RCR Tech
NEWS. INSIGHT. IMPACT.

MARKET PULSE REPORT

TELCO AI

REPORT PRODUCED BY RCRTECH
EDITED BY CHRISTIAN DE LOOPER, EDITOR

Sponsored by

amdocs CISCO EMERSON Red Hat ZTE

Contents

RCRTech Takeaways

Operators should treat the gap between frontier AI and telecom as the first problem to close. General-purpose models fall short on 3GPP specifications and network tasks out of the box, and building frontier models from scratch fails on cost. The proven path is fine-tuning strong foundation and open-weight models, measuring them against telecom-specific benchmarks, and collaborating on shared problems, digital twins and synthetic data rather than on raw network data.

Data readiness, not data volume, is the gating factor for telco AI. Operators are drowning in data that is fragmented, missing context and not machine-ready. Consolidate, contextualize and correlate that data, embed shared ontologies so semantic understanding lives in systems rather than in people's heads, and target narrow, high-value use cases first instead of trying to prepare everything in advance.

AI in the RAN is commercial now, but it must clear a telco-grade bar. Embedded models have to beat algorithms the industry has tuned for three decades, infer within sub-millisecond windows, scale to trillions of daily operations and run on installed hardware. Gate every feature with an "AI worthiness" test that weighs gains against added cost and complexity, and keep a fallback to rules-based functions in case models drift.

Autonomy must be earned stage by stage, with probabilistic intelligence kept away from load-bearing actions. Follow the assist, recommend, act, orchestrate path — skipping stages produces impressive demos and nervous operations teams. Use AI freely for insight, correlation and diagnosis, but reserve network-touching actions for deterministic playbooks, confidence thresholds and human approval. Accountability stays with the operator, whatever an agent decides.

Measure outcomes, not AI activity, and build the instrumentation in from day one. Success is incidents resolved faster and opex reduced, not agents deployed or tokens burned, and value must be traced through an explicit causal chain to the CFO. Choose use cases for business value rather than technical feasibility, minimize the cost of failure instead of chasing success rates, and remember that most of what determines success is people and process, not technology.

The endgame is a network that sells intelligence alongside connectivity, and positioning starts now. The business model is shifting from bytes to tokens, and an AI-native 6G network will orchestrate communication and compute as one system serving hundreds of billions of agents. Operators should monetize strengths they already have, such as edge inference, sovereign infrastructure and deployment muscle, and unify network, cloud and AI operations before the wave arrives.

Introduction

Telco AI is no longer just a proof-of-concept. AI is no longer an overlay on the network. It is being embedded in the network itself, from telco-grade models running inside basebands and radios to agents working the incident chain. At the same time, the network's customers are changing. AI-generated traffic, physical AI and, eventually, hundreds of billions of agents will bring demands that today's networks were never designed around, including heavier uplink, deterministic latency and compute on tap.

That changes the equation. The industry no longer needs convincing that AI works in telecom – AT&T alone is processing tens of billions of tokens a day, and live RAN trials are measuring double-digit gains at the cell site. The task now is to industrialize it, with models that clear a telco-grade bar, data that machines can actually use, autonomy earned stage by stage, and value traced all the way to the CFO. This report, which draws on insights shared during RCRTech's Telco AI Forum, examines how operators, vendors and researchers are converging on that challenge, and what has to happen next to turn AI's momentum into durable operational and commercial advantage.

1

Open telco AI – foundations of telco-grade AI

Telco AI has an uncomfortable starting point. The frontier models driving the broader AI boom do not deeply understand telecoms. Closing that gap is the foundation on which everything else in this report is built, and it runs through domain benchmarks, telecom-specific models, disciplined adoption programs and shared industry problems.

The Telco AI Forum opened with a panel that put the problem in plain terms. Moderator Louis Powell, director of AI technologies at the GSMA, noted that when general-purpose frontier models are tested against 3GPP specifications or network troubleshooting tasks, they fall short out of the box. "That is a cost for the industry, whether that be inaccuracy, too expensive models, too slow, or models we don't trust," he said. The GSMA's response has been its Open-Telco LLM Benchmarks, launched in early 2025, which rate frontier and open-source models against seven telecom-specific tests. Increasingly, operators are launching models of their own, with SoftBank, China Telecom and AT&T among them.

Adoption at 30 billion tokens a day

AT&T's numbers suggest the adoption question has already been answered, and the real question is scale. Mark Austin, the operator's VP of data science and AI, said AT&T is processing "about 30 billion tokens a day" through its large language models, with "close to 150 use cases in production." The operating model behind that rests on two mechanisms. The first is an internal innovation fund on which business units compete for backing based

on return on investment. “Two years ago we had 2x ROI, last year we had 5x ROI,” Austin said. The second is a constant training drumbeat, with weekly generative AI office hours, multiple internal summits a year and, over the past two years, a deliberate push to teach teams to build agents as workflows on Ask AT&T, the operator’s generative AI platform.

NTT Docomo’s Yoshihiro Nakajima, whose focus is the operator’s cloud-native network transformation and autonomous network work, described a deliberately operations-first approach that starts from real operational pain points rather than the technology itself. Deploying and operating a mobile core has traditionally meant hundreds of documents, complex configurations and manually aligned parameters. Docomo is working to collapse that process by combining agentic AI with GitOps. Agents generate configurations from natural-language intent, Git serves as the single source of truth, and CI/CD pipelines automate the network lifecycle end to end. At MWC, the operator demonstrated a fully automated, AI-powered deployment of its core network on public cloud, from intent to execution.

Fine-tune, don’t build

If the panel agreed on one thing about model strategy, it was that building a frontier model from scratch is the wrong answer. Mérouane Debbah, professor at Khalifa University in Abu Dhabi and senior director of its Digital Future Institute, traced the origin of his team’s Telecom GPT series to exactly that calculation. Off-the-shelf models performed “relatively poor[ly]” on the telecom use cases his team defined, and a from-scratch frontier build was ruled out on cost. Fine-tuning strong open-source models instead produced results he called “quite striking” on 3GPP datasets, telecom Q&A and coding benchmarks, strong enough to spawn regional variants, including an Arabic-language Telecom GPT built with local operator data.

Austin framed the same discipline as a “tokenomics question” and described four levers AT&T pulls once a use case is proven on a foundation model. Because open-weight models catch up to frontier quality within six to 10 months, roughly a quarter of AT&T’s tokens run on foundation models more than a year old that can be migrated to open weights with little more than prompt tuning. High-volume use cases on newer models get fine-tuned. Smart routing addresses a human weakness: “Given the choice, people don’t make very good routers themselves… you’re always going to use the smartest model, the most expensive model… but most of the time you don’t need that.” And once models are open-weight or fine-tuned, hosting them on-premises squeezes out further cost. Nakajima described a similar hybrid discipline at Docomo, where hyperscaler foundation models supply general reasoning, augmented with telco-specific knowledge such as standard operating procedures and past operational records. Keeping answers stable and narrow enough for production networks doubles as the operator’s main defense against hallucination.

Debbah then pushed the domain-model argument beyond text entirely. “Telecom is all about waves,” he said, and his team asked “what would it take to start chatting with waves?” The result is RF GPT, which tokenizes radio-frequency signals so that a model can identify the technology being measured, characterize its features and suggest optimizations. Trained on a mix of real measurements and a large synthetic dataset, it is the first in a research line he calls Large Perceptive Models — LLMs for IoT that could eventually tokenize sensing modalities from RF to temperature and pressure.

