not need a separate forecast for AI RAN.

Go back to the top

A positive ROI is the primary driver for AI RAN, and one that the operator can control by choosing the appropriate adoption strategy.

In addition, AI adoption among users and enterprises enables and drives new services or applications, or enriches existing ones, creating more demanding performance requirements, e.g., for capacity, latency, reliability and security. AI will help operators meet the new requirements by increasing network efficiency, instead of investing in non-AI infrastructure (e.g., adding new cells) which is more expensive.

Monetization. It is not clear if and to what extent operators will be able to monetize AI directly by offering new services. For instance, AI is going to be a crucial enabler for sensing (ISAC), and will widen the opportunity for drones and network APIs, but AI will not be the differentiating feature of the service. Operators have to sell a drone service to get revenues, not AI.

Enterprises will be heavy consumers of AI, and increasingly physical AI, and some of this use of AI will require connectivity over public mobile networks. But it is still unclear how much the role of mobile operators will extend, if at all, beyond that of providers of connectivity.

The AI-on-RAN monetization opportunity from leasing unused GPU capacity is the most problematic:

- In the short term, it would require operators to buy GPUs they do not yet need to lease capacity. It is doubtful that this could be a sound business model unless it is part of edge AI enterprise services, not specifically tied to the RAN.
- The RAN has extremely tight reliability and performance requirements, and operators will be cautious in opening up precious and vital GPU resources to third parties, especially in a competitive market where revenues will be constrained.
- Leasing spare capacity assumes that there will be demand outside RAN peak hours. But it is likely that third-party demand is correlated in time and peaks at similar times, and operators will not be able to meet third-party requests unless they overprovision the RAN, which is an expensive proposition.

Takeaways

Use AI where you need it. Not everywhere (yet).

AI does things differently. We need to learn to do things differently, too.

Integrate AI gradually, layering it from the top.

Choose the AI tools you need wisely to get a positive ROI

No hurry to get AI-native. No need to deploy GPUs to have an AI RAN initially.

Expect cost savings. Be wary of too optimistic monetization opportunities.

Read the edited transcripts below, or watch the conversations here

Go back to the top

AI in the RAN: A data-first path to full automation

How AI is transforming the RAN, from hype to deployment. A conversation with Karsten Gaenger, Principal Product Line Manager for the RAN at NETSCOUT and Monica Paolini, Principal at Senza Fili

Integrating AI into the RAN is no longer a fanciful promise: it has become an unavoidable tool for managing complexity, increasing network efficiency, and reducing costs in an environment where revenues are stagnant.

Karsten Gaenger, Principal Product Line Manager for the RAN at NETSCOUT, outlines a strategy for integrating AI into the RAN, starting with traditional AI/ML models and gradually integrating LLMs and AI agents, assisting operators on their path to Layer 5 automation.

- **Data reliability: Success depends on high-quality, correlated, AI-normalized datasets**
- **Gradual adoption path: Operators can start with 4G or 5G, and move forward at a manageable pace that fits their needs**
- **Benefit from a diverse ecosystem: APIs, telco-specific LLMs and MCP give operators end-to-end, vendor-independent visibility across capacity, mobility, and services**
- **Beyond KPIs: AI enables scalable root-cause analysis, prediction, and faster repair, with start-to-finish analysis and monitoring of procedures**

Monica: Artificial intelligence (AI) in the radio access network (RAN) is a hot topic, but one that is often approached from a high-level perspective that does not address the implementation details.

In this conversation, Karsten Gaenger outlines concrete ways in which AI can be integrated into the RAN and the benefits it can deliver. He has been at NETSCOUT for a long time, but he has a new role.

Karsten, could you tell us about your background and what you do at NETSCOUT?

Karsten: I oversee the RAN products from NETSCOUT on the RAN product line management team. I have been working in product management at NETSCOUT for more than a decade. Previously, I worked at Tektronix Communications for more than a decade, starting in a technical role in engineering. I started my career in the industry over 20 years ago in the academic domain at the Fraunhofer Heinrich Hertz Institute in Berlin. My research there focused on making video transmission robust in mobile networks.

I am happy to oversee the RAN products for NETSCOUT right now. It is a great time to be here, as the industry is changing significantly with AI and machine learning (ML).

Monica: I have been working with NETSCOUT for a long time. What I like about companies like yours is that you constantly try to improve existing networks. Clearly, AI offers an opportunity to take you one step forward by providing more tools to improve network efficiency, performance, and costs.

Let's start with the basic drivers. Why do we need AI in the RAN? What challenges are we facing, and how can we solve them?

