Ericsson White Paper
GFTL-26:000788 Uen
June 2026

ERICSSON

# Trustworthy AI for Telecom Systems

# Content

# Abstract

Artificial intelligence (AI) is becoming integral to next-generation telecom systems, but it brings risks. The recent AI advancements in large language models (LLMs) and agentic AI introduce new dimensions to that risk. To trust AI-enabled systems, we must be able to trust AI itself and comply with regulations such as the European Union (EU) AI Act. Trust means ensuring the system works as intended and does no harm.

The key message of the paper is that for AI to be integrated into the telecom domain, including 5G and 6G networks, it must move beyond mere performance metrics to a holistic framework of trustworthiness. These capabilities should be embedded by design in the telecom domain. Trustworthiness is a core requirement for adopting AI-based systems, especially in live networks.

# Introduction

As networks evolve into AI-native architectures, AI moves from a recommended trait to a foundational requirement, embedding trust at the core. In 5G, 6G, and autonomous network management, trustworthiness encompasses safety, security, transparency, reliability, and ethics. This integration of AI introduces a complex paradox: while AI is essential for managing the scale and complexity of modern traffic, its black box nature and susceptibility to adversarial manipulation pose risks to safety, transparency, and security. If not addressed, these risks might affect its adoption, delaying its positive benefits.

At Ericsson, we build trust into our AI portfolio by design. Our solutions respect human values and aim for outcomes that are both optimal and sustainable.

This vision aligns with the EU AI ethics guidelines [1], which Ericsson adopted in 2019 and which set out requirements for human agency and oversight, transparency, privacy and data governance, diversity, non-discrimination and fairness, technical robustness and safety, societal and environmental wellbeing, and accountability.

Our previous whitepaper on trustworthy AI outlined this strategy and presented how to address trustworthiness requirements in AI-enabled systems. In Europe, the regulatory process culminated in 2024 with the EU AI Act [2], which defines rules for both providers and deployers of AI and applies a risk-based approach with stricter obligations for higher-risk systems. Provisions of the act are now coming into force, and supporting materials such as guidelines, codes of practice, and harmonized standards are also emerging.

For creators and users of AI in telecom, trustworthiness is now a must, for compliance and for market confidence. It also opens space for innovation in a trustful manner. Operators and subscribers expect it, and competitors will offer it.

In addition, the recent explosion in the use of LLMs, generative AI, and agentic systems brings a new era for AI, which puts a new spin on trustworthiness. AI now interprets and generates natural language, reasons, operates autonomously, and can act in the real world. Thus, AI trustworthiness as a moving target calls for new techniques. The ethics guidelines still apply but their interpretation must evolve to fit the nuances of modern AI.

For Ericsson, this means ensuring that our current and future networks—5G, 6G, and beyond—become more autonomous, resilient, predictable, aligned with the regulations, and worthy of our customers' trust.

# Trustworthiness for traditional/ generic AI/ML

As networks progress towards higher levels of autonomy, AI/ machine learning (ML) functions are being integrated across mobile network architecture, from lower-level layers to high-level management [3]. It becomes essential to ensure these AI entities can be trusted.

Here, "traditional" refers to AI models that require structured data with a singular modality and rely heavily on feature engineering. These systems are often thought of as ML rather than AI. Trustworthiness requirements are addressed through the development and application of technical methods specific to telecommunication use cases.

## Transparency/explainable AI

Transparency in AI-based systems provides interpretable, relevant, and understandable reasoning for why a decision was made and/or how AI arrived at this decision, using explainable AI (XAI) techniques. This builds customer confidence and trust and supports the adoption of black-box AI models executing in live networks.

XAI gives developers and testers a window into the model's internal logic and helps verify correctness. These techniques help engineers better understand the AI model and optimize it with efficient feature engineering and data collection.

Explainability can be applied to ensure correctness and enable trust at the model, component, or system level. At the model or component level, it provides model-specific insights in terms of input features, model parameters, or internal logic. This is useful for validating the black-box models and improving them by identifying corner cases.
It can also be applied to discover and describe unusual or new behavior exhibited by the model. Insights from model-level explainability can be used for developing autonomous networks, to perform root-cause analysis, and to proactively catch issues before they appear in live systems.