Shared problems over shared data

That synthetic-data dependency shaped the panel’s view of collaboration. Debbah was candid that operator alliances built around sharing data sets have “not been so fruitful,” arguing the more promising path is

standardized, large-scale digital twins that can generate extensive training datasets as shared community tools, with pooled compute so that countries which cannot afford it can still innovate. Austin agreed data-sharing is hard but argued the industry should not discount the value of posing good common problems. The GSMA's AI Telco Troubleshooting Challenge, built around its TeleLogs dataset, gave vendors a construct they could simulate and AT&T a framework it could apply internally, and Powell noted the challenge drew some 1,500 applications, with an agentic troubleshooting extension now underway.

Asked to name the open problem that matters most for agentic AI and 6G, Debbah described a coming "internet of intelligence" in which the network's primary users are no longer people. "We're talking about a telecommunication network which will be serving hundreds of billions of agents," he said, arguing the bottleneck will shift from compute to the interconnect — the lesson, he noted, of Nvidia's Mellanox acquisition in the data center. The foundations of telco-grade AI, in other words, are being laid for a customer base that barely exists yet

2

Fixing data bottlenecks before scaling AI

If the opening panel established that telco-grade AI needs telco-grade models, the second established that neither matters without telco-grade data. The bottleneck, the panel agreed, is not access to data, which operators have in overwhelming volume, but data that AI can actually use.

Moderator Christopher Silberberg, who leads IDC's telecom operations and monetization research, noted that every wave of digital transformation has turned on the same question of getting data into the right formats, with the right accessibility and the right governance models, to create new value. His opening challenge to the panel was to name their single biggest data bottleneck, and the three diagnoses quickly converged. For Jose Carlos Mendez, director of product marketing for network and OSS at Amdocs, the problem is not volume but readiness. The data available to CSPs "isn't AI-ready… it's not machine-ready. It is often fragmented, it's missing context, it's often disconnected, so it's hard for AI to make sense out of it." Diego Lopez, senior technology expert at Telefónica and chair of ETSI's Technical Committee for Data Solutions, named heterogeneity: "There is heterogeneity in users, heterogeneity in use cases, heterogeneity in devices, heterogeneity in technologies, and even heterogeneity in purposes" — including inside the operator itself. And Ed Fox, CTO of MetTel, said he had lived the problem since 2016, when training models on unstructured network data forced the company to build a data stream and data lake where everything the network produced could be subscribed to in a common format.

Breaking silos that were designed in

Lopez argued that the silos frustrating AI teams are not accidents; they were designed in. The 3GPP architecture and the industry's layering models were built to separate concerns and separate access, and AI now demands precisely the opposite — share what you know, export what you have, because something or someone downstream will take decisions and actions on it. Breaking those silos is hard for technical and organizational reasons at once. Technically, telecom's vocabulary is overloaded: "Think about session, think about domain, think about address. Everything is depending on the context." Fourteen years after joining Telefónica from academia, Lopez said he still encounters parts of the company using entirely different terms for the same thing. Organizationally, the reluctance to share data is often legitimate, and it is a governance problem. How do you guarantee data will be used the right way, for the right purpose, and keep proof of it? Fine-grained solutions, he conceded, do not yet exist, though Telefónica is working on them internally and through ETSI.

Mendez translated the semantics problem into a delivery method. Consolidate the data, then contextualize it, then correlate it. Context is where the trouble lives, because a "site" means something different to inventory management, assurance and asset management. That is why ontologies that define both terms and the relationships between them, horizontally across domains, are becoming essential; if AI is to deliver any autonomy across domains, it needs that common understanding embedded in systems rather than in people's heads. "Up until now, we humans have been kind of the semantic glue," Mendez said. TM Forum has begun exploring the move from its shared information data model toward true ontologies, though he expects several domain-specific ontologies rather than one universal model.

Automation moves to the front

Fox offered the panel's longest view of how the tooling has evolved. When MetTel brought in machine learning academics a decade ago with a hundred candidate projects, the most useful thing the ML told them was what to automate. The result was what MetTel calls intelligent process automation, built from single-purpose machine learning models, each trained on one question, nested together through automation. Now the pattern is inverting, with automation moving to the front of the pipeline to clean and structure data so that frontier and open-weight LLMs can act on it. Silberberg captured the shifting goalposts with a borrowed line: "AI is whatever we haven't automated yet."

Mendez described Amdocs applying agents to the data problem itself, with data-ingest agents trained on industry- and CSP-specific ontologies that check quality, identify gaps, fix them where possible and escalate to humans when needed. That iterative loop of agents, governance and human review, he argued, is how trust gets built, particularly with TM Forum benchmarks ranking data quality among operators' highest challenges. Lopez punctured the terminology even while endorsing the approach: "People are using agents right now for things that some time ago were called microservices, some time ago were called modules or functions." Whatever the label, he said, distributed workloads that enhance and share data while enforcing policy are "a general and very wise design principle" — one that matches both the distributed nature of the network and operators' insistence on sovereignty, meaning retaining control of what is theirs.

Standardize interfaces, not architectures

The closing exchange turned to trust and where to invest. Fox's answer on data integrity was a philosophy of speed. MetTel's per-vendor collectors keep source data clean, and on incident management, "it's better to make a

wrong decision in a couple milliseconds than it is to wait 30 minutes for someone, a human, to pick it up and be too busy and make a wrong decision at that point." Mendez put the investment case for the data layer in one line: "if you get data right, AI will just multiply its effect." He added the neglected people factor, arguing that staff need to understand and be able to explain how agents treat data before they will trust what comes out. Lopez, for his part, warned against over-standardization. He is "strongly fighting against" standardizing architectures, because "standards are for interoperability." The goal is to identify the interfaces where interoperability mechanisms genuinely apply, or risk "a forest of useless specifications."

Asked for the AI equivalent of the industry's G-by-G roadmap, Fox pointed to the RAN, where he sees AI changing the industry and the consumer experience most profoundly. It is where this report turns next.

3

From AI models to measurable impact – making AI-RAN work at scale

After two panels of diagnosis, the forum's first presentation offered a prescription, and it arrived with fresh news behind it. Just days earlier, Ericsson had announced AI in RAN, bringing telco-grade AI models directly into basebands and radios to deliver gains in performance, efficiency and energy savings without additional hardware. Gabriel Foglander, head of strategic RAN leadership at Ericsson, used the session to explain what that means in practice, and why AI in the RAN is a fundamentally different engineering problem from the AI most of the industry has been experimenting with.

Foglander framed embedded AI as a paradigm shift that touches three domains at once. First, the vendor domain, with AI models built into the network equipment itself. Second, operations, since the RAN produces data that agentic AI will increasingly consume, augmenting how networks are run and optimized. And third, traffic, because the capacity, spectral efficiency and user-experience gains that AI-RAN unlocks will help carry a coming wave of AI-generated traffic from physical AI, multimodal interactions and machine-to-machine communication. That traffic shares a common denominator, he noted, in far heavier reliance on uplink than today's mobile broadband and a need for more consistent latency.