Karsten: AI can play a comprehensive role in the future of our industry. I am quite bullish that it will help our customers be successful. We have seen for decades that network complexity is increasing at an accelerated pace. However, it is no secret that service revenues are flat and not growing in line with the adoption of new technologies.

Automation and machine learning are required to manage the network's complexity. 5G is a recent example. This was the first time that 3GPP had released a new generation that could be integrated into existing legacy networks. I think it is the right approach because operators did not want to deploy a whole new network to have coverage to generate

Watch the video of the conversation

commercial revenues. Instead, with close interworking with anchor 4G sites, 5G non-standalone (NSA) allows for 5G coverage in focus areas like downtown cities based on the 4G network. We are now moving into 5G standalone deployments that support more sophisticated use cases.

However, getting two network generations to interwork is hard. The 5G logo may appear on the screen, but the phone may not be connected to a 5G carrier. With carrier aggregation and dual connectivity, for instance, the user experience may not be as good as expected.

RAN teams need more visibility and better tools to pinpoint problems, predict issues, and propose solutions. We are on an accelerated journey toward the industry vision of Level 5 automation, which includes zero-touch and fully automated networks.

Monica: Most of AI and automation in the RAN is for the future because we are just starting. Adoption of automation and AI in today's networks is limited. But as complexity grows, networks would become unmanageable without AI. We could not hire enough people to manage them, even if we had enough qualified people to do so, and we do not.

How do you see the adoption process? What should operators do first? What is the low-hanging fruit?

Karsten: Operators do not have to go from zero to 100 in a step function; they have a choice to deploy AI and ML gradually. They can initially adopt machine learning modules in selected areas first and then gradually grow toward agentic AI. Large language models (LLMs) can eventually help non-technical staff access datasets and understand complex analytics.

Monica: What is NETSCOUT's proposed path for the operators?

Karsten: We have learned in the past that it is essential to manage the timing correctly. There is a risk of being too quick and making wrong decisions, but we also do not want them to be unsuccessful late followers. The challenge is to find the right middle ground.

From the early adoption of ML, we learned that input datasets are crucial to success. We need correlated and AI-normalized data where the correct correlation is embedded in the data. If datasets are uncorrelated, even the most complex AI cannot make sense of the information and extract the relevant relationships between data points. With the right datasets, we can feed models efficiently and get the best return on our investment.

Monica: That is a crucial part. If the data is inaccurate or unreliable, you are not going anywhere. It is a lack of good data that often causes hallucinations.

How do you ensure an operator has the right data, especially in a complex network with different vendors and varying formats and levels of granularity? How do you make sure it all fits together?

Karsten: This is an area where NETSCOUT concentrated early on. We saw that feeding industry data blindly into models resulted in a garbage-in-garbage-out outcome. To avoid this fate, we have worked for years on defining AI-normalized datasets to feed dedicated modules and agents. This helps RAN, device, performance, and optimization teams do their jobs more efficiently.

With AI, ML, and the right datasets, we can address tasks that were previously impossible to do at scale.

One example is single-call analysis. This requires state machine processes to analyze calls from start to finish and monitor all procedures, including handovers. You must understand what happened beforehand and what happens after a handover to see how it affects service quality. If you isolate data using only performance counters, the correlation is lost, and the outcome will be poor.

At NETSCOUT, we use a comprehensive view of the call from start to end, correlating the procedures and RF messages with service level insights to generate a real, holistic view of the network. From here, models can pinpoint root causes of service issues rather than just giving indications. They can also identify good neighbor relations and handover zones based on quality of service, latency, and throughput.

We have created AI-normalized datasets that use agentic AI algorithms that generate outcomes for capacity, mobility, and services — three areas where user experience is key. Common interface adoption and, when needed, definition allow for AI agents and datasets to communicate. We are also investigating how to implement the Model Control Protocol (MCP) server and client architecture to integrate agentic AI into our networks.

Monica: What makes NETSCOUT's approach stand out?

Karsten: We focus on vendor-independent datasets in the areas of capacity, mobility, and services. NETSCOUT's cloud-native solution serves a diverse set of customers, ranging from small mining companies with small 5G or 4G private networks to

the largest mobile networks with more than half a billion subscribers.

We work hard to preserve dataset integrity to retain causality. Solutions based on uncorrelated statistics or KPIs lose intelligence. For example, handover KPIs may indicate that the network works properly. However, even after an apparently successful handover, a subscriber may experience call drops or interference. If you do not correlate these procedures throughout the call, you will have an imperfect data view, which will cause even the best AI models to underperform.