System-level explainability makes sure the entire system behaves as expected. It enhances traceability for the entire flow from inputs, through the model and other components,
to outcomes delivered to the user. These include dynamic, model-agnostic explainability techniques such as decision flow trees or graphs, system traces, and case-based summaries. Surrogate model techniques can also be applied to explain other models. System-level explainability also helps automated AI and ML lifecycle management through compositional, generalized explanations.

Both external users—customers and non-expert stakeholders—and internal users—developers, testers, and monitors—can all benefit from explainability insights provided at the model or system level.

Please see Section 4 for detailed examples.

## Data and privacy

ML models are usually derived from large quantities of training data, so the quality of the data directly affects the quality of the model. Corrupted training data can result in incorrect or biased model results, and corruption of query data can taint the results of specific queries. Model developers need to analyze data for correctness, bias, and representation of the whole input domain. The statistical properties of the training data should match those of the real world. Tracking data provenance is important, to ensure data has not been tampered with between its source and the training pipeline. Where possible, training data should be retained to allow forensic analysis in the event of incidents and to better support retraining and model updates.

Depending on the domain, the data may contain sensitive content such as personally identifiable information (PII), so de-identification may be required. Even where data has been de-identified, attacks may be possible if a model itself is used to correlate with external data sources and infer aspects of members of the training set. Privacy principles by design and default should be applied to ensure security throughout the model lifecycle.

## Robustness and security

A robust and secure telecom network requires that the AI on which it depends is also robust and secure. Attacks shown to be possible against traditional ML include:

- data poisoning, where the attacker can influence the overall model to introduce bias or can change model outputs for specific input cases
- adversarial examples, where the attacker makes minor, often undetectable changes to query data to change the output
- extraction attacks, where query data is collected to steal or copy the model weights
- inversion attacks, where queries are used to learn aspects of the training data, possibly by obtaining a significant subset of it
- membership inference attacks—that can be used to subvert de-identification controls—where the attacker attempts to determine if certain data points exist in the training set

Since these attacks require differing levels of access to the training data, query data, and query interface, applying security controls throughout the model lifecycle is important. These controls must work together with quality assurance mechanisms—robustness and security are two sides of the same coin. For example, adversarial training, in which we intentionally include malicious examples in the training data so the model can learn to resist them, improves both security and robustness. Automated quality assurance is essential, using metrics selected to meet the joint needs of output quality and security. Conducting invariance and directional expectation tests is also essential to assess and assure the model's robustness. In an invariance test, model predictions are expected to remain unchanged when label-preserving perturbations are applied to inputs. In a directional expectation test, a set of perturbations is defined for the input, which should have a predictable effect on model output. In production, guardrails can help constrain model output and protect against harmful actions. Human-in-the-loop architecture should be used for deployments where errors can have drastic effects.

# Trustworthiness for foundation models and LLMs

Foundation models and LLMs present new challenges to trustworthiness. The training data used for LLMs is often unknown, yet it fundamentally determines model behavior and reliability. Often, that data comes from unreliable or potentially malicious sources, such as scrapes of the public internet. LLMs are inherently non-deterministic, which means they can generate different text for the same input, may generate factually incorrect content, and can be sensitive to the phrasing of the input or prompt. Even with large context windows, their performance degrades as the context length increases [11]. Architectures of LLMs are often unpublished, which makes it difficult to explain their behavior. LLM explainability remains challenging, even if the model architecture is known, because of its black-box architecture, billions of parameters, and emergent behavior.

One approach is to ask the model to produce step-by-step natural language explanations for its output; however, this is prone to common LLM reliability challenges. Attention visualization shows the relative strength of attention weights between the tokens in the model and can provide insight into which tokens were used, or given attention, when predicting the next token.

Both model accuracy and explanation quality can be enhanced by integrating knowledge bases that include semantic graphs and ontologies. By explicitly modeling the semantics of telecommunications concepts and storing actual network configurations within this framework, we can more effectively ground LLMs in the telecom domain and, for example, improve the accuracy of telecom-specific reasoning. This approach allows for the generation of explanations based on well-defined concepts derived from the underlying ontology.

As with traditional ML, LLM training data may include sensitive content such as PII, raising similar privacy concerns. The model can also absorb sensitive data if the system is structured such that it learns continuously from user input. Even if such data is de-identified, the analytical and highly correlative nature of the LLM opens the possibility of using it to attack that de-identification. This may be an easier attack to mount for an LLM than for a traditional ML model because of the typical conversational nature of LLM systems. Bias in outputs is also a concern, since the model will generally reflect any bias in the training data. Researchers have shown that only a small amount of training or fine-tuning data needs to be affected to impart bias to the overall model [13].

Implementing a data governance framework is essential for training, fine-tuning, testing, and using LLMs, so that their outputs can be trusted. Data sources must be trustworthy, collected data must be properly curated, and licenses, intellectual property, and copyright requirements must be respected. Data retention policies and regulatory compliance are mandatory.

End-to-end data traceability or data provenance must be implemented from data sources to their use in training, testing, and inference. It is essential to be able to assure the desired data quality. Because LLMs may regurgitate training data, anonymization of training data, removal of PII, and implementation of data leakage prevention mechanisms are necessary. If Retrieval Augmented Generation (RAG) or knowledge bases are used, the data governance framework should cover them as well.

LLMs share the same security concerns as any large application and introduce some unique ones. One notable attack is prompt injection. The input fed to the model consists of three parts. The system prompt is instructions to the model written by its creators or deployers. The user prompt is instructions or questions to the model from the user. User data is any data provided for the model to process. These three components are concatenated and fed to the model as a single stream. As a result, if the data, for example, contains instructions, the model may interpret them as such and act on them. These instructions might override the system and user prompts and cause the model to act incorrectly. There is particular danger when the user data is retrieved from an untrusted source, such as content pulled from the internet. Similarly, the user may attempt to put content in the user prompt that overrides controls or guardrails in the system prompt. Prompt injection is an example of a classic vulnerability known as confusion of control and data channels, where data is mixed with instructions and therefore interpreted as commands.

The only sure prevention for prompt injection would be to separate the control and data channels, but the inherent nature of LLMs makes this impossible. Various techniques have been devised to protect against prompt injection, but none have been found to work with a high degree of success [14]. Best practices currently involve using a layered defense strategy, based on a risk analysis. Controls include instructions in the system prompt to constrain model operation (also known as guardrails), filtering of the user prompt, restrictions on allowed data sources, filtering of user data, using multiple models separated into trusted and untrusted zones, and human-in-the-loop architectures where outputs are checked before use. In situations where model output can affect other systems or the physical world, for example, in agentic systems, surrounding the model with deterministic or algorithmic code that constrains its actions can be useful.

# Trustworthiness for agentic AI

Agentic architectures use agents and tools to create systems that execute complex tasks autonomously, given natural language requests. Agents, the brains of the system, are based on LLMs and can perform reasoning to understand a request, break it into pieces, and execute them in order, calling other agents and tools as necessary. Tools, the hands of the system, perform desired operations. A wide body of tools is emerging that can do almost anything, from reading and writing data to operating physical systems.

Standardized protocols such as Agent to Agent [15] and the Model Context Protocol—between agents and tools [16] link the components together. Tools publish natural language descriptions of their capabilities, so agents can intelligently select the appropriate ones for a given task. This represents an enormous increase in system power and flexibility. Rather than a defined set of functions developed into the system, an agentic system with a sufficiently powerful set of tools can take a natural language request and do something new. This power means that agentic systems will be a key component of future telecom systems, supporting autonomous operation of networks and powerful management capabilities.

Increased power, however, brings with it increased risk. Risks related to trustworthiness include:

- the highly parallel nature of agentic systems, coupled with automation bias, often leading to inadequate testing and limited visibility into system-level performance and robustness
- incorrect or malicious agent actions, due to hallucination or resulting from prompt injection
- excessive power of tools, such as the use of general-purpose tools with broad capabilities when only a small set of functions is needed
- inadequate sandboxing of code generated by tools
- inadequate or uncoordinated authentication and authorization, particularly when integrating agents and tools from multiple providers
- chaining effects, where errors move from agent to agent, or are remembered in system memory and affect future operations
- immature implementations - because these protocols and their implementations are new, their security characteristics are not fully understood, and they lack long periods of production use and testing
- the dynamic and nondeterministic nature of agentic systems makes it difficult to test thoroughly, because not all possible flows of agent actions can be predetermined

It is important to do thorough risk analysis on agentic systems, applying zero-trust principles. Data sources and their potential for prompt injection should be analyzed. Human oversight and approval before execution of dangerous operations can help, but in some cases is not practical if the system operates faster than human perception and reaction. In those cases, consider putting algorithmic guards—not LLM-based and therefore not corruptible—in place to check and constrain actions.

Security of agentic systems is an active area of research, and many good resources are emerging, such as the Open Worldwide Application Security Project Top 10 for agentic applications [17]. Multi-agent collaborations in an agentic AI framework may often showcase good and bad emergent behavior more than in generic multi-agent systems. These systems are being increasingly considered for deployment in the telecommunication domain, where safety and performance are critical. Thus, establishing methods for identification, specification, detection, prevention, and mitigation of harmful emergent behaviors has become essential to ensure overall trustworthiness requirements.

# Example use cases

A typical case where explainable AI is used to perform root-cause analysis is network slice assurance. For slice assurance, the agreed network properties or key performance indicators in the service level agreement must be guaranteed. Otherwise, a penalty has to be paid. One use of AI and XAI is to predict a potential violation that could occur in the future and then resolve it in an automated fashion. Model-level, post-hoc explanation techniques that are attributive and counterfactual, are used to perform root-cause analysis and understanding the key-factors involved [4], [5].

In mobile network optimization, reinforcement learning (RL) has been implemented for the Remote Electrical Tilt (RET) use case, which automates tilting of the antenna. RL is again a black-box method. Multiple explainable RL (XRL) methods are being researched. One of them is our in-house developed Both Ends Explanations for Reinforcement Learning (BEERL) method [6] that generates a detailed explanation to check the correctness of the model and to justify the selection of an RL action taken. For example, explaining why an uptilt action was chosen instead of downtilt.

Other methods being researched generate interactive explanations, such as Autonomous Policy Explanations (APE) [7]. This generates explanations in natural language format. It is interactive and useful for a high-level user or management person in checking the AI model's correctness. A user can inquire about an action, for example, "When is uptilt performed?" The method responds back with the response "Uptilt is performed when the cell has good signal quality and the cell has high throughput". Or the user can ask about a condition, for example, "What happens when coverage is low?" and the method responds back with "The antenna will uptilt". While APE is retrospective, Temporal Policy Decomposition is another novel in-house developed method that captures even expected future outcomes [8]. This information is important to clarify and explain future situations such as when an agent might compromise its current action in order to maximize its future cumulative rewards.

Figure 1 presents a RAN energy saving use case, where feature importance and APE methods have been researched at different lifecycle phases of an RL agent, and later deployed in a customer network [9]. Feature importance using Shapley Additive Explanations was used before deploying the agent to the real network to ensure the correctness of the model. After the deployment phase, APE was used to validate the outcome—whether the agent behaved as expected or not. In addition to ensuring correctness and providing transparency of the models, it established customer trust for moving the model from the simulator environment to the live network.

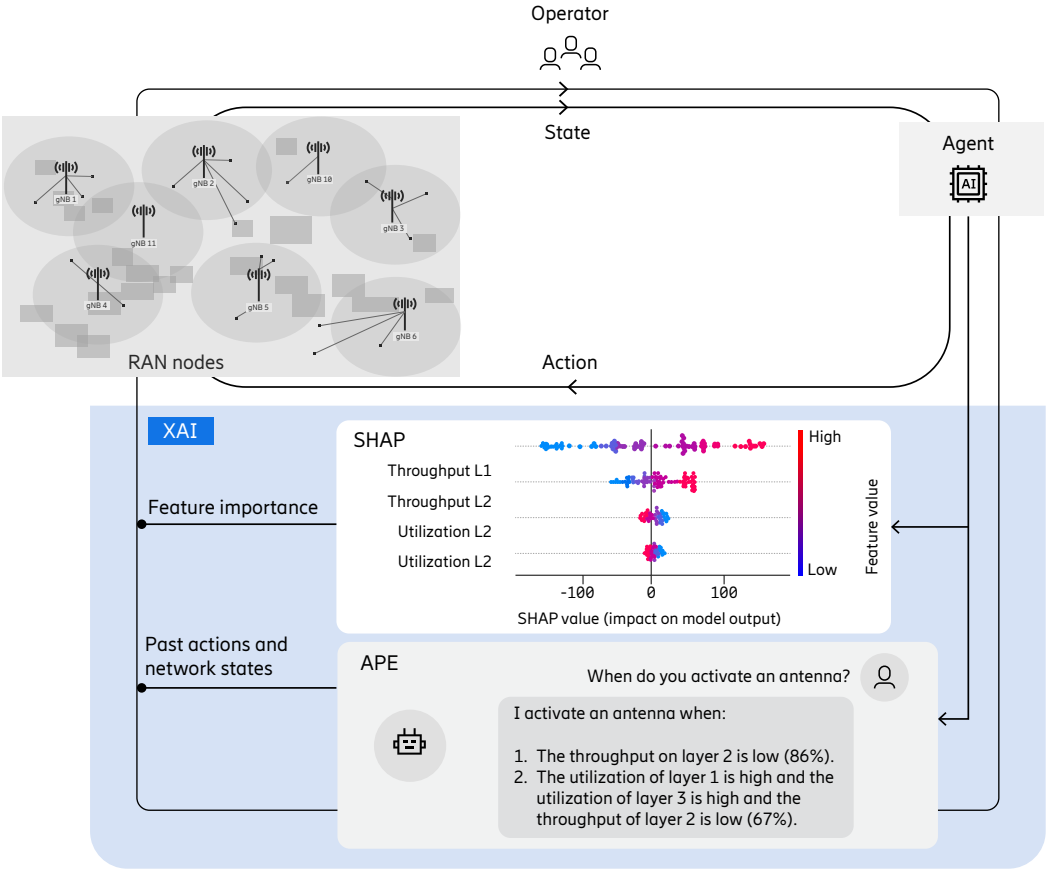



Figure 1. Explainable RL on energy saving use case to analyze agent's behavior and to evaluate deployment results.

Furthermore, we have shown that LLMs can be used to enhance explainability for RL models, which is more dynamic and flexible than APE explanation. This translates feature importance outputs into natural language explanations, as shown in Figure 2. [12].

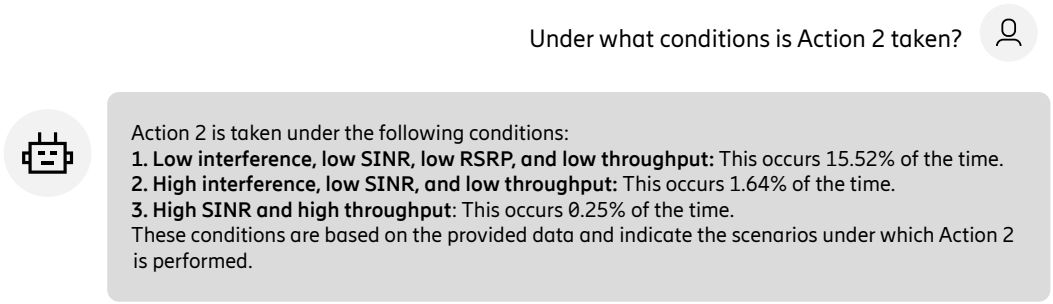



Figure 2. LLM implementation to explain a trained RL model for RET use case.

More reliable and robust answers from LLMs can be achieved by analyzing their input-output behavior. This can be done by measuring the amount of change observed in the output prompt in relation to perturbations in input prompt. LLM-powered agents are being widely adopted in intent-based service management or other similar use cases in the telecommunication domain.

Intent-based networking [10] leads to automatically configurable network operations, abstracting away the complexities through Intent Management Functions (IMFs) used by intent owners or intent handlers at different levels of operation—business, service, or network (RAN/Transport/Core). Transparency is a key requirement in the decision-making process by IMFs at the system level. Here, explanations can be provided as part of intent reports addressing multiple scenarios: why an intent was rejected, why an intent's requirements cannot be fulfilled, why an action taken was necessary, or why an observation was required to fulfil an intent. As shown in Figure 3, given the decomposed intent through the different layers, northbound explanations directed from intent handler to intent owner are provided to state why an intent was or was not fulfilled in accordance with the standardized format for intent reports.

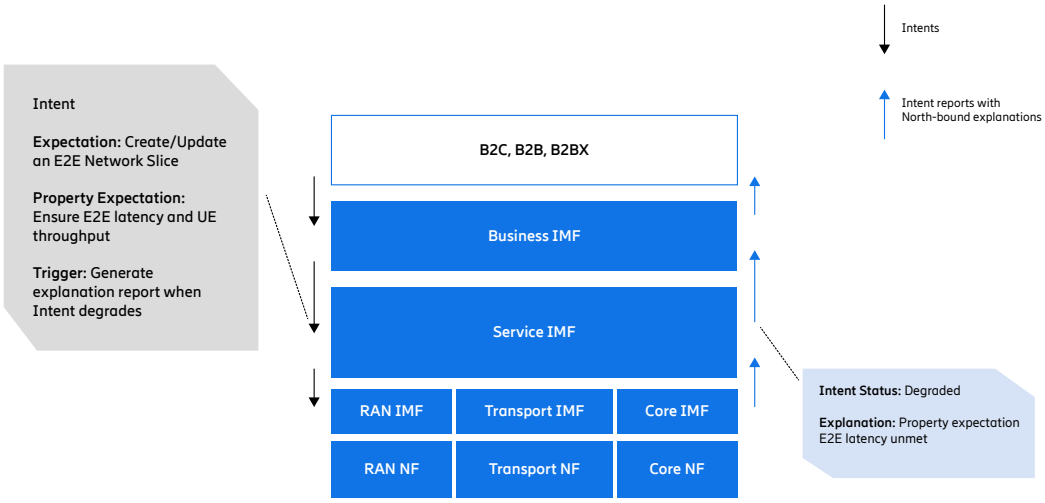



Figure 3. System-level explainability through the network domains - business, service and resource layers.

# Conclusion

Modern telecommunications systems will depend heavily on AI and, as they become more autonomous, will leverage the latest AI developments such as LLMs, generative AI, and agentic architectures. For these systems to be trustworthy, the AI inside them must be trustworthy. Developers and deployers must take this into account, consider the entire AI lifecycle, and adopt an approach of trustworthy-by-design.

Comprehensive risk analysis must be applied, covering traditional and modern AI, with controls starting at the earliest development phase. Appropriate controls against traditional attacks include authentication and access control; data governance, protection and de-identification; adversarial training; algorithmic security controls; and invariance and directional expectation testing. For LLMs and agentic architectures, important controls are prompt guardrails, trust-zone separation, data source restrictions, and restrictions on the duration and scope of agent powers.

For all types of AI, explainability techniques can help achieve both the security and quality aspects of trustworthiness. Human-in-the-loop controls, along with transparency measures and auditability, ensure the AI remains answerable to its users.

In 5G and 6G systems, and beyond, AI-based architectures will provide unprecedented features and power. With appropriate attention during their development and deployment, they can provide unprecedented trustworthiness as well.

# References

1. EU AI ethics guidelines
2. AI Act
3. Autonomous networks
4. White Paper on 'Explainable AI – How humans can trust AI?' https://www.ericsson.com/en/reports-and-papers/white-papers/explainable-ai--how-humans-can-trust-ai
5. A. Terra, R. Inam, S. Baskaran, P. Batista, I. Burdick, and E. Fersman, "Explainability Methods for Identifying Root-Cause of SLA Violation Prediction in 5G Network," in GLOBECOM 2020 - 2020 IEEE Global Communications Conference, Dec. 2020, pp. 1–7. doi: 10.1109/GLOBECOM42002.2020.9322496.
6. A. Terra, R. Inam, and E. Fersman, "BEERL: Both Ends Explanations for Reinforcement Learning," Applied Sciences, vol. 12, no. 21, p. 10947, Jan. 2022, doi: 10.3390/app122110947.
7. M. Hefny, A. Terra, and A. Valencia, "Comprehensive Reinforcement Learning Explanations Using Queries," in Explainable Artificial Intelligence, vol. 2580, R. Guidotti, U. Schmid, and L. Longo, Eds., in Communications in Computer and Information Science, vol. 2580. , Cham: Springer Nature Switzerland, 2026, pp. 27–40. doi: 10.1007/978-3-032-08333-3_2.
8. F. Ruggeri, A. Russo, R. Inam, and K. H. Johansson, "Explainable Reinforcement Learning via Temporal Policy Decomposition," Jan. 07, 2025, arXiv: arXiv:2501.03902. doi: 10.48550/arXiv.2501.03902.
9. "Telenor and Ericsson successfully co-create Agentic AI - Telenor Group." [Online]. Available: https://www.telenor.com/media/newsroom/announcement/telenor-and-ericsson-successfully-co-create-agentic-ai/
10. "Autonomous networks with multi-layer, intent-based operation." [Online]. Available: https://www.ericsson.com/en/reports-and-papers/ericsson-technology-review/articles/autonomous-networks-multi-layer-intent-based

11. Y. Du et al., "Context Length Alone Hurts LLM Performance Despite Perfect Retrieval," Oct. 06, 2025, arXiv: arXiv:2510.05381. doi: 10.48550/arXiv.2510.05381.

12. M. U. Ahmed, LLM for AI Explainer. 2025. [Online]. Available: https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-564822

13. "A small number of samples can poison LLMs of any size." [Online]. Available: https://www.anthropic.com/research/small-samples-poison

14. S. Gulyamov et al., "Prompt Injection Attacks in Large Language Models and AI Agent Systems: A Comprehensive Review of Vulnerabilities, Attack Vectors, and Defense Mechanisms," Information, vol. 17, no. 1, p. 54, Jan. 2026, doi: 10.3390/info17010054.

15. a2aproject/A2A. Shell. Agent2Agent (A2A) Project. [Online]. Available: https://github.com/a2aproject/A2A

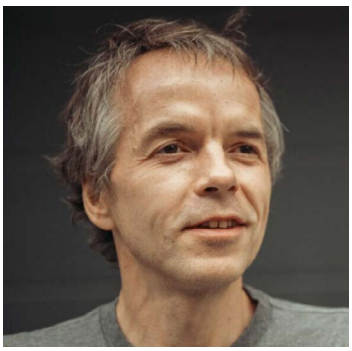
16. "What is the Model Context Protocol (MCP)?," Model Context Protocol. [Online]. Available: https://modelcontextprotocol.io/docs/getting-started/intro

17. Owaspg. Editor, "OWASP Top 10 for Agentic Applications for 2026," OWASP Gen AI Security Project. [Online]. Available: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

# Authors

**Rafia Inam** is a Senior Research Manager at Ericsson Research in Trustworthy AI and Adjunct Professor at The Royal Institute of Technology (KTH), Sweden. She has conducted research for Ericsson for the past eleven years on 5G for industries, network slices, and network management; and AI for automation. She specializes in trustworthy AI, Explainable AI, Ai regulation, risk assessment and mitigations using AI methods, and safety for cyber-physical systems for telecom and CPS. She is also contributing to trustworthy AI based standardization specially to European standards in CEN/CLC based on EU AI Act. Rafia has co-authored 55+ refereed scientific publications and 60+ patent families, and 2 best paper awards. She won Ericsson Top Performance Competition 2021 on her work on AI for 5G network slice assurance and was awarded multiple Ericsson Key Impact Awards.

**Jim Reno** is a Distinguished Engineer at Ericsson, where he works on security aspects of Artificial Intelligence as applied to telecommunication systems. He has more than 40 years of industry experience in fields including system software (operating systems, networking, system management, and cloud native systems), payment system security, authentication, authorization and identity management.



**Attila Ulbert** joined Ericsson in 2015 and he is currently Artificial Intelligence System Manager. In his enthusiastic journey with Ericsson, he lead the development of Ericsson's AI platform, and worked on fundamental AI studies on security, trustworthiness, and industrialization. Attila has a PhD in Informatics from Eötvös Loránd University. He is a marathoner.

**Ahmad Terra** is an experienced researcher at Ericsson, specializing in explainability for telecommunications. Terra holds a PhD in Machine Design from KTH Royal Institute of Technology and previously worked at ABB as a robotics engineer. At Ericsson, Terra completed an industrial PhD, developing an explainable reinforcement learning method (BEERL) that has been adopted in an Ericsson product. Terra has co-authored more than ten peer-reviewed publications and has more than ten filed patents. Current research focuses on exploring explainability within intent management functions.

**Vandita Singh** is an Experienced Researcher at Ericsson Research, Sweden, specializing in Explainable Artificial Intelligence for autonomous networks. She holds her master's degrees from Uppsala University, Sweden and C-DAC, India, with prior experience in academia. Her research work is primarily focused on system-level explainability for intent-driven networks, explainable AI evaluation metrics, and trustworthiness aspects for emerging AI technologies. She has contributed to 6G standardization, with several filed patents and co-authored journal and conference papers.