What "telco-grade" means at the cell site

The heart of the presentation was a definition. A telco-grade AI model for the RAN, Foglander argued, must satisfy four criteria simultaneously. It must be outcome-driven, delivering measurable gains over algorithms the industry has built and tuned for 20 to 30 years — a high bar for very specialized tasks. It must operate on a strict time budget, with sub-millisecond inference windows that require models to be deterministic in the time domain, a sharp contrast with the relaxed timing of large language models. It must offer scalable inference, because network-wide deployment can mean trillions of inference operations per day, and the economics collapse if each one draws meaningful energy. And it must fit into existing hardware, which pushes toward compact models that leverage the installed base rather than demanding new silicon. "It's not really comparable in any way, shape or form to a large language model that sits in a data center somewhere," he said. "This is very specific to RAN needs."

The launch puts models against exactly those specialized tasks. There is a coverage-prediction model that smooths multi-layer spectrum transitions without flooding the network with control signaling, an AI-native scheduler for link adaptation that finds the most efficient link between network and user, smarter beamforming schemes that unlock capacity in massive MIMO radios, and an AI enhancement to 5G Advanced positioning that significantly improves precision in non-line-of-sight conditions. Alongside the models, Ericsson is opening machine-readable interfaces to raw RAN data and richer observability — feedback the models need to understand their own performance, but also data operators can consume northbound to design networks around more sophisticated, AI-derived metrics. Measured in live networks, Foglander said, inferencing the models draws "very little additional, if any, energy consumption in the existing equipment." And because it all runs on hardware customers have already deployed, scaling doesn't wait on a hardware cycle.

Train on the job, or train for the world

Asked how training strategy separates approaches that scale from those that don't, Foglander described a spectrum rather than a single method. At one end sit highly context-dependent problems, such as coverage prediction, where every cell in the world has its own nuances; those models are designed to "train on the job," learning from the network data captured in their natural deployment conditions. At the other end sit physics-based problems low in the Layer 1 stack, such as radio propagation, where a global model can generalize well, provided it is exposed to a diverse enough set of situations to "smoke out the corner cases." After that, it can be deployed anywhere without retraining. In between sit problems that borrow from both, with models trained on complementary real-world deployments to accumulate learnings that transfer across many environments. The differentiator, in his telling, is not just the right data but the right type of exposure.

No substitute for practical experience

Ericsson enters commercial scaling, Foglander said, with roughly 15 customers already running live AI-RAN deployments and trials, and models deployable at commercial scale starting in the second quarter of 2026. Foglander was direct about the robustness bar. No operator is in the mood to compromise operational performance and KPIs, so embedded AI must demonstrably augment the network, never detract from it. His near-term advice to operators came in two parts. On traffic, standalone 5G networks already have the feature set to tailor connectivity for AI's demands, so operators should stand up an AI-optimized slice now, not least because doing so generates early insight into how that traffic gets monetized. On embedded AI, "there's really no substitute for practical experience." Deploying the models available today surfaces the operational questions that conceptual architecture work alone won't. If models get more control, operational processes must change around them;

if models are separated from features, weights may need updating independently of the running system. That journey toward AI-native networks, he argued, can only be worked out in close collaboration between vendors and CSPs.

For an industry eager to move from pilots to production, the closing message was strikingly direct. The models are commercial, the hardware is already in the field, and "it's there for you to start right now." Whether operators take that invitation, and how they govern what they switch on, is where the forum turned next.

4

Agentic AI – from pilot to production, risk & governance

If Ericsson's message was to start now, the fireside that followed asked the harder question of how to start without breaking things, or people's trust. Philippe Ensarguet, Orange's vice president of cloud and software engineering, opened by rejecting the most common way of measuring agentic AI. Success, he argued, is not AI activity, measured in agents running and tokens burned, but operational outcomes end to end — incidents resolved faster, fewer customer-experience disruptions, operations teams spending less time coordinating manually.

Getting there, he said, follows a maturity path from assist to recommend to act to orchestrate, and each stage exists for a reason. Assistive AI makes the human faster and better informed while cognitive load shifts progressively toward the agent; bounded autonomy comes next; orchestration across domains comes last. "Telcos do not succeed by jumping directly into full autonomy," he said. Trust is earned stage by stage, and skipping one produces either resistance from teams who don't trust the system, or a production failure nobody can seriously explain. Those who skip straight to autonomy, he warned, will end up with a "very, very impressive demo" and very nervous operations teams.

Business value first, techno-push last

Orange's live agentic work spans the stack. In security, the operator demonstrated an agentic solution at Orange Open Tech last November that manages security events at 5G scale, with multiple agent ecosystems listening to network activity, connected to the SIEM and Orange's knowledge bases, and generating new signatures for undetected or unwanted traffic live. In telco cloud operations, Orange is working through the Sylva project with fellow operators in Europe and beyond on root cause analysis across the cloud-native stack, using a dedicated MCP ecosystem to triage layers that have become too numerous for manual investigation. In the RAN, agents are adjusting energy consumption in real time against traffic patterns and grid constraints, automating configuration and upgrade decisions across the lifecycle, and correlating, diagnosing and proposing ticket resolutions for support engineers. A Sylva white paper, he noted, goes deep on four of these use cases, covering autonomous RAN optimization, AIOps with autonomous remediation, intelligent network slicing and security.

What separates the pilots that scale from those that don't, Ensarguet argued, is how they were chosen. Orange applies a "high value scenario" methodology that starts not from what is technically feasible but from what would create the most business value across its affiliates, with cross-country alignment so a proven case can be leveraged everywhere. "The use cases that scale are the ones chosen for business value, not for technical implementation," he said, adding a broader charge that the telecom industry is "way too much techno-push. We need to start with the problem and not with the technology."

The causal chain to the CFO

The industry's mood on AI economics has swung, in a matter of weeks, from token maximalism to token discipline, and Ensarguet was candid that anyone offering a clean method for measuring token ROI "is certainly oversimplifying." His instinct is to treat tokens as a unit cost, but to measure them at the level of real outcomes — token cost per assisted decision, per resolved case, per completed workflow. Aggregate token consumption is meaningless without context.

From there, value must be traced through an explicit causal chain. The AI was used, decisions changed, operational KPIs such as mean time to repair, first-time fix rate and churn risk improved, and measurable business value followed in lower opex, protected revenue or avoided penalties. The flashy alternative doesn't survive scrutiny. "We introduced AI, and revenue went up — this is correlation, it's not proof," he said. Getting closer to causation requires A/B testing where possible and staged rollouts with control groups. "Each link needs evidence. If one link is weak, the whole chain is dead." And the instrumentation has to be built in from the start: "If you're trying to answer retroactively after your CFO is asking the question, you're not solving the right equation."
Safe enough to trust, simple enough to adopt

Asked to balance governance against trust, Ensarguet rejected the premise. The goal is not governance over trust but governance that enables it, and he framed the design problem as avoiding two failure modes. Governance that is too light leaves people worrying about risk: "No one wants to be the person who approved the agent that caused an outage." Governance that is too heavy stalls adoption, because using the tool takes longer than doing the job manually — at which point engineers route around it, and shadow AI becomes the enterprise's biggest problem. The answer is risk-proportional control, with human-in-the-loop approval and audit trails for high-impact actions, log-and-proceed with sample reviews for low-risk ones, and observability by design across every layer. "The best governance model makes AI safe enough to trust and simple enough to adopt," he said. "If your governance framework is something people run around, it's not governance, it's just documentation."

His closing assessment split into what the industry is getting right and what is still missing. On the first list, model

capability is no longer the bottleneck it was two years ago; the intent-based paradigm at the heart of cloud-native transformation is proving to be the right paradigm for AI too, shifting integration from programmatic to semantic; and the open source momentum around frameworks, tooling, governance and serving stacks is "extremely genuine." What's missing is the substrate for scale — security, observability and above all the operating model. Telcos are brownfield environments, and an agentic ecosystem that isn't coherent with the layers around it, including agreement on intent when different domains declare different things, will not hold. The GitOps discipline operators have built becomes newly important here, enabling human-in-the-loop validation, agent runtime isolation and policy guardrails. The production-grade enablers exist, Ensarguet concluded; the work is assembling them. Skip it, and the industry gets brilliant agents running on something totally unstable.

5

Q&A with ZTE

ZTE's answers were provided in writing and are lightly edited here for style

The operator business model is shifting from selling connectivity to selling intelligence. What does "AI for telco" concretely mean in 2026?

The essence of "AI for Telco" in 2026 is a paradigm shift in the global telecom business model: from "selling connectivity (Bytes)" to "selling intelligence (Tokens)." Telecom operators are transforming from "connectivity pipe providers" into "intelligent service providers." AI is no longer merely a tool for network cost reduction and efficiency gains (improving spectrum efficiency, energy efficiency, and O&M efficiency), but a driver of new revenue streams.

Under this trend, "selling intelligence" comprises two progressive layers:

Layer 1: Connectivity Upgrade — Intelligent Connectivity. By embedding AI into the network foundation, operators build an intelligent pipe that is "perceptive, cognitive, and orchestrated." Through differentiated experience assurance and tiered SLA offerings, they address the deterministic requirements of emerging services such as live streaming, cloud gaming, AI glasses, and embodied intelligence — improving connectivity ARPU and solidifying the subscriber base.

Layer 2: Beyond Connectivity — Tokenized Intelligence Supply. Token becomes the new pricing unit, enabling operators to directly deliver AI compute and model inference capabilities to users. This realizes a business-model leap from "connectivity value" to "intelligence value."

China's three major operators are pioneering this dual-track path. On the Intelligent Connectivity side, they are promoting 5G-A experience-guaranteed packages with tiered SLAs to boost connectivity ARPU and high-value user stickiness. On the Tokenized Intelligence side, they are actively exploring Token business models: China Mobile unveiled its Token operation ecosystem at the 2026 China Mobile Cloud Congress, introducing a unified metric of "one Token account for every user." Its MoMA platform integrates 300+ models, cutting per-Token costs by ~30% through consolidated operations, with a commitment to invest billions in Token ecosystem resources over the next 3–5 years. China Telecom launched its Xingchen TokenHub 1.0 platform, featuring multi-model aggregation and intelligent routing, and rolled out three pilot commercial Token packages on May 17, 2026, offering integrated "Token + connectivity + security" services. China Unicom introduced the "Agent + Token + AI Cloud" paradigm at MWC 2026, upgrading its Token offering into dual-track personal and team editions, delivered through its Yuanjing MaaS platform.

The core logic of the "connectivity to intelligence" transition lies in deeply converging network and AI capabilities, using Token to restructure the "connectivity–compute–intelligence" value chain, and upgrading connectivity subscribers into intelligent service consumers. Yet the real challenge is twofold: Layer 1 requires a technical closed loop of deep AI-network integration as the foundation, with a commercially scalable business model still taking shape; Layer 2 lies in building a measurable, billable, and deliverable Token business loop — rather than merely expanding compute infrastructure.

ZTE draws a distinction between "AI for RAN" and "RAN for AI." Can you unpack that, and where is the real near-term value today versus what's still roadmap?

From a compute-supply perspective, "AI for RAN" and "RAN for AI" should be decoupled when evaluating AI-RAN integration: "AI for RAN" leverages AI to enhance RAN efficiency and performance, bringing universal value to all sites, which operators are eager to deploy in live networks today; "RAN for AI" utilizes RAN compute for general-purpose AI applications, where the business model remains immature and demand is limited to a few niche scenarios. Therefore, a decoupled, heterogeneous architecture is the optimal path for ROI.

The real near-term value for telcos is reducing OPEX and improving revenue, which is primarily achieved through AI for RAN. ZTE has conducted extensive exploration and practice in this field: on spectrum efficiency, AI-powered M-MIMO boosts cell capacity by 15–20%; on energy efficiency, AI delivers an additional 10–15% power savings; on O&M efficiency, MTTR is reduced by 20%. Meanwhile, by leveraging AI to deliver tiered and differentiated experiences across diverse scenarios and services, operators can upgrade their business paradigm from "selling traffic" to "selling experience," driving ARPU growth of 5–20%. What's still on the roadmap includes renting out computing power for third-party inference, which is still seeking a closed business loop.

The industry's talking about Level 4/5 autonomous networks that self-heal and self-evolve. What does "minimal human intervention" really look like in a live network today, and where do you draw the line on what an agent is allowed to decide on its own?

In mobile communication networks, equipment vendors and operators are working together to advance toward L4/L5 autonomous networks. Progress has already been made in the dimensions of O&M efficiency and energy efficiency, with full-autonomy closed-loop operations achieved in certain scenarios.

Specifically:

- In fault management, agents are integrated with ticketing systems to automatically analyze and identify root causes, providing handling recommendations. Routine alarms can be remotely repaired automatically, while complex faults are escalated to human engineers for resolution.

- In network optimization, built-in base station intelligence, combined with network optimization large models, enables automatic identification of poor-quality call records and root cause localization. Multi-dimensional network digital twin models then generate optimization plans. For certain mature scenarios, solutions can be executed in a closed loop automatically, while others require human review before implementation.

- In energy efficiency, fully automated closed-loop operations are already realized. Managers only need to set policy objectives; the system automatically analyzes and predicts network load, and dynamically formulates and deploys energy-saving configurations for each cell based on KPIs — requiring no manual intervention throughout the process.

The boundary for agent autonomous decision-making is governed by a dual verification framework: "accuracy meets threshold" and "impact is controllable." Agents are authorized to execute closed-loop actions when the problem domain is well-structured, solution accuracy reaches or exceeds 90%, and the operational impact is predictable and reversible. For example: for intelligent energy saving, adjusting parameters based on balanced load and KPIs already achieves high accuracy, and since the basic coverage layers remain active, performance is safeguarded. For remote reboot of out-of-service cells, the operation carries low risk and will not trigger more negative impacts.

Operators are weighing whether to become AI infrastructure and compute providers themselves. Where do telcos have a real structural advantage there, and what do operators most often underestimate about what it takes?

Telcos may enjoy more unique advantages than some realize, for good reasons. The AI industry, like many groundbreaking industries before it, does not transform our lives and business overnight; instead it focuses first on its "engine" and a very limited range of applications and then on a much broader range of applications. In other words, the adoption and diffusion of AI are just getting started.

So the first and foremost advantage of telcos is connectivity that enables AI cloud and AI agents to get connected to people and things and to one another, and telcos can and should do so with AI-awareness so that certain types of AI traffic can get their SLA guaranteed.

The second advantage of telcos is that they are already in a unique position to support AI infrastructure deployment and development with their data centers, access to the power grid and expertise in this area.

The third advantage, and an increasingly important one, is that telcos sit right at the literal fingertips of billions of users who use AI so they are the first to facilitate the flow of a massive scale of valuable data. It is of utmost importance for telcos to safeguard data, privacy and network security, which only telcos can help address from where it begins.

Where is telco-AI adoption moving fastest right now, and what have live deployments taught you that the lab didn't?

The areas where AI has been most rapidly adopted in the telecom industry are energy saving, fault management, and network optimization. These scenarios have a common characteristic: traditional manual methods either cannot achieve precise control (such as second-level load forecasting or cell-level energy-saving scheduling), or require significant manpower and resources (such as root cause analysis of massive alarms). AI delivers immediate and measurable benefits in these areas, making them the first to see large-scale deployment across the industry.

Lessons from real-world network deployment:

- Data is the bottleneck. The quantity and accuracy of data directly determine how well AI performs. AI needs access to more data sources to reach its full potential. This includes data collected more frequently (like every second) and data that goes deeper into details (such as root cause information or data at the service/call record level). Missing data directly limits what AI can do. Furthermore, technologies like network digital twins and AI-driven optimization rely heavily on reliable basic data (for example, accurate neighbor cell relationships or up-to-date configuration baselines). Inaccurate data can lead to systemic risks in the live network. Therefore, ensuring the accuracy of basic data must become a strict, everyday operational standard.

- AI is redefining what "good" means. AI helps us do old tasks much more precisely and improves our old practices. For example, in the past, we often ignored small, specific issues with network coverage. Now, with AI, we can combine fine-tuning at these tiny spots (micro-optimization) with overall network improvements (macro-optimization). This achieves both precise results in specific areas and the best possible performance for the entire network.

- Technology deployment does not equal value creation, and business loop is the final gap. Take tiering and differentiated experience as an example: this is a key chance to redefine premium pricing in the 5G era. Operators understand this, but progress is much slower than in areas like saving energy. Why? Because it's not enough to just have the technology ready. It requires strong teamwork between network, marketing, and customer service teams. Most importantly, customers must clearly feel the difference — they need to see, measure, and value the better experience — so they are willing to pay for it. Only then can AI capabilities be truly transformed into sustainable business revenue and drive a shift in operating business models.

Looking to 6G and "native AI," what changes when AI is designed into the air interface from the start rather than bolted on?

The AI-native air interface marks a fundamental departure from conventional rule-based radio design, ushering in data-driven intelligence. In 5G, AI is deployed in an overlay fashion — it serves as an auxiliary optimization module without altering the core air-interface logic. In contrast, the 6G native-AI air interface embeds artificial intelligence into the base station architecture from the very outset of design, holistically integrating protocol signaling, algorithm stacks, hardware interconnects, and computational resource allocation into a unified framework. Leveraging AI's ability to learn from massive real-world network data, the system autonomously discerns patterns within complex environments and makes superior decisions. Anchored in this end-to-end native intelligence framework, AI deeply empowers critical real-time air-interface procedures, delivering substantially greater wireless performance gains compared to the overlay-AI approach of 5G.

6

Scaling autonomous operations – from AIOps to agentic networks

The forum's largest panel took Philippe Ensarguet's maturity path into the operational trenches. It convened David Kypuros, global principal solutions architect at Red Hat; Wasim Azhar, head of solution innovation and business growth at Amdocs; David Palacios, telco product director at Tupl; Rana El Desouky Kazamel, senior director of product management at Cisco; and Alexis Koalla, director of operations strategy and autonomous network transformation at Orange.

It opened where every conversation at the forum seemed to start, with data readiness, and produced a remarkably consistent answer. Don't try to fix everything at once. Palacios said trying to prepare all data in advance is one reason programs stall; target narrow problems where data can be made clean, accurate and real-time, then grow. Koalla described Orange's discipline of starting from the business case and working back to "the right data from the right source to serve the right business at the right time." Kazamel added the funding reality. Know your North Star architecture so today's choices don't have to be redone, but get to impactful use cases fast, because "if you don't show outcomes fast enough, you might not be even able to progress and continue funding it." Azhar said the tier-one CSPs pursuing broad unification are building a common data fabric — "the automation oil," as he put it — while others attack specific data gaps directly. When a fault appears fewer than five times across months of alarm data, operators are synthesizing realistic data to fill the gap, and using chaos engineering to simulate rare double-

fault scenarios, like a fiber cut coinciding with a core software crash, so agents learn remediations for events the network has barely seen.

Agents in the incident chain

The panel's vendors are all building agentic operations offerings, and the proof points are getting specific. Amdocs positions aOS, its agentic operating system, as a "cognitive core" above any vendor's BSS and OSS, with a pre-built agent library spanning sales, billing and care through to network planning and service delivery. The metric operators start with, Azhar said, is not mean time to resolve but mean time to detect, cutting alarm noise and the lag between alarm and trouble ticket. At least three CSPs Amdocs works with have seen detection-time reductions of 65 to 85%. Kypuros described a Red Hat project with Ericsson and Intel on one of the RAN's nastiest time sinks, PTP synchronization troubleshooting, which typically drags multiple companies into a room, each staring at its own data. Agents outfitted with skills did the fetching across all of them — MCP running across corporate boundaries rather than inside a single company, and a first taste of what disaggregated multi-vendor operations will demand. Cisco, meanwhile, used Cisco Live to launch Crosswork AI, a multi-agent framework with out-of-the-box agents and room for operators to build their own. Kazamel framed the destination as "intelligent operational teamwork at scale," where agents "have identities like humans… and they get performance reviews like humans." Cisco's AI impact report projects agents generating roughly 450% more traffic than humans performing the same tasks, a load she argued only agentic operations can absorb. Koalla's response captured the operator mood. Add accountability to the list alongside explainability, traceability and trust, and mind the timeline. "If it is next year… I'm not confident. But if it is three, five years — yeah, definitely we are stepping in."

Who owns the failure?

Asked who is responsible when a production agent degrades service, Azhar was clear: "The operators are accountable." His prescription is to make that accountability workable through confidence thresholds, where an agent takes real autonomous action only with something like 99% confidence in both the root cause and the remediation path and escalates anything less, plus guardrails, human-in-the-loop models and agents that learn from outcomes. Kypuros sketched the human's seat in that loop, with data to troubleshoot on one side and, on the other, the remediation environment — a "kind of danger zone" where humans supply the intuition an agent lacks about blast radius, like impact on neighboring cells. Standards bodies, he argued, "are our friends" here, providing hard boundaries through TM Forum intent models and MCP gateways, and the industry's agentic harness "should have a lot of reasons to tell agents no." Azhar flagged the gap. TM Forum defines APIs and data models, but no real telecom standard yet exists for agent-to-agent collaboration, work Amdocs is now pursuing with partners and CSPs.

Probabilistic brains, deterministic hands

The panel's conceptual heart was the tension between deterministic, rules-based telecom systems and probabilistic AI. Nobody argued for retreat. "It would be quite a loss for the telco industry if we don't leverage the AI probabilistic models," Kazamel said. Use them freely for insight, correlation and deep troubleshooting, and reserve determinism, or human approval, for actions that touch the network. Palacios described Tupl's deliberately conservative variant, which uses generative AI at build time to construct fully deterministic automation pipelines that then run the network. The result is more traceable and more palatable to risk-averse operators, with runtime AI to follow as governance matures. Kypuros offered a complementary pattern. Let LLMs bring their full stochastic intelligence to analysis, but keep remediation "non-LLM load-bearing," executed through discrete, pre-approved

playbooks a human initiates. Azhar predicted convergence through mutual adaptation, with AI constrained by guardrails and OSS accepting probabilistic decisions mediated by policy layers. He also tempered digital twin expectations. A true twin of the entire network isn't coming, but scenario-specific twins, fed by the right data and sharpened through chaos engineering, will become accurate enough for agents to trust.

The closing lightning round revealed the real bottleneck, and it wasn't technical. Koalla named organizational willingness, admitting "it's new, it's a little bit scary," and a culture change Orange is addressing through training and coaching. Palacios' diagnosis came down to one word — trust. Kazamel pointed to upskilling teams to co-develop agents rather than merely govern them. And Azhar went a step earlier than trust, to acceptance. Autonomous agents will redraw the boundary between operations and engineering itself. A panel that began with data readiness ended somewhere more uncomfortable — the technology, the panel broadly agreed, is maturing faster than the organizations meant to run it.

7

AI-RAN – embedding intelligence into the RAN

Where Ericsson's presentation made the case for AI models inside today's basebands, this fireside zoomed out to the full AI-RAN arc — and put dates on it. Shujaur Mufti, director of telco ecosystem solution architecture at Red Hat, worked from the taxonomy that has come to organize the space. AI for RAN uses AI to improve the radio network itself. AI-and-RAN runs AI and RAN workloads on common infrastructure. And AI-on-RAN turns the RAN into a platform carrying AI traffic.

The sequencing, in his telling, is not in dispute. AI for RAN comes first because the value is immediate and visible in energy savings, network efficiency, spectral efficiency and fault detection, and because "it applies on the traditional radio network as well. You don't have to modernize your RAN network." SON systems managing traditional radio networks are being enhanced with AI nativeness, and the SMO layer born in Open RAN is developing rApps and xApps while starting out managing traditional RAN. AI-and-RAN comes next. SoftBank, a Red Hat partner of several years, started that journey early, and T-Mobile has recently gone public with POCs and trials. AI-on-RAN comes last, unlocking new monetization as AI-bound traffic from consumer and enterprise use cases passes through the network — "another bearer," as Mufti put it, flowing through the radio network.

Not GPUs everywhere

Asked when GPU-accelerated compute actually lands in the RAN footprint, particularly with the 6G era bringing upper mid-band spectrum, higher-density massive MIMO arrays and compute-hungry use cases like integrated

sensing, Mufti started with a corrective: "We should not think GPU everywhere in the RAN." The transition will be gradual and distributed, likely starting with distributed inferencing at the edge in an AI-first approach, because the RAN "is a premium workload and… a premium asset for any mobile operator."

Red Hat's early work with SoftBank, with Fujitsu among the initial partners, produced a finding that still shapes the engineering conversation. When the RAN application runs at L1 and L2 on the GPU, "you don't need the real-time kernel" — the subject of a Red Hat publication comparing CPU- and GPU-based RAN three years ago. The characteristics that favor GPU compute map neatly onto 6G's demands for terahertz-range spectrum, extreme massive MIMO and sub-millisecond latencies. His timeline runs in three phases. Through roughly 2027, AI for RAN dominates, with more rApps and xApps launching and, in a detail worth watching, early vendor moves toward subscription-based models for RAN optimization. From 2027 to 2030, as 3GPP incubates initial 6G standards, AI-and-RAN moves through POCs and trials on the hard problems of common infrastructure, from resource utilization to running both workloads simultaneously with security, isolation and segmentation, to managing separate lifecycles for AI and RAN workloads on shared hardware. From 2030 onward, with 6G standardized, AI-on-RAN arrives. With autonomous operations maturing in the core and OSS in parallel, the operator network becomes AI-native end to end.

From AI factories to the AI grid

The integration groundwork is further along than the timeline might suggest. Red Hat has worked with NVIDIA for more than three years, publishing jointly at MWC 2023 on cloud RAN deployed on NVIDIA's Aerial SDK for Layer 1 — again without a real-time kernel, where today's CPU-based cloud RAN and Open RAN deployments still demand careful CPU optimization. The SoftBank proof-of-concept delivered the benefits the operator had envisioned and added an orchestrator, integrated with Red Hat's platform, to schedule RAN and AI workloads together. The ecosystem has since widened considerably. "Every RAN vendor today is looking into… how to enable GPUs in the RAN application," Mufti said, pointing to the publicly announced three-way alignment of NVIDIA's investment in Nokia, with Red Hat as Nokia's platform. Beyond PHY- and MAC-layer enablement, the partners have launched an AI factory and are now extending the AI-RAN concept into what Red Hat calls the AI grid: "RAN-ready AI infrastructure at the edge." For some operators, he noted, that inverts the deployment logic — start with the AI workload at the edge, prove the inferencing revenue, and add the RAN application when it makes sense.

Turning capex into a monetization machine

The business case discipline follows from the RAN's status. It absorbs most of an operator's capex and determines network quality, so AI-RAN "would only make sense if it has technology and economic benefits" — better network quality, or new revenue that lets operators "transfer their capex into a monetization machine." The AI grid model is the hedge. Begin with selected sites, monetize edge inferencing first, then decide how much GPU capability to carve out between RAN and AI workloads, expanding only as the economics prove out. The unifying requirement, in Red Hat's vision, is a common cloud-native platform and a common AI fabric stretching from the data center and core through the edge and far edge to the radio network — one operational envelope, one way of managing agentic and model frameworks alike.

That framing shaped his parting advice, which could stand for the forum as a whole: "Don't think implementing AI-RAN as a silo." AI is already being deployed at the core, in OSS and BSS, and in autonomous operations; the task is to carry those lessons into RAN design, sizing and lifecycle management rather than starting over. The RAN may be the premium asset, but on this account it is the last domain AI reaches — and the one where the economics have to be proven, not assumed. That proof lives in operators' networks, which is where this report turns next.

8

Evolving RAN with AI – CSP lessons on deployment, scale, value

Ericsson's AI-RAN case was put to the test here by two of the operators actually running the features. Moderating for Ericsson, Melike Erol-Kantarci framed the vendor's newly announced AI in RAN product, with its telco-grade models, continuous learning and agentic AI support, as the output of a long joint learning journey spanning more than 15 trials of RAN AI features around the globe, with North American CSPs among the first to test. Her summary of what those years taught Ericsson could serve as a thesis for the whole forum: "It's never about having more and more data, but it's about having the high quality data." From there come specialized models capable of real-time inference, then observed actions, measurable outcomes and only then scale.

The operators came with results. Adam Loddeke, AVP of RAN technology at AT&T, said the carrier already uses AI in the RAN at scale — predominantly above the network through SON applications and rApps, spanning day-zero planning and forecasting, day-one deployment and configuration, and day-two optimization and lifecycle management. The newer shift is AI-native features embedded in the RAN stack itself, and AT&T has been trialing Ericsson's. Results for the AI-native link adaptation feature have been "very impressive to date. We've seen significant improvements in both throughput and spectral efficiency." Cindy Dong, associate director of RAN planning at Verizon, described a "dual RAN AI strategy" that balances centralized, non-real-time RF optimization with local AI integrated in the gNodeB, running on what she called a world-leading, fully virtualized RAN that gives Verizon an AI-ready head start. Verizon has published RAN AI guidance to all its vendors in recent months, she

said, has an IEEE ComSoc paper on the way, and proved in a live Ericsson trial late last year that AI link adaptation improves downlink spectral efficiency, with solid coverage gains in further field trials this year. Ericsson's own trial numbers put the link adaptation feature at up to 20% downlink throughput improvement and up to 10% better spectral efficiency.

The AI worthiness test

The panel's most useful contribution to the industry's vocabulary came from Dong. Verizon's gatekeeper for any RAN AI feature is what she calls AI worthiness. Any performance or user-experience gain must explicitly outweigh the added cost and complexity of the extra compute it demands and, more pointedly, must beat the best classical rule-based baseline vendors have already built. "We don't want to just deploy and scale the AI just for sake of doing AI," she said. The published guidance enforces safeguards and lifecycle discipline to match, including the ability to switch from an AI function back to the rule-based function if the model detects drift. Loddeke's framing was different in vocabulary and identical in substance. It comes down to cost-benefit analysis, and "not every application that needs to reside in the network needs to be agentic in nature, because there's a cost associated with doing so." AI in the RAN, he argued, is really the business of constantly optimizing competing constraints — capacity versus performance, peak performance versus energy savings, speed versus stability.

Where the models live

Both operators see the frontier moving from centralized optimization down into the gNodeB — for Verizon, touching Layers 1 and 2 in real time — which is precisely where the trade-offs sharpen. Loddeke walked through the questions his teams weigh. What is the performance benefit versus the compute resources a model consumes? How accurate is it, and what update frequency does it need to stay accurate? Should it be generalized or specialized, given that a model that excels in dense urban environments may perform poorly in rural ones, and what works for one operator's network may not transfer to another built differently? The operational side worries him at least as much as the engineering — is it one model at every site or a handful across the network, and how does an operations team troubleshoot a network with embedded models in place while managing their lifecycle? Erol-Kantarci pointed to Ericsson's answers, with a TCO-optimized architecture designed so features run on readily available, already-installed hardware, and algorithms that increasingly transfer learnings, so a corner case optimized in one network can improve another. On use cases, Loddeke put today's value squarely in AI-driven optimization and automation. Link adaptation, beamforming, channel estimation, interference optimization and traffic steering sit on the performance side, with fault detection, root cause analysis toward closed-loop remediation and the industry's long-running energy efficiency push on the cost side. New AI-driven services, he said, are a strategic wave still to come.

From blueprint to production

Asked to pick the single capability that moves AI-RAN from pilots to production, whether data, automation or architecture, the two operators split in a revealing way. Loddeke declined to separate them but admitted where his team's time goes. The answer is architecture, meaning an open, programmable platform that enables ecosystem innovation "so that we're not just dependent on any one person for developing AI in our network," with third parties, Ericsson and AT&T's own automation all able to build against it. Dong held her ground. Architecture is fundamental, but identifying the exact problem worth solving with AI matters more, because it determines the investment; the target is corner cases today's rule-based layers struggle to address, delivering extra gain at

minimum cost. Their 12-to-18-month priorities followed suit. Verizon will keep its long-term planning momentum while aggressively executing the handful of AI areas already identified with its vendors, every deployment passing the AI worthiness gate with controlled cost. AT&T will scale new use cases at the gNodeB while building an architecture ready for "the things that we can't predict" over the next two to three years.

Loddeke's summary line marked how far the conversation has moved: "We've both proven that AI works in the RAN space." The question at the cell site is no longer whether — it is where the models run, what they cost, and whether they clear a bar the rules-based network spent three decades setting.

Telco AI infrastructure – foundation for scalable AI

The infrastructure panel opened with a question the industry has been burned by before. Is running AI across distributed telco sites actually new, or a rerun of the telco edge and MEC conversations of the last decade? The consensus from Tom Davis, principal systems engineer at Everpure, the company formerly known as Pure Storage; Alexandre Guilbault, vice president of artificial intelligence at TELUS; Rob Hughes, head of wireless marketing at 1Finity; and Jake Katz, vice president of AI products and solutions at Cisco, was evolution, but with the stakes transformed.

Katz put it in one word — stressors. The mechanisms aren't new, but "the stressors are much higher" on the network, on GPU and CPU infrastructure, and on security and data privacy. Hughes drew the sharpest contrast with mobile edge computing. MEC was a build-it-and-they-will-come bet on developers who never quite came, whereas "with AI it's coming whether you're ready or not." And unlike MEC's idle capacity problem, an operator deploying AI-RAN can be its own anchor tenant, running the RAN on the infrastructure and selling the excess. Guilbault added the geopolitical layer. AI is becoming as critical as electricity, something no country wants fully dependent on another. "We need to own this future, we need to control end to end our workload in AI — and who else than telco is better positioned to do that?" For TELUS, that means edge AI for network use cases plus a centralized, sovereign "token factory" serving defense, health and government workloads.

LLMs in the core, expert models at the edge

On the centralized-versus-distributed question, the panel converged on a hybrid, but one with unusually clear

sorting logic. Guilbault split AI by model type rather than treating it as one thing. Forecasting, optimization and lighter deep learning models belong at the edge, where acting fast on network data matters and compute needs are modest. But LLMs doing high-level reasoning already carry seconds of thinking time, so shaving transport latency at the edge buys little: "I don't see a future where we would necessarily need LLMs on the edge." Davis made the latency case concrete from the other direction. A voice AI agent served from a central site suffers a two-second round trip, a pause anyone who has endured one in conversation knows too well, while the same capability served by a small language model at the edge comes back in roughly 400 milliseconds and "feels much more natural." Expert models for voice-to-text or computer vision, latency- and jitter-sensitive or generating immense backhaul, are the edge's natural tenants. Physical AI splits the work, triaging in real time at the edge and sending data back for deeper analysis. Katz added robotics, autonomous systems and clinical use cases to the decentralized column while noting industry agreement that foundational model training stays central. He also flagged an infrastructure subtlety operators should watch. Today's AI builds are split between a traditional front-end network and a costly, power-hungry back-end scale-out fabric for multi-rack GPU training, but inference often needs no scale-out fabric at all, running on a single GPU or a scale-up server. That means edge inference "is going to allow you to do more combinations than we actually see today." Hughes's advice was correspondingly incremental. Don't put GPUs at every site; centralize in enterprise-dense pockets, and over time use application awareness to move workloads wherever latency, cost, sovereignty or spare capacity dictates.

Fail in half a day

Guilbault brought the panel's most sobering numbers, drawn from 20 years of AI adoption at TELUS. Historically, he noted, only 20% of data science projects reached production, and the agentic era has made it worse, with studies he cited putting the failure rate for agentic projects at 95%. TELUS's response is not to chase the success rate but to minimize the cost of failure — test the weakest link first and pull the plug fast. "Even if we fail 95% of the time, if it only takes half a day to fail," the sunk costs stay small and the winners get the investment. He paired it with a statistic he said the consultancies' State of AI reports agree on, that roughly 70% of project success is people and process, not technology. "If you build the best AI in the world, but nobody listens to it, nobody changes their behavior, no process changes, it won't bring any value." Davis mapped where the failures actually occur. Customers buy infrastructure first, choose a model second, and only then discover the real problems. Data is siloed, dirty and ungoverned. Data a model can reach may still lack business meaning and policy context, producing answers that are less useful and riskier. And pilots that work as one-offs resist operationalization at scale. On the trust side, Guilbault described TELUS's two-directional approach, with an AI board including the CEO and leadership governing safe adoption from the top, and the latest tools placed in the hands of all of its employees — 150,000 worldwide, by his count — in a safe environment from the bottom — a hands-on "idea tank" whose best discoveries his team hardens into enterprise-grade production.

Sell what you already have

The monetization discussion kept returning to the same principle of extending existing strengths rather than speculating on new ones. TELUS already sells GPU-as-a-service, priced per GPU per hour, with demand strongest from teams training and fine-tuning models and from universities. But most of its Canadian SME customers want agents and RAG applications, not hardware, which is what the token factory abstracts away behind an API, with token-based pricing to follow. Hughes sketched the enterprise ladder operators should offer. GPUaaS suits sophisticated customers, AI-as-a-service over the top serves enterprises that have applications but no appetite for managing infrastructure, and an application development platform with customizable starter apps helps those who need the most support. Katz framed the whole opportunity as an extension of managed services operators already run, hosting data today and GPUs tomorrow, backed by the two structural advantages of security and

the deployment muscle enterprises lack for standing up GPUs and the networks that feed them. Davis closed by crediting a colleague, Everpure's Shawn Rosemarin, with the analogy that summed the panel up: buying the GPUs first "is akin to buying the machinery for your factory before deciding what you were going to build in said factory." Scalable telco AI, Davis argued, won't be won by whoever buys the most hardware, but by operators that build the right data foundation, give models the right context and simplify the path from pilot to production.

Guilbault got the last word, and it pointed straight at the forum's finale: "Context is king." An AI may someday reason at Einstein level, but like a brilliant new employee without knowledge of the company, it can do little without context — so document everything, build the data infrastructure, and accept it will never be perfect. "Don't wait for it to get started. Just get started now."

Towards 6G AI-native architecture

The forum saved its most fundamental question for last. What does AI-native actually mean? Dr. Yue Wang, chief technologist at China Telecom, answered with a definition before touching the architecture. An AI-native network, she said, has three layers. In the infrastructure layer, network, compute, storage and AI resources are planned and deployed. In the operational layer, those resources are optimized and orchestrated. And in the service layer, the network supports AI services — services that will ask for connectivity, compute and latency adaptation at the same time.

Hold that picture up against today's networks and the architectural limitation is obvious. Current systems were built around deterministic procedures, predefined interfaces and human-engineered control logic — "not originally designed around AI." The data is not AI-ready, the APIs were not always designed for closed-loop control, and network functions were not built with AI lifecycle management or real-time decision-making in mind. The industry's problem, in her diagnosis, "is not a lack of AI algorithms put into a software system, but rather the ability to orchestrate the network and the compute together to support the AI services."

A hybrid convergence

Asked whether the industry should adapt AI to telecom's ways or rethink telecom to accommodate AI, Wang refused the either/or: "It has to go both ways." AI must bend to telecom first, because the network is critical infrastructure where uncontrollable decisions cannot be tolerated, and Wang said this is already being addressed

with policy engines that validate AI-driven actions before they can breach SLA boundaries or threaten service continuity. But telecom must bend too, because the future network is "too complex and too dynamic" for every decision to be manually specified in advance. Her formulation balanced the two obligations precisely: "We should make AI reliable enough for telecom, but at the same time we need to allow telecom architecture to be flexible enough to benefit from AI."

The impediments to fully autonomous operations, she argued, are both technological and organizational, and China Telecom's own history shows how far ahead of the AI wave the organizational work has to start. The operator has pursued cloud-network convergence since 2019, on the logic that "you can't have two separate systems" each getting optimization and AI bolted on independently; they must be unified. Crucially, that meant more than integrating technical components. China Telecom fully merged its network and cloud operations teams, and that single organizational decision, made years before agentic AI existed, is what now lets the operator insert AI and upgrade network and cloud capabilities as one system rather than two.

Deciding where intelligence runs

For the service layer, Wang borrowed the AI industry's own roadmap, showcased at CES. 2026 is the year of agentic AI, with agents collaborating to solve complex tasks, and 2027 brings the emergence of physical AI, from robotic dogs to factory robot arms, in applications demanding very low latency and compute together rather than connectivity alone. From that trajectory she derived the first capability a truly AI-native 6G network must have — joint orchestration of communication and compute. Today's 5G network is still mainly designed to transport data. An AI-native 6G network "should be able to decide where I run this intelligence," whether in the far cloud, at the edge to save latency, or spread across both, based on latency, cost, energy and the service's requirements.

Her five-year milestone followed the same logic. The industry will know it is on the right path, she said, when it has delivered a flexible capability platform — a network that "can seamlessly expose and orchestrate" communication, compute, data, intelligence and assurance as capabilities offered to customers, AI models and devices alike. And because those AI models and devices will keep changing, "the network fabric must be able to adapt right along with them." The network, in short, must expose its capabilities and evolve in step with AI's.

It made a fitting bookend for the forum. The opening panel ended with Mérouane Debbah forecasting a network whose primary customers are hundreds of billions of agents; the closing interview described the architecture that would have to serve them — one that no longer treats AI as a workload riding on the network, but as the thing the network is built around.

Conclusion

The path forward is clear, but not simple. Telco AI will scale on commercial foundation models, embedded telco-grade models and agentic operations, but those technologies have to be adapted, benchmarked, governed and proven for networks where failure carries real operational and commercial consequences. Telecom-specific models, machine-ready data, embedded RAN intelligence, staged autonomy, outcome-based measurement and AI-native architecture are interdependent parts of a larger operational system, one that must earn trust stage by stage and prove value all the way to the CFO.

The next phase requires disciplined execution built on narrow, high-value use cases, honest gatekeeping of what deserves AI at all, instrumentation designed in from day one and deeper collaboration on shared problems, benchmarks and synthetic data. No single operator, vendor, standards body or research lab can solve this alone. Advantage will come from building together, measuring together and industrializing AI into networks ready to serve customers that barely exist yet — hundreds of billions of agents, and the intelligence economy they bring.

Acknowledgements

Amdocs

We help those who build the future to make it amazing. In an era where new technologies are born every minute, and the demand for meaningful digital experiences has never been so intense, we unlock our customers' innovative potential, empowering them to transform their boldest ideas into reality, and make billions of people feel like VIPs. Our approximately 30,000 employees around the globe are here to accelerate our customers' migration to the cloud, differentiate in the 5G era, digitalize and automate their operations, and provide end users with the next-generation communication and media experiences that make the world say wow.

Learn more

Cisco

Cisco is a worldwide technology leader revolutionizing how organizations connect and protect in the artificial intelligence (AI) era. Through its industry-leading AI-powered solutions and services, the company enables customers, partners, and communities to unlock innovation, enhance productivity, and strengthen digital resilience.

Learn more

Emerson

Emerson provides an advanced automation portfolio for the world's most essential industries, delivering technologies and driving innovation to help companies optimize their operations, safety and sustainability.

Learn more

Red Hat

Red Hat is the world's leading provider of enterprise open source software solutions and services, using a community-powered approach to deliver reliable and high-performing Linux, hybrid cloud, container, and Kubernetes technologies. Red Hat helps customers develop cloud-native applications, integrate existing and new IT, and automate, secure, and manage complex environments.

Learn more

ZTE

ZTE Corporation is a global provider of information and communications technology (ICT) solutions, serving telecom carriers, enterprises and consumers across more than 160 countries. Founded in 1985 and headquartered in Shenzhen, China, the company operates across three core business areas: carrier networks, government and enterprise, and consumer devices. Its portfolio spans 5G wireless and wireline infrastructure, core networks, optical transport, cloud and data centre technologies, chipsets and a broad range of smart terminals.

Learn more

MARKET PULSE REPORT

TELCO AI

RCR Tech
NEWS. INSIGHT. IMPACT.