Monica: As network complexity grows, you need an end-to-end view to understand why a user might be unhappy, even if individual elements appear to work well and KPIs are good.

How much of the AI, ML and GenAI models do you develop in-house versus using what is already available? How do you avoid reinventing the wheel while still being useful to operators with different networks and needs?

Karsten: NETSCOUT follows a twofold approach. We are not an AI research company; we apply AI to make sense of network data. We use existing modules, models, and algorithms to generate insights from mobile datasets.

We differentiate where AI is used. NETSCOUT plays a role in service assurance, optimization, and pinpointing issues to reduce mean time to repair. We look toward predictive analytics and self-healing for Level 4 and Level 5 autonomous networks.

The AI train has left the station at the Physical and Media Access Control (MAC) Layers. While NETSCOUT does not play a large role there, we see AI adoption in network equipment and Open RAN for interference coordination and scheduling. This is already appearing in products.

Regarding the journey of AI, we are still partly stuck in the hype phase, slowly realizing where AI is useful and what it cannot do.

Companies now realize the importance of keeping ground truth and AI-generated datasets separated to avoid pollution with misleading hallucinations and ensure that models learn from real network data.

NETSCOUT helps the industry to be more successful. For example, operators can increase network capacity by bundling more carriers and bandwidth. To have visibility into how these carriers and band combinations are performing per device and per subscriber, operators need a multidimensional analysis. KPIs alone cannot provide this correlation. NETSCOUT looks at the call from start to finish to establish the correlations needed to make sense of the data.

In the past, radio engineers had to spend hours analyzing a single call. This is not possible at scale for millions of calls. AI and automation can automatically analyze call behavior to provide engineers with situational awareness. From situational awareness, we can move into prediction and proposed problem resolution.

Monica: Operators, as well as vendors and analysts, want to adopt AI, but they must do so in organizations where many do not yet understand the technology. You need to build knowledge and trust to integrate AI successfully with what you already have.

How do you see this integration for your existing clients?

Karsten: There are two dimensions to this: technology and change management. There are risks associated with AI, as well as human adaptation issues.

NETSCOUT is committed to securing existing operators' investments and offers two deployment options: the AI Sensor and the AI Adapter. The AI Sensor is a data collector and processing engine that correlates information for the world of AI. The AI Adapter is for operators who wish to plug into their existing cloud-native processing chains as the first step in their AI journey.

At our 2025 ENGAGE user conference, our CEO, Anil Singhal, announced a NETSCOUT telco-specific LLM, based on the AI Sensor and AI Adapter, that will help organizations become more efficient and give their staff easy access to data that previously required expert knowledge.

A dedicated telecom LLM can provide RAN expert knowledge to everyone in the organization. You could type in a prompt asking how a video conferencing service is performing today. The model will inspect the relevant datasets, analyze them, and provide a well-crafted answer without requiring the user to have deep RAN expertise. It will provide detailed latency and throughput reports with graphs, predictions, as well as proposals for next steps and recommended actions. This can be used in various departments, from call center agents helping complaining customers, to RAN engineers finding root causes.

Monica: How do you deal with power consumption in the AI Sensor, AI Adaptor and LLM models? Some

say that adding AI to the RAN will lead to an explosion in processing and power requirements.

Karsten: Our customers do not need to build large-scale data centers for this. We use pre-trained models. However, we expect AI models will need to be trained within customer networks, as each network behaves differently. The models will learn in a specific RAN environment to provide tailored outcomes.

One example is our AI-based geolocation algorithm, which has been on the market for a few years. We have a dedicated geolocation team with many PhDs. Depending on the network's size, we may have hundreds of thousands of AI geolocation agents. Each individual cell has its own agent that learns the physics of that cell to provide the best geolocation outcome. We geolocate all calls in real time to make spatial sense of performance. The AI algorithm can accurately predict geolocation and recognize known locations, such as a specific street corner, over time, based on the RF environment.

We have achieved excellent results in power consumption efficiency and geolocation based on network footprint. While initial learning consumes power, the models are efficient once they are in a steady state.

Monica: After the learning phase, the model no longer needs to repeat the process continuously. For instance, the model retains the information tied to a specific location, such as the street corner you mention. In this context, what needs to be centralized versus what can be pushed to the edge?

Karsten: While we are totally flexible in this area, operators typically deploy NETSCOUT's products in data centers where they can holistically aggregate distributed data sources. Automated optimization tasks need a broader context, and the real-time nature of an edge deployment with millisecond latency isn't practical.

NETSCOUT supports distributed systems with AI agents at the MAC and PHY layers that do not necessarily need to be fed back to centralized systems.

One challenge we have in distributed systems is solving colliding interests. You might have one application dedicated to energy saving and another dedicated to performance. If you want to increase performance, you usually do not shut down carriers, which is exactly what the energy-saving application might do. This needs to be resolved at a centralized, higher level.

Monica: What is next? What are the highest priorities for NETSCOUT and the industry to move forward successfully?

Karsten: The industry needs to expand and strengthen interoperability. We need APIs and MCP to allow AI agents to talk to each other.

NETSCOUT is committed to providing domain knowledge using these methods for AI initiatives, such as those supported by the AI-RAN Alliance, which includes major industry players.

AI will play a role in 6G standardization, both at the network layer and the data layer. For example, the Fraunhofer Heinrich Hertz Institute is working with Ericsson and Nokia to define standards for 6G video codecs. This will allow source coding and network coding to work together, combining RAN, transport, and data layers in 6G networks to make the video codec network aware.

However, when asked about 6G, operators often say they do not yet have a fully functional 5G standalone network and want to leverage 5G before making the next major investment. Increasing efficiency through AI and automation is applicable to 5G and 4G today. Focusing on AI agents across various domains today will help us move toward a more holistic future with 6G.

Go back to the top

About NETSCOUT

NETSCOUT is a leading provider of service assurance, cybersecurity, and advanced AI/ML analytics solutions. Our portfolio spans both communications service provider and enterprise environments, unlocking the power of IP-based, Deep Packet Inspection (DPI) data to deliver unmatched visibility, performance, and security across hybrid networks.

With the rise of AI-driven operations (AIOps), NETSCOUT is helping customers simplify network complexity and improve user experience by delivering real-time, actionable intelligence. Our solutions enable faster mean time to knowledge (MTTK), proactive issue detection, and automated insights that support smarter decision-making and more resilient infrastructures.

Our market-leading service provider product portfolio is designed to provide end-to-end observability across physical, virtual, and cloud environments. By combining scalable monitoring with AI-powered analytics, we empower operators to assure service quality, secure applications, and accelerate innovation. At NETSCOUT, we help make every user experience exceptional by accelerating the connected world through Visibility Without Borders®.

About Karsten Gaenger

Karsten Gaenger oversees NETSCOUT's RAN product line and portfolio with a focus on analytics and service assurance sectors. He holds an electrical engineering degree in telecommunications and has worked extensively in research and test and measurement roles for more than 20 years. Prior to joining NETSCOUT in 2006, he published several IEEE research papers with the Fraunhofer Institute. Karsten also holds several patents in the telco domain and is the co-author of the Wiley book "LTE Signaling - Troubleshooting and Performance Measurement."

During this interview, NETSCOUT may have made forward-looking statements, which are subject to risks and uncertainties. These statements may cover a range of topics such as NETSCOUT market opportunities, market share, business initiatives, future product plans, and potential impacts on NETSCOUT performance. Actual results may differ materially from those expressed or implied in these forward-looking statements due to various known and unknown risks, uncertainties, assumptions, and other factors. NETSCOUT cautions you not to place undue reliance on these statements, which speak only as of the interview recording date.

Go back to the top

From AI in the RAN to AI for the RAN: Making innovation safe

Striking a balance between benefiting from AI and protecting networks, users, and society from its excesses. A conversation with Nirlay Kundu, Head of Technology Standards at IMDEA Networks, and Monica Paolini, Principal at Senza Fili

The role of AI in the RAN is evolving from AI in the RAN, to incrementally optimizing RAN operations and increasing cost efficiency, to AI for the RAN, to reimagine the RAN architecture, ensuring accountability and resiliency and protecting networks, users, and society from biases and unintended outcomes.

Nirlay Kundu, Head of Technology Standards at IMDEA Networks, guides us through a strategic approach to benefiting from and safeguarding against AI, drawing on his research and standardization experience, as well as his previous work at Verizon.

We talked about:

- **Accountability and governance with AI**
- **Predictive, adaptive, autonomous and explainable intelligence**
- **The essential role of humans in the loop**
- **Minimizing bias, maximizing learning**
- **AI at the edge and in the data center**
- **The latest innovations in agentic AI**

Monica: Today, we are going to talk about how AI is changing the RAN with Nirlay Kundu at IMDEA Networks in Spain. Nirlay has been working on the evolution of the RAN for a long time and, before joining IMDEA Networks, he was part of Verizon's Open RAN standardization efforts and held a leadership role in the O-RAN ALLIANCE.

Nirlay, how has your involvement in the RAN transformation changed over the years?

Nirlay: As I moved from Verizon to IMDEA Networks, I continued my standards work, and I represent IMDEA Networks in the O-RAN ALLIANCE. I also help research students at the institute in a technology-agnostic role, working on networking interfaces such as 6G or 5G, as well as on technologies such as AI and blockchain.

Can you tell us more about IMDEA Networks?

Nirlay: IMDEA Networks Institute is a research organization on data networks whose multinational team is engaged in cutting-edge fundamental science and technology. As an English-speaking institute in Madrid, Spain, IMDEA Networks offers a unique opportunity for pioneering scientists to develop their research vision. IMDEA Networks has established itself internationally at the forefront in the development of future network principles and technologies. Our team of highly reputed researchers is designing and creating today the networks of tomorrow.

Monica: We all agree that AI is going to have a major impact on the RAN. But will it abruptly redefine the RAN, or will it foster a gradual transformation?

Nirlay: Today, we are deploying AI in the RAN. Most AI deployments in telecom are geared toward optimizing existing architecture. Think of AI as a technician that tunes parameters, detects anomalies, schedules resources, and predicts traffic. These are often narrowly scoped and data-driven tasks, bound by existing protocols and infrastructure. This is smart automation, not strategy transformation.

What is emerging now is AI for the RAN. This is where the conversation becomes more transformative. We are starting to ask what a RAN should look like if AI is part of the design process from day one. This includes AI-informed spectrum policy and spectrum-sharing models, AI-guided slicing based on societal, economic, and mission-critical applications, and AI-driven architectural

Watch the video of the conversation

choices, such as when to deploy microcells, when to use satellite, and when to use edge compute. AI can also help rethink coverage, investment decisions, and even fairness using intelligent simulations and reward models.

AI for the RAN is not just about efficiency. It is about changing the goals, not just achieving them faster. Today's AI-in-the-RAN discussion is mostly about vendor capabilities, while the emerging AI-for-the-RAN discussion focuses on high-level regulatory, boardroom, policy, and strategy considerations. If we do not ask AI to help us redefine what we want the RAN to do, including societal goals such as digital inclusion, energy justice, and public service access, we risk building faster networks that replicate past inequalities. That is why it is important to discuss AI for the RAN now.

Monica: We often think about using AI to enable new applications and services, especially for the enterprise. With AI for the RAN, the network can fundamentally transform itself and become more closely tied to physical, social, political, and economic environments, increasing its usefulness and fairness.

But what kind of intelligence are we looking at for the RAN? Is artificial general intelligence (AGI) going to be relevant for the RAN if and when it becomes available?

Nirlay: Your question gets to the core of what we mean by AI for the RAN. The term "intelligence" is often used loosely in this context. For clarity and strategy planning, it is essential to understand what kind of intelligence we are dealing with. We should think of RAN intelligence not as a single capability, but as a spectrum of functions, each with different implications for design, control, accountability, and risk.

AI gives us predictive intelligence. It anticipates future states using time-series, historical, or real-time data. Example use cases include predicting traffic surges, anticipating handovers, and forecasting peak demand. Instead of reacting to events after they happen, operators can adopt a proactive approach. We still need human intervention and validation because outcomes depend on data quality and generalization error.

In the next phase, autonomous intelligence allows AI to make and execute decisions without human intervention. Use cases include self-tuning of antenna parameters, automated load distribution, and closed-loop network optimization. AI models can act faster than human operators, scale more easily, and reduce operating expenses by removing manual intervention.

The challenge is that AI can be difficult to govern. Because decision-making is often opaque, there is a risk of cascading failures. We still need to determine who is responsible if autonomy causes a service disruption or complication.

Another form of intelligence is adaptive intelligence, in which behavior evolves in response to changing conditions through a feedback loop. It enables networks to adapt to environment-specific usage shifts, new devices, and new mobility patterns, reducing model retraining costs and improving robustness over time. However, adaptive intelligence can drift in unintended ways, requiring robust monitoring and rollback mechanisms.

We also have explainable intelligence, which provides a traceable and understandable rationale for its actions and outputs. How do you decide how to allocate spectrum? When do you generate a dropped-connection alert? These capabilities are critical for compliance, debugging, and trust-building, especially with regulators and customers. The trade-off is between model complexity and performance.

Most AI deployed today is predictive and semi-autonomous. We are rapidly moving toward systems that are adaptive and increasingly opaque, which makes explainability and accountability non-negotiable.

Monica: AI is not just about crunching more data; it can also identify patterns that humans miss because of their experience bias. If something works today, I am going to do the same thing tomorrow. AI continues to explore alternative approaches across domains and use cases. How should we manage this AI-driven continuous exploration?

Nirlay: There is a learning curve. You need to manage interactions between autonomous systems and humans. A Responsible, Accountable, Consulted, Informed (RACI) matrix is required to define who is responsible for what. This is an extension of software entities that learn autonomously. There should be a feedback loop with continuous evaluation, in which humans assess correctness and intervene to adjust and improve algorithms when outcomes are misaligned with objectives.

Monica: We need to recognize that AI can bring benefits but also cause harm. How do we manage

that, and who takes responsibility? For example, autonomous driving can cause accidents, but the accident rate may be lower than with human driving. Still, people tend to be more critical of mistakes in AI-driven decisions or services, and we need to develop a more deeply grounded approach.

Nirlay: This is a critical issue. While AI promises greater efficiency, it introduces complexity, accountability challenges, operational risk, and legal issues. Ultimately, consumers use operator services, and operators own the customer experience, network performance, and quality of experience. Even if operators outsource decision-making to AI systems or third-party vendors, they are ultimately legally and contractually responsible for Service Level Agreement (SLA) compliance, Quality of Service (QoS) and Quality of Experience (QoE) guarantees, and regulatory compliance for things such as lawful intercept, net neutrality, and spectrum usage.

Right now, if AI changes something in the network, the operator is accountable. For example, if an operator reduces transmission power to save energy, it is responsible for any resulting service degradation, even if the decision was made by a vendor's AI. Operators must decide whether to deploy AI recommendations automatically or with a human in the loop, ensuring AI outputs are interpretable and traceable within the network operations center. Operators need a monitoring model to detect drift or abnormal behavior after deployment. Even when vendors supply AI, operators decide how and where they apply it and are liable for the outcome.

Governance, regulatory compliance, and risk management are also critical. Operators must maintain AI risk logs, implement fallback systems, and retain the ability to revert to legacy or rule-based optimization when things go wrong. We need to establish compliance mapping between AI decisions and telco regulations, and to be able to audit AI actions for forensic purposes. For example, we need to find out why a service outage or a spectrum violation happened: Was it AI or was the human in the loop? Regardless of the systems used, operators remain responsible for customer harm, such as unfair throttling, uneven access, or dropped calls.

Vendors, by contrast, provide the AI software and hardware and develop the algorithms, models, and platforms that inform and automate RAN control decisions. Unless specified in contracts, vendors are not legally liable for operational outcomes. They are responsible for the model architecture, the model training pipeline, validation in the test environment, such as simulation labs, and synthetic data. They track performance metrics, such as accuracy, convergence speed, and impacts on KPIs, and must satisfy SLAs. But this is where their responsibility stops. Models are trained on general datasets. Vendors do not have, or have only partial, real-world datasets from the operator, and cannot conduct full-fledged testing because operators will not share much of their data. This can lead to a generalized approach from vendors.

When vendors cause an AI-driven failure, they should document the failure modes and limits of model performance. They should clarify what data the AI model expects and how to avoid garbage-in-garbage-out outcomes. They should provide some error margins at regular intervals, specifically for high-stake decisions, for example, for spectrum assignments or handovers. Vendors are responsible for the overall security and robustness of the model and for ensuring that it is not vulnerable to manipulation, intentional or unintentional. They need to protect the model's intellectual property and data pipelines against attacks, certify compliance with standards such as those from the International Organization for Standardization (ISO) and the O-RAN ALLIANCE, and follow security best practices. That is the vendor's responsibility.

Monica: You raise crucial issues on how using AI creates the need for a new approach to assigning responsibility when things go wrong. But can we use AI to check AI? And maybe we will have to do that because the models are too complex for humans to analyze.

Nirlay: We are just at the beginning. A slight distortion of the data yields synthetic data for model training, initially resulting in a minor aberration. But when you run it in a loop 100 times or 1000 times, it just goes wild. That is what I am afraid of: how much real data we have and how much is the delta per loop. It ends up with a nasty prediction. This makes me worry about the capacity of AI monitoring AI.

Monica: This is similar to what happens with GenAI in the wider domain, beyond telecoms. LLMs are regurgitating what is already out there. If what is out there is false or inaccurate, AI amplifies the distortion. If you feed an AI model to optimize the network with inaccurate data, it is likely to go down the wrong tangent and may or

may not be able to self-correct. To be fair, this happens to humans, too, but the risk is on a different scale with AI.

This raises the question of how we should invest in and adopt AI. Not only do operators need to ensure everything works as planned, but they also need to decide where and when to integrate AI into the network. Where should operators deploy AI, at the edge or in the cloud? Are they overinvesting or being too slow? How much computational power do they need at the edge? Where do they need GPUs?

Nirlay: My advice is to start slow and small, and see whether AI produces results, whether deployed at the edge or in a data center. Recent developments show we can achieve high-performance results without high-power GPUs through algorithm optimization. Software performance is going to improve quickly right now because everyone is thinking about optimizing their software and using less GPU power, compute, and energy.

At the end, you need to figure out how much you are saving. Are you saving on opex? Are you saving on capex? Mostly, it will be opex savings. If the savings are on the bottom line, or if the numbers are bigger on the top line, that is where we should focus. Where is the biggest bang for the buck? We are still learning in this game. Focus on the areas with the greatest return on investment, and avoid concentrating all efforts in a single area. Experiment across different parts of the network to identify where you create the most value.

Monica: Not all operators will reach the same conclusion. A common concern is that AI requires so much power that the higher energy costs offset any opex savings from increased network efficiency. I would argue that as models improve, training and inference will become more energy efficient.

Nirlay: Yes, this is a long-term effort. Another way to address this is to use SLMs, which are less energy hungry and can address some privacy issues, as the data can be confined to within the company premises.

Monica: This creates challenges because in telecoms, as well as in other sectors, we look at the short-term financial impact. We look at quarterly results, and AI is not a quarterly kind of thing.

Another potential benefit of AI is its ability to discover things we did not know. But because we did not know them, it is difficult to assess their value. What is the best way to transfer our expertise, accumulated over a long period across multiple domains, to an AI model?

Nirlay: This is a crucial question, especially in the context of AI for the RAN, where decisions can affect millions of users, critical infrastructure, and regulatory compliance. AI lacks visibility into many physical realities of the RAN, i.e., what AI can't see and what sensors don't capture, such as local topography shifts, new buildings that block signals, microclimate or seasonal changes, and seasonal interference. Let's say a field technician's actions are not logged in, or that hardware faults or fiber cuts are not reported. AI doesn't know exactly what is going on. The AI may optimize based on an incomplete or outdated view of reality. That is an urgent issue right now.

AI does not understand human context and intent. AI doesn't understand why we set the policies or thresholds we do. For example, do AI models understand why emergency coverage requirements for hospitals are as they are now? Do they understand the legal and political implications of prioritizing one region over another? AI still doesn't understand the context well. Socioeconomic factors are also a big challenge because they are subjective, for example, in spectrum allocation. AI may optimize technically, but it may do so unethically or illegally because of the lack of understanding of the wider socio-economic human context.

Another concern is that AI is trained on time-series past data and may fail in new scenarios. Suddenly, there are mass gatherings, protests, or festivals. Out of the blue, something is happening. Let's say there are cyberattacks that cause interference. Recently, in Spain, there was a power outage. If AI had acted on that, it would have been a failure. Humans intervened to take the situation under control. AI is not intelligent enough today to respond to malicious interference and cyberattacks. With drones and autonomous cars, we may suddenly see congestion that AI cannot predict. AI would not know why new mobility patterns, generated by devices it does not know about, happen. The AI may not know how to handle these events and may make the wrong decisions. Humans are still required to intervene in such cases.

The training data fed to AI models is biased. Under-representation of certain communities, for example, low-income or rural communities, in the

training data may be unintentional but have undesired results. Vendor data silos across the RAN and the core, clerical mistakes, and manual errors may amplify AI biases, leading to uneven performance or unintended outcomes.

To address these limitations, a human-in-the-loop approach remains essential. AI systems must allow expert override in high-impact situations. We need a backup plan for human intervention to enforce quality of service (QoS). We need to develop confidence in the model to resolve this uncertainty.

AI should express its confidence. Maybe there could be a flag for out-of-distribution scenarios and support rollback mechanisms. Every autonomous loop should include clear, well-documented, safe-mode mechanisms that revert to conservative, well-tested methodologies.

AI in the RAN is powerful, but not ubiquitous. It still has its blind spots, and we cannot ignore them. They must be anticipated, monitored, and governed. Good AI does not assume that you don't know everything. The model should not be the be-all and end-all. Ideally, AI models should recognize their limits and request human intervention when they encounter something they do not know.

Monica: AI models need guidance, just as kids need education. Just as you do with kids, you need to push models in the right direction. Humans must provide strategy and direction to prevent something bad from happening. For instance, you may want to guide AI models to recognize and address bias in the data, for instance, to ensure you do not worsen the digital divide as we discussed earlier.

Do you know when you are not overcompensating for the bias in the data? Are we substituting the initial bias with another bias?

How is this different from what we are already doing to correct for bias in our assumptions and data? AI amplifies bias, and we need transparency to be aware of and minimize its impact. It is a challenge to devise where to draw the line because there are security and ethical tradeoffs. For instance, do you balance security with a potential reduction in the quality of experience?

How does agentic AI fit into all this?

Nirlay: Agentic AI is a crucial area of AI development today because it enhances the network's autonomy. It operates independently, deciding and acting on its own without constant human guidance. It is able to learn and adapt to new information in collaboration with humans. It can provide assistance and insight into complex situations. By relying on pattern recognition and probabilistic reasoning, it provides flexibility and adaptability in dynamic conditions.

Go back to the top

About IMDEA Networks

IMDEA Networks Institute is a research organization on data networks whose multinational team is engaged in cutting-edge fundamental science and technology. As an English-speaking institute in Madrid, Spain, IMDEA Networks offers a unique opportunity for pioneering scientists to develop their research vision. IMDEA Networks has established itself internationally at the forefront in the development of future network principles and technologies. Our team of highly reputed researchers is designing and creating today the networks of tomorrow.

Some keywords that define us: 5G, 6G, Big Data, blockchains and distributed ledgers, cloud computing, content-delivery networks, data analytics, energy-efficient networks, fog and edge computing, indoor positioning, Internet of Things (IoT), machine learning, millimeter-wave communication, mobile computing, network economics, network measurements, network security, networked systems, network protocols and algorithms, network virtualization, software defined network (SDN), network function virtualization (NFV), privacy, quantum communication, social networks, vehicular networks, wireless networks and more.

About Nirlay Kundu

Nirlay Kundu is a telecom thought leader with 25+ years of experience spanning global operators, vendors, startups, and academia. He has led strategic partnerships, standards initiatives, and innovation programs across 5G, Open RAN, cloud, and emerging 6G technologies. A former Distinguished Engineer at Verizon, he has worked extensively with Tier-1 operators and standards bodies, including the O-RAN ALLIANCE, ITU-T, and 3GPP. Nirlay currently bridges industry and academia as Head of Technology Standards at IMDEA Networks and Professor of Practice at leading universities. He holds an MBA from Babson College and an MS in Engineering from the University of Leicester.

© 2026 Senza Fili

Go back to the top

About Senza Fili

Senza Fili provides advisory support on wireless technologies and services. At Senza Fili, we have in-depth expertise in financial modeling, market forecasts and research, strategy, business plan support, and due diligence. Our client base is international and spans the entire value chain: clients include wireline, fixed wireless, and mobile operators, enterprises and other vertical players, vendors, system integrators, investors, regulators, and industry associations. We provide a bridge between technologies and services, helping our clients assess established and emerging technologies, use these technologies to support new or existing services, and build solid, profitable business models. Independent advice, a strong quantitative orientation, and an international perspective are the hallmarks of our work. For additional information, visit www.senzafili.com.

About Monica Paolini

Monica Paolini, PhD, founded Senza Fili in 2003. She is an expert in wireless technologies and has helped clients worldwide to understand the technology and customer requirements, evaluate business plan opportunities, market their services and products, and estimate the market size and revenue opportunity of new and established wireless technologies. She frequently gives presentations at conferences, and she has written many reports and articles on wireless technologies and services. She has a PhD in cognitive science from the University of California, San Diego (US), an MBA from the University of Oxford (UK), and a BA/MA in philosophy from the University of Bologna (Italy). You can contact Monica at monica.paolini@senzafili.com.

© 2026 Senza Fili. All rights reserved. The views and statements expressed in this report are those of Senza Fili, and they should not be inferred to reflect the position of the sponsors or other parties involved in the report. The document can be distributed only in its integral form and acknowledging the source. We edit transcripts for clarity and readability. No selection of this material may be copied, photocopied, or duplicated in any form or by any means, or redistributed without express written permission from Senza Fili. While the document is based on information that we consider accurate and reliable, Senza Fili makes no warranty, express or implied, as to the accuracy of the information in this document. Senza Fili assumes no liability for any damage or loss arising from reliance on this information. Trademarks mentioned in this document are the property of their respective owners. Photos by Senza Fili and Gemini, unless otherwise noted.