JUNE 2025

COMMUNICATIONS SECURITY, RELIABILITY, AND INTEROPERABILITY COUNCIL IX

Report on Threats Posed by Artificial Intelligence/Machine Learning Systems to the Security, Reliability and Integrity of Networks and Recommendations on How to Overcome Them

PREPARED BY
WORKING GROUP 1: HARNESSING ARTIFICIAL INTELLIGENCE/MACHINE LEARNING TO ENSURE THE SECURITY, RELIABILITY, AND INTEGRITY OF THE NATION'S COMMUNICATIONS NETWORKS

Table of Contents

Executive Summary

The telecommunications industry is embracing the use of Artificial Intelligence (AI) in its sophisticated next generation networks, driven by a surge in connected devices and data-intensive applications. AI is emerging as a critical enabler to complement traditional, manual network management processes to meet dynamic demands, including automated resource allocation and network optimization. While AI can significantly enhance efficiency, performance, and security, its deployment also introduces new security risks and vulnerabilities that require careful mitigation. The Federal Communications Commission (FCC or the Commission) has tasked the ninth Communications Security, Reliability, and Interoperability Council (CSRIC IX) to identify these risks and recommend mitigations.

Approach. This report examines the technical and operational implications of integrating AI into telecommunications by identifying AI use cases, threats, and mitigations in five distinct areas of wireless networks: End User Equipment, Radio Access Networks (RAN), Backhaul, Core Network, and Operations Support Systems. This analysis is not meant to be exhaustive but provides concrete examples and a general framework for operators and vendors to consider when developing and implementing AI solutions and hardening networks against potential threats.

Risks & Vulnerabilities. To frame the analysis, the report adopts the U.S. Department of Homeland Security's (DHS) Framework for Artificial Intelligence in Critical Infrastructure, which groups AI risks into three primary categories.[1] First, adversaries can exploit AI to augment their attack capabilities by automating cyber compromises, evading detection, or launching targeted physical and digital assaults on critical infrastructure. Second, AI systems themselves are susceptible to attacks, ranging from input perturbations, prompt injections, and denial-of-service via external users, to insider threats that poison data, alter algorithms, or manipulate evaluation benchmarks. Third, failures in AI design and implementation, such as brittleness under unforeseen conditions, inherent inscrutability, statistical biases, and inconsistent system maintenance, can exacerbate these risks. Recognizing that these vulnerabilities can manifest at any stage of the AI model lifecycle, from planning and data preparation to deployment, inferencing, and ongoing monitoring, operators should integrate robust risk mitigation strategies, standardized protocols, and continuous evaluation processes into their network management practices to mitigate these risks and enhance security, reliability, and interoperability.

Mitigations & Recommendations. This report provides recommendations to enhance the security of communications networks, addressing both AI-specific risks and broader relevant network security concerns. Best practice recommendations outlined in this Report specifically should be considered voluntary and implemented in a manner that is appropriate to the needs, resources, and capabilities of each individual organization.

- Implementing a zero-trust security strategy with verification, least privilege access, and an assumption of breach, protecting all network layers with strong access controls and standardized mechanisms such as those defined by the 3rd Generation Partnership Project (3GPP).

- Promoting comprehensive AI education and awareness training, especially related to the use of AI in telecommunications. Training should include considerations for evaluating and adopting AI-based technologies in a secure and responsible manner. Appendix A contains a good list of resources as a starting point, and a forthcoming report from this CSRIC will address "Recommended Best Practices for the FCC and Industry on the Ethical and Practical Use of Artificial Intelligence/Machine Learning."

[1] U.S. Department of Homeland Security (DHS), *Roles and Responsibilities Framework for Artificial Intelligence in Critical Infrastructure*, https://www.dhs.gov/publication/roles-and-responsibilities-framework-artificial-intelligence-critical-infrastructure.

- Conducting a thorough threat analysis, considering how multiple, low-risk vulnerabilities can be combined to create high-impact threats, which AI can facilitate, especially when interfaces are internet-exposed.

- Promoting access to training and test data, and sample models for key telecommunications AI use cases to foster a robust ecosystem of model providers for telco scenarios. Standardized interfaces, datasets and schemas will help vendors create standardized AI solutions for the telecommunications network, and accelerate operators' ability to evaluate, contrast and securely deploy AI systems.

- Promoting collaboration among equipment manufacturers and network operators to conduct rigorous monitoring of the multivendor ecosystem through bills of materials, root-of-trust frameworks, secure data pedigree tracking, and continuous risk assessments to enhance transparency, mitigate AI-related threats, and protect telecommunications infrastructure.

As part of adopting the aforementioned topics, some key themes emerge including:

- Protecting data -- both during AI development and deployment-- with safeguards for personal data; encryption; and/or integrity safeguards to prevent tampering as well as controls to track the origin and if possible, the identity and provenance of AI data such as AI training data.

- Continuous monitoring and anomaly detection to promptly identify and respond to unusual events in the network behavior.

- Implementing overload protection to prevent AI functions from consuming excessive resources and causing system failures.

- Regularly assessing and updating AI systems with fresh datasets to stay ahead of potential performance drifts and emerging threats.

- Thorough output testing and validation protocols to verify the reliability of AI-generated responses before full-scale deployment.

Overall, this report underscores the dual nature of AI in telecommunications: its potential to revolutionize network operations and enhance efficiency, and the need for proactive, robust safeguards to address the multifaceted risks it introduces. By aligning with rigorous industry standards and adopting a risk-aware approach throughout the AI model lifecycle, both regulators and industry stakeholders can better protect public safety and ensure resilient, secure telecommunications networks in the evolving digital landscape.

1 Introduction

The increasing complexity of networks, driven by the exponential rise of the number of devices and data-heavy applications, challenges traditional network management. Manual processes and static automation—standing alone—may not be sufficient to keep up with the dynamic demands of modern networks, especially as networks transition to 5G Advanced and beyond.[2] AI offers an opportunity to adaptively allocate resources and optimize networks for network efficiency, performance, and security. The communications industry is already developing standards for the use of AI.[3]

While AI has great potential to contribute to network protection, it also can introduce new attack surfaces. Adversaries and other bad actors might use AI as an attack vector, making it crucial to secure critical AI assets such as training data, models, and their parameters from unauthorized access and tampering. As AI models depend on training data and because of the probabilistic nature of AI model outputs, AI developers and deployers must take a risk-based approach to design, develop, and deploy applications for outputs to be trustworthy and secure.

AI's rapid evolution makes it difficult to predict where it will be in 5 years; hence, we view the challenges in a realistic manner based on what we know today so as to alert future developers of relevant issues that could impact telecommunications deployments of the future.

1.1 CSRIC Structure

CSRIC IX was established at the direction of the Chairperson of the Federal Communications Commission (FCC or Commission) in accordance with the provisions of the Federal Advisory Committee Act.[4] The purpose of CSRIC IX is to provide recommendations to the FCC regarding ways the FCC can strive for security, reliability, and interoperability of communications systems. CSRIC IX's recommendations will focus on a range of public safety and homeland security-related communications matters. The use of AI in telecommunications networks is a new focus area for CSRIC.

The FCC created informal subcommittees under CSRIC IX, known as working groups, to address specific tasks. These working groups must report their activities and recommendations to the Council as a whole, and the Council may only report these recommendations, as modified or ratified, as a whole, to the Chairperson of the FCC.

[2] A recent 5G Americas white paper gives a detailed view of the use of AI in current and future wireless networks. *See Artificial Intelligence and Cellular Networks*, https://www.5gamericas.org/artificial-intelligence-and-cellular-networks/.
[3] *See, e.g.,* O-RAN Alliance, *Principles and Methodologies for AI/ML Testing in Next Generation Networks*, https://www.o-ran.org/research-reports/principles-and-methodologies-for-ai-ml-testing-in-next-generation-networks.
[4] 5 U.S.C. App. 2.

Communications Security, Reliability, and Interoperability Council IX		
Working Group 1: Harnessing Artificial Intelligence/Machine Learning to Ensure the Security, Reliability, and Integrity of the Nation's Communications Networks	**Working Group 2: Ensuring Consumer Access to 911 on All Available Networks As Technology Evolves**	**Working Group 3: Preparing for 6G Security and Reliability**
Co-chairs: Vijay Gurbani, Vail Systems Jason Hogg, Microsoft	Co-chairs: Brandon L. Abley, NENA Stephen Hayes, Ericsson	Co-chairs: Brian Daly, AT&T George Woodward, Rural Wireless Association
FCC Liaison: Zenji Nakazawa/ Kurian Jacob	FCC Liaisons: Gerald English, Ryan Hedgpeth, Zachary Dileo	FCC Liaison: Jeffery Goldthorp/Kenneth Carlberg

Table 1 - Working Group Structure

1.2 Working Group 1 Team Members

The members of Working Group 1 include a mix of experts from industry and government:

Name	Company
Vijay K. Gurbani * (Co-Chair)	Vail Systems, Inc.
Jason Hogg * (Co-Chair)	Microsoft
Mark Annas *	City of Riverside (CA) Office of Emergency Management
Praveen Atreya *	Verizon
Mike Barnes *	Mavenir Systems, Inc.
Richard Barron	MITRE Corp.
Chris Bennett	Motorola Solutions
Craig Bowman	Futuri
Matt Carothers	Cox Communications
Christina Chaccour	Ericsson, Inc.
Andrew Drozd *	ANDRO Computational Solutions, LLC
Luiz Eduardo *	Hewlett-Packard Enterprise
Bob Everson *	Cisco Systems
Ben Goldsmith *	U.S. Department of Justice
Mark Grubb	Cybersecurity and Infrastructure Security Agency
Ankur Kapoor	T-Mobile
Yong Kim *	VeriSign, Inc.
Lauren Kravetz *	Intrado Life & Safety, Inc.
Salman Marvasti	Advanced Computer Concepts
Steve Mathesius*	ACA Connects
Tim May *	National Telecommunications and Information Administration
Martin McGrath *	Nokia
Brian Murray *	Harris County (TX) Office of Homeland Security & Emergency Management
Jonathan Petit	Qualcomm
Abir Ray	Expression Networks LLC

Name	Company
Travis Russell *	Oracle Communications
Peter Santhanam	IBM
Narothum Saxena *	UScellular
Peter Scott	PBS
Rikin Thakker	NCTA
David Valdez	CTIA
Henry Young *	The Business Software Alliance
Dongsong Zeng	U.S. Department of Commerce
**CSRIC Member*	

Table 2 - List of Working Group Members

Working Group members nominated alternates from within their organizations. Although these alternates are not members of the Working Group and may not vote, they provided valuable input towards the completion of this report that should be acknowledged. The Alternates include:

Name	Company
Anmol Agarwal	Nokia
Patrick Arsenault	Intrado Life & Safety
Michael Beirne	CTIA
Robert Cantu	NCTA – The Internet & Television Association
Devin Christensen	Cybersecurity and Infrastructure Security Agency
Sean Donelan	VeriSign, Inc.
Nars Haran	UScellular
John Hunter	T-Mobile
Jithin Jagannath	ANDRO Computational Solutions, LLC
David Marcos	Motorola Solutions
Olga Medina	The Business Software Alliance
Jennifer Oberhausen	Microsoft
Jim Reno	Ericsson, Inc.
John Roznovsky	Mavenir Systems, Inc.
Joseph Smetana	Vail Systems, Inc.
Mourad Takla	Verizon
Bill Tortoriello	ACA Connects
Lei Yu	Expression Networks LLC

Table 3 - List of Working Group Alternates

2 Objective, Scope, and Methodology

2.1 Objective

The FCC tasked CSRIC IX to provide recommendations on: (1) threats posed by AI systems to the security reliability and integrity of networks and how to overcome them; (2) best practices for the FCC and industry on the ethical and practical use of AI; and (3) best practices for the use of AI systems specifically intended for public safety networks. In this report, CSRIC IX addresses the first line of these efforts.

In creating this task, the FCC directed CSRIC IX to consider how AI increases the risks to the security, reliability, and interoperability of communications networks and how best to mitigate the risks caused by

the introduction of AI. In its study, CSRIC IX was directed to consider how the FCC and industry can: promote sound policies and practices that support public safety, network security, and resilience; promote the responsible use of AI; and prevent and mitigate harms associated with the use of AI. Among other issues, CSRIC was to consider current trends, developments, and related standards work in "hardening" AI, protecting data used for training, to identify gaps in efforts to develop AI's cyber-readiness and trustworthiness.

2.2 Scope

Communications networks encompass a broad scope. There are three broad segments where AI may be leveraged to increase performance and efficiency for a wide range of consumer, business, and public safety applications: first, the increasing use of AI in existing 4G, 5G and fixed-line networks; second, a potentially central role for AI in the development and operation of emerging and future networks (e.g., 6G) and applications; and third, the promise of AI to cater to the unique needs of the public safety community and the networks they rely on, including 911/Next Generation 911 and emergency alerting platforms. Across these three segments, there are significant implications to consumers from the use of networks deploying AI technology, including ethical and practical use of AI. Because of the breadth of this AI landscape, this report will assess AI principally in the context of current network design and operations, the implications for businesses and consumers that rely on them, the threats, if any, posed by the use of AI in these contexts, and recommendations to overcome potential threats.

2.3 Methodology

Given the broad AI landscape even within the currently deployed networks (4G, 5G and wireline), CSRIC decided to focus on the most relevant AI use cases, threats, and mitigations in five distinct areas: End User Equipment, RAN, Backhaul, Core Network, and Operations Support Systems. The deployment of 5G networks will drive the evolution of 6G; therefore, this Report's Appendix C includes an overview of the potential impact on 6G network design and deployment. While not exhaustively addressing every potential AI use case or threat to any type of communications network, this distillation serves to identify key threats and a set of recommendations for the FCC and industry to consider in the journey to use AI in current and future communications networks.

Communications networks used by and for public safety merit focused consideration regarding AI risk and safe use practices. Within the context of public safety, the manner in which information is used (where "use" broadly construed refers to how information is collected, stored, accessed, protected, and used to make accurate and timely decisions) is immediately rights-impacting and often has life or death implications for emergency responders and the people they serve. If applied correctly, evolving AI technologies promise new ways to use information in support of public safety missions. However, adversaries may leverage this technology to infiltrate and disrupt public safety operations, causing serious consequences including erosion of public trust in government services, decreased effectiveness of public safety service provisioning, and loss of life and property. For these reasons, in a later report, CSRIC will focus on risk mitigation best practices for AI use within public safety networks.

3 Introduction to AI and Machine Learning

In his pioneering 1950 paper, mathematician and computer scientist Alan Turing introduced the modern formulation of AI -- computing machines emulating human intelligence in some context.[5] He devised the "Turing Test" to measure a machine's ability to exhibit intelligent behavior indistinguishable from that of a human, through a conversational task. Over the next few decades, significant progress was made in various areas of AI, such as perception, knowledge representation, reasoning, search, common sense,

[5] A. M. Turing, *Computing Machinery and Intelligence*, Mind, New Series, Vol. 59, No. 236 (1950), pp. 433-460.

rule-based systems, and planning.[6] Symbolic AI focused on using human understandable symbols and logic (e.g., rules, lists, graphs) to represent knowledge and reason about it. An example of this was called "Expert Systems" that relied on explicit "if-then" rules that are defined by humans *but lacked the capacity to learn the rules automatically*. Such rule-based systems are best suited for narrow, well-structured domains with known patterns where deterministic behavior is expected, such as to set up policies or guardrails that cannot be violated.

Most contemporary AI systems, built on machine learning (ML) techniques,[7] are best understood as highly sophisticated statistical models that identify complex patterns in large datasets, called "training data." AI systems use those patterns to make probabilistic predictions of outputs for given inputs to perform a specific task. This is made possible because of the advances in hardware, including Graphics Processing Units, access to vast computing infrastructure (e.g., cloud architectures), and the availability of massive quantities of data from the internet and other digital sources. Typical AI models have huge numbers (i.e., millions to billions) of parameters that are optimized during the training process to provide the required accuracy and expected performance, thus rendering them far beyond human comprehension.

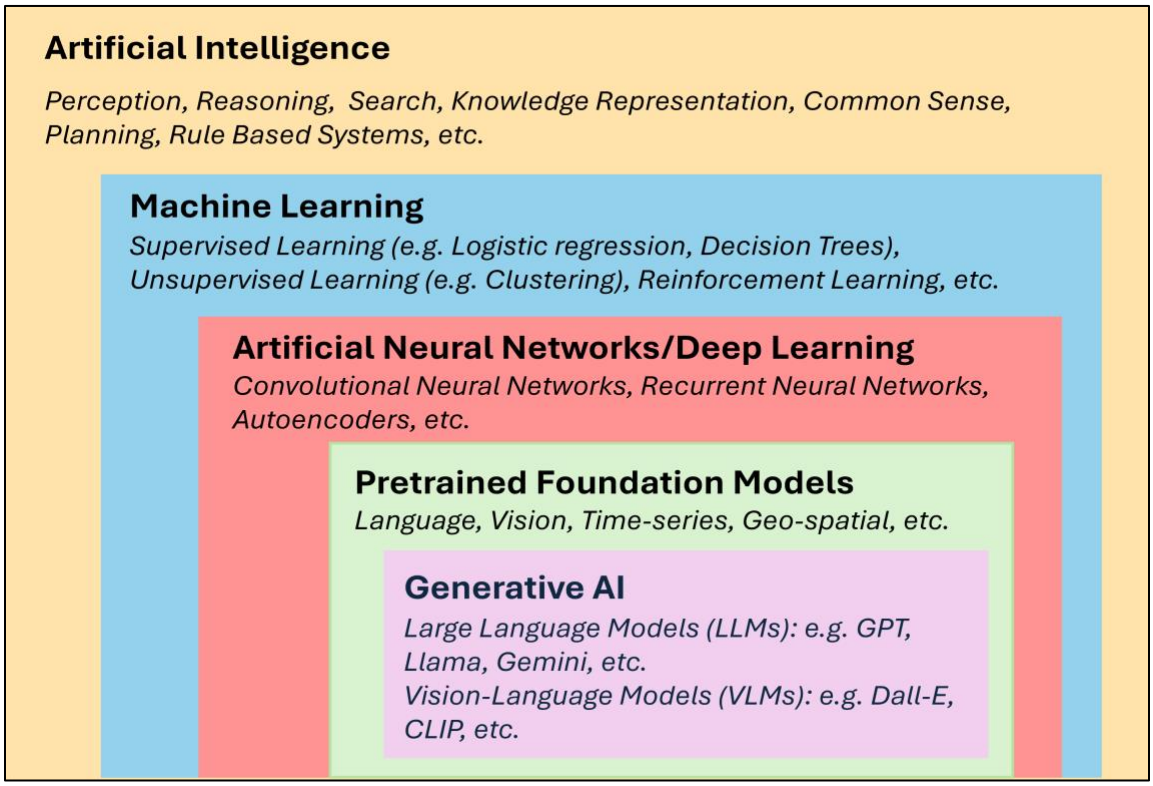

Figure 1: Various concepts in the AI discipline and their relationships.
Adapted from "Beyond Algorithms: Delivering AI for Business", J. Luke, D. Porter and P. Santhanam, CRC Press (2022)

The AI discipline encompasses a wide variety of concepts and techniques, tailored to specific tasks and objectives. Figure 1 depicts these concepts and their relationships. Due to significant advances in ML

[6] Stuart Russell & Peter Norvig, *Artificial Intelligence: A Modern Approach (Pearson Series in Artificial Intelligence),*(Pearson, 4th ed. 2020).
[7] Machine learning techniques refer to the computational methods and algorithms that enable systems to learn patterns from data and make predictions or decisions with minimal human intervention.

techniques using artificial neural networks, they have become the predominant choice to build AI models.

Predictive AI. Predictive AI refers to the use of ML to perform a classification or prediction task based on the input. Examples of a classification task include identifying an object contained in an input image or assessing the sentiment of a customer based on what she said about a product or service. This technique, called "Supervised Learning," requires a human to provide the training data, such as pictures of animals with corresponding labels of the data, such as "dog" or "cat". In this scenario, the ML model learns what a "dog" or "cat" image would be by learning the intrinsic features in the training data images with their associated labels. Depending on the specific use case, labeling large datasets can be expensive. This is where pre-trained models -- also known as "Foundation Models" -- can be helpful.[8] They are large, self-supervised models trained on vast unlabeled datasets. For example, predictive AI for customer support can be implemented using models trained on customer data from an organization or using a pre-trained model, such as a language model, with some local adaptation for specific customer support needs.

Generative AI. The primary objective of generative AI is to produce outputs (e.g., text, image, audio) that closely resemble those created by humans in response to a user request, typically made through text. The recent acceleration in generative AI capabilities is largely due to the development of foundation models using so-called "transformer" architecture focused on attention mechanisms.[9] This enables models to focus selectively on the most relevant parts of the input during output generation. This breakthrough significantly advanced performance in natural language processing. Foundation models with transformers, commonly called Large Language Models (LLMs), now underpin many of today's leading generative AI systems, including GPT, Gemini, and LLaMA. These systems can generate human-like responses to natural language prompts and exhibit strong generalization capabilities with minimal fine-tuning.

Notably, the quality and effectiveness of contemporary AI models to provide responses to user requests are dependent on four factors:

1. Quantity and quality of the training data
   .
2. Similarity between the training data and the data the model will encounter in production.

3. Uncertainties in model outputs due to intrinsic noise in the training data and a lack of knowledge about which model explains the observed data best.

4. The model's ability to learn from new observations.

This intuition is broadly applicable to all probabilistic AI methods and helps provide a foundation for understanding model failure in many contexts, including not only language applications but all environments within which AI models are expected to make predictions, recommendations, or decisions.

Overall, each of the AI models—rule-based, predictive, and generative—possesses unique strengths and characteristics. There is growing interest in hybrid approaches that combine symbolic reasoning with statistical learning. These methods seek to leverage the structure and interpretability of symbolic AI

[8] Jacob Devlin, et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,* Computation and Language, v2, 1-16 (2019).
[9] Ashish Vaswani, et al., *Attention is all you need,* NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, 6000 to 6010 (2017) (presenting the Transformer model architecture that is based entirely on attention, replacing the recurrent layers most used in encoder-decoder architectures with multi-headed self-attention that enables faster training).

alongside the adaptability and scalability of ML. New research suggests that the next generation of AI may increasingly draw on encoded knowledge to reason effectively while learning from smaller datasets, enhancing both transparency and data efficiency.[10]

4 Understanding AI Risk

There is extensive literature discussing the risks in developing and deploying AI systems that use ML models. For completeness, they are summarized in Appendix A of this report.

4.1 DHS Framework

We adopt the approach that DHS recommended in its April 2024 report on safety and security guidelines for critical infrastructure owners and operators of AI systems.[11] The report lists three overarching categories of risk:

- **Attacks Using AI**: The use of AI to automate, enhance, plan, or scale physical attacks on, or cyber compromises of, critical communications infrastructure.

- **Attacks Targeting AI Systems**: Focuses on targeted attacks on AI systems supporting critical communications infrastructure.

- **Failures in AI Design and Implementation**: This risk category stems from deficiencies or inadequacies in the planning, structure, implementation, execution, or maintenance of an AI system leading to malfunctions and/or exploitation by adversaries.

Of the three categories, it is important to understand that adversarial attacks using AI from outside the network cannot be controlled by the organization. However, it is critical that network operators understand the new threat vectors clearly (similar to traditional cybersecurity) to be able to protect against them through appropriate defense actions implemented in the AI model lifecycle. Understanding the AI model lifecycle is critical to assessing the potential threat vectors and mitigation plans.

4.2 AI Model Lifecycle Activities

Building an AI model to perform a specific task, such as object recognition or text classification, requires multiple steps. The following discussion applies to both predictive and generative AI models, even though the details of the specific step can be different. Figure 2 represents the various steps to create and deploy a ML model in practice. *For simplicity, Figure 2 does not include any other software engineering activity related to the integration of the model with the business application (e.g., access control).*

[10] Christina Chaccour, et al., *Less Data, More Knowledge: Building Next Generation Semantic Communication Networks*, IEEE Communications Surveys & Tutorials, Volume 27, Issue 1, 37-76 (2025) (Encoded knowledge is pre-stored information inside an AI model that helps it understand concepts, make decisions, and learn efficiently without needing a massive amount of new data.).

[11] DHS, *Mitigating Artificial Intelligence (AI) Risk: Safety and Security Guidelines for Critical Infrastructure Owners and Operators*, Apr. 2024, https://www.dhs.gov/sites/default/files/2024-04/24_0426_dhs_ai-ci-safety-security-guidelines-508c.pdf (*DHS Report*). DHS also provides a full list of risk subcategories and mitigations.

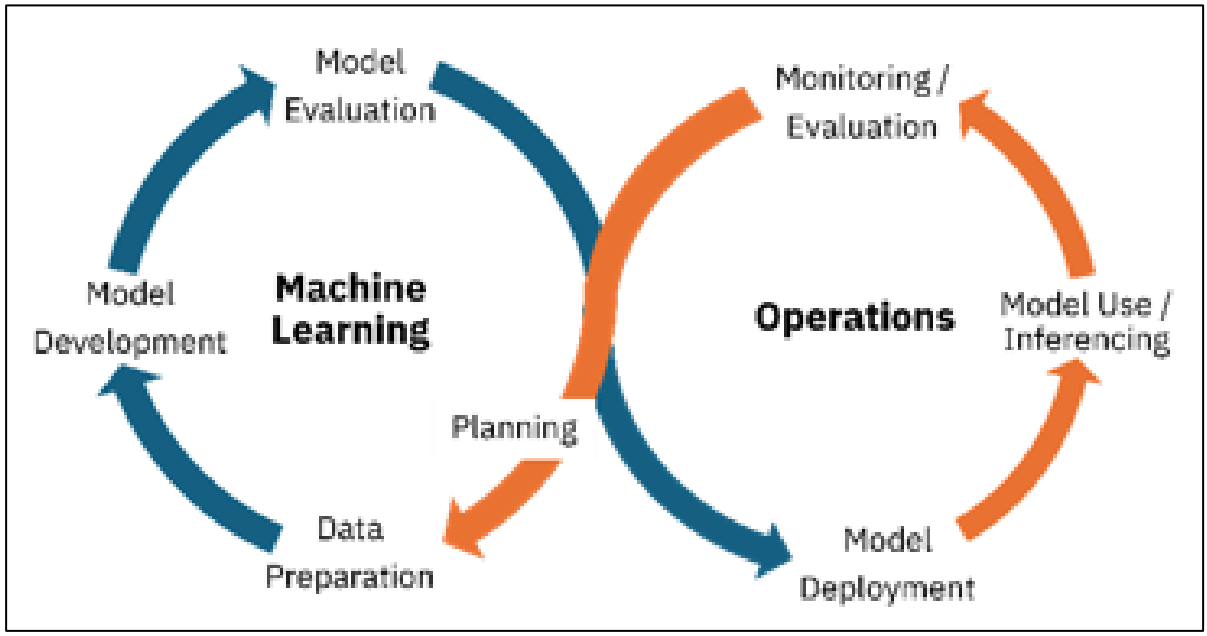

Figure 2 -- AI Model Lifecycle

Planning. This step includes (1) selecting the task(s) for the potential use of AI that meets the business requirements (e.g., accuracy, explainability, response time) and (2) the availability of necessary training data, technology infrastructure, and skills in the organization.

Data preparation. Quantity and type of data required (e.g., labelled vs. unlabeled, structured vs. unstructured, text, images) depends on the specific type of model to be built (see the next paragraph). This data may need preprocessing to remove noise, handle missing values, and normalize features.

Model Development. This step aims to find the best pattern in the training data that matches the output to the input data, using various algorithms. Details depend on the requirements of the business task. In telecommunications, most of the current AI use is predictive. Some emerging use cases, such as answering customer support queries in natural language, involve generative AI. Typical ML model choices are (1) predictive AI for a narrow task with dedicated data, (2) predictive AI using a pre-built foundation model with customization for a specific task, or (3) generative AI using a commercial or open source LLM, with some adaptation (e.g., retrieval augmented generation) for the specific use case (e.g., question-answering) and domain (e.g., telecommunications). Predictive models for narrow tasks can use any of the well-known algorithms, such as decision trees, support vector machines, or artificial neural networks.

Model Evaluation. Due to the non-deterministic nature of AI outputs, businesses are left with only benchmarks for evaluating the AI models. Depending on the specific areas of concern (e.g., accuracy, bias, robustness), businesses need to use appropriate benchmarks that align with their priorities.[12] Total accuracy is not possible.

Model Deployment. Deployment of an AI model (preferably automated) for production use includes the

[12] Here, benchmarks means standardized tests designed to measure performance across key areas. Benchmarks serve as structured evaluations that assess a model's ability to predict outcomes accurately (e.g., classification tasks); process information efficiently (e.g., response time); handle unexpected inputs robustly (e.g., adversarial testing); and generalize well to new data (e.g., ability to apply learned patterns from training data to previously unseen data while maintaining accuracy).

integration with the rest of the software system to support the business.

Model Use/Inferencing. In this step, users are interacting with the AI model to perform the intended task such as predict user location or radio channel quality.

Monitoring/Evaluation. During operational use, unexpected inputs can lead to degrading model performance (i.e., drift) that require monitoring and retraining of the model depending on the performance needs of the business. Depending on the nature and severity of the drift, a business may need to return to the planning step for the next level of improvements to the model, finishing the lifecycle loop.

4.3 Mapping Risks to AI Model Lifecycle

The purpose of this section is to explain how the categories of risk relate to the life cycle activities for creating and managing the deployment of an AI model.

4.3.1 Attacks Using AI

The following are non-exhaustive examples extracted from the DHS Report of how adversaries can use AI to augment their capabilities and operations to impact network operations. Key enablers of these attacks are the levels of automation possible with ML systems and their ability to create realistic network artifacts. AI also makes it faster and easier for attackers to identify critical vulnerabilities by using multiple, lower risk issues to accelerate vulnerabilities, especially if the interface is available over the internet.

- Injection of autonomous malware into networks.
- Automatic parsing of publicly available documentation for network vulnerability insights.
- Unauthorized data access by clever network indirections and deceptive access profiles.
- Evasion of cyberattack detection by network owners.
- Collecting and analyzing data to find and monitor physical targets (e.g., cell towers) for potential attacks.
- AI-enabled social engineering to manipulate users to obtain sensitive information that compromise security controls, such as the use of deep-fakes or AI-enhanced phishing / vishing attempts.[13]

As noted, it is not possible to prevent malicious actors from initiating these attacks, but they can be systematically addressed by appropriate technology and processes in the creation and deployment of AI systems, either through normal network design and operations or by defensive use of AI.

[13] Phishing is a cyberattack where scammers trick people into giving personal information, like passwords or credit card details, by pretending to be trustworthy sources, usually through fake emails or websites. Vishing is a type of phishing that happens over phone calls, where scammers use voice manipulation or false identities to deceive victims into revealing sensitive information.

4.3.2 Attacks On AI

Attacks on AI can happen under two distinct scenarios. First, an adversary is outside the network and has access to the system much like any other user. Second, an adversary has broken into the network or there is an insider in the organization with malicious intent. Figure 3 and Table 4 highlight these two scenarios.

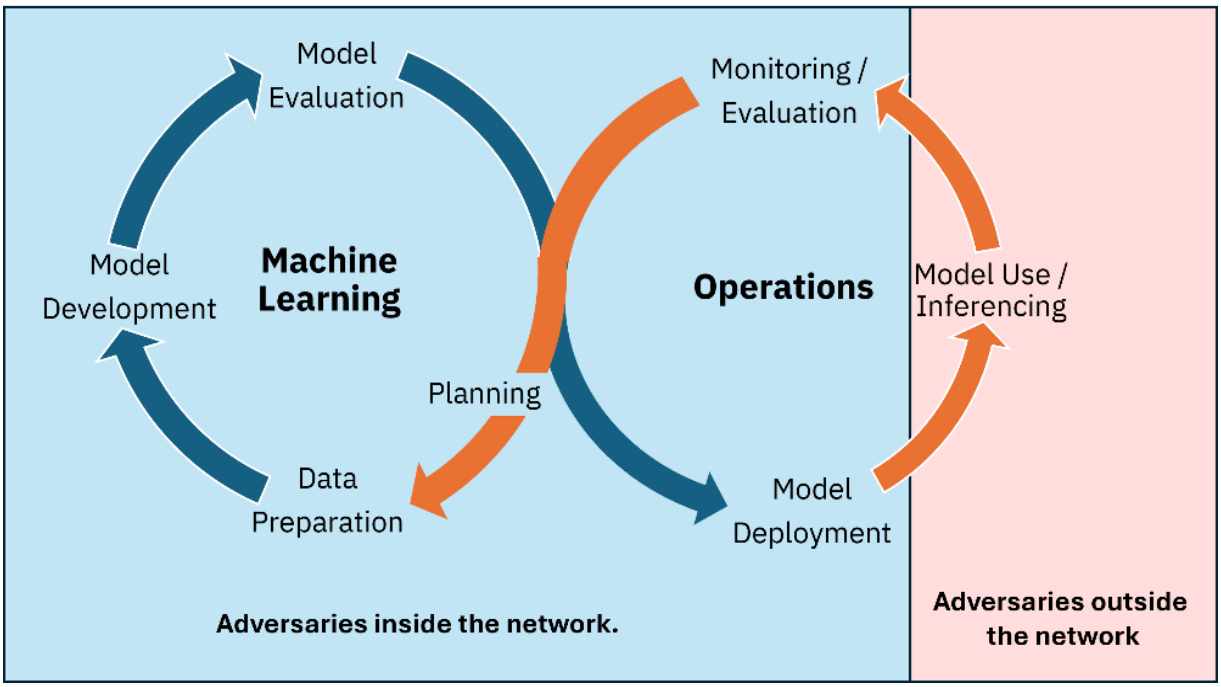

Figure 3 – Activities available inside (left) and outside (right) the network.

Adversary Inside the Network	Adversary Outside the Network
Once an adversary is inside the network, the level of harm they can cause is directly related to the level of access and the extent of knowledge they have about the AI system, related processes and artifacts. The adversary can affect at least three of the lifecycle activities, namely, "Data preparation," "Model Development," and "Model Evaluation."	The only way the adversary can access to impact the AI system is through the "Model Use/Inferencing" activity.
Examples	**Examples**[14]
Data poisoning, by injecting intentionally corrupted, false, misleading, or incorrect samples into the training or fine-tuning datasets.	Evasion attack that makes a model output incorrect results by slightly perturbing the input data that is sent to the trained model.
Intentionally modifying algorithms, data, or sensors to cause AI systems to behave in a way that is harmful to the infrastructure they serve.	Prompt injection attack on a generative AI model to produce unexpected output by manipulating the structure, instructions, or information contained in the prompt.
Manipulating benchmarks for evaluation.	Interruption/denial of service attack by flooding the system with input requests.
Theft of confidential or sensitive critical infrastructure data from AI systems and other supporting systems.	Model inversion and extraction attack to steal the training data or parameters of a model, or reverse engineer the functionality of a model.
	Jailbreaking attack to break through the guardrails that are established by the model owners to perform restricted actions.
	Prompt injections to reveal sensitive information from the IT systems.

Table 4 - List of Adversary Inside and Outside the Network Scenarios

4.3.3 Failures in AI Design and Implementation

Any IT system is affected by the quality and thoroughness of the underlying engineering development process. However, AI systems introduce additional, if not novel, responsibilities for the system owners. Unlike traditional software, these systems possess two new characteristics: (i) ML from large amounts of training data, which if not appropriately managed, may raise questions about data quality, data rights, and provenance among other issues, that can affect the quality of the end application; and (ii) the non-deterministic nature of AI model outputs.[15] These characteristics, if not anticipated and managed appropriately, could introduce engineering challenges that may manifest in unexpected ways to make the system vulnerable. Examples under this new class of risk include:

[14] *See* IBM, *AI Risk Atlas*, May 29. 2025, https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas.
[15] The non-deterministic nature of AI model outputs refers to the fact that AI systems, particularly those using machine learning and neural networks, can produce different results when given the same input under slightly varying conditions. Unlike traditional deterministic algorithms, which always yield the same output for a given input, AI models operate with probabilities, randomness, and learned patterns, making their behavior less predictable. Factors such as model architecture, training data variability, stochastic processes (like dropout in neural networks), and even hardware differences can contribute to this unpredictability. As a result, AI-generated responses or predictions may vary between runs, reflecting the inherent complexity and adaptability of such systems.

- **Brittleness**: Unintended/unexpected behavior of AI systems when confronted with circumstances outside of their original problem context.
- **Inscrutability:** Limited transparency and inherent uncertainties in AI systems that make diagnosing and correcting AI system anomalies difficult.
- **Statistical Bias:** The reproduction or amplification of biases in the training data and algorithms leading to erroneous decision-making.
- **Inconsistent System Maintenance:** Failure to regularly update and maintain AI models and supporting systems, potentially leading to malfunction of service disruptions.
- **Over/under reliance on AI:** Inadequate human oversight or under-utilization of AI.

Since these risks are the results of failures in AI design and implementation, and not due to any specific adversarial action, their consequences and the related risks will depend on the specific context/use case.

4.3.4 Generic AI Threats & Mitigations

Many threats to AI systems and associated mitigations can be considered generic and applicable across a broad range of use cases and industries where AI is leveraged. An overview of these generic threats and mitigations related to different phases of the AI Model Lifecycle is described in Appendix B.

5 AI in Telecommunications Networks

Artificial intelligence is rapidly transforming telecommunications networks, particularly within 5G systems, where automation, optimization, and predictive analytics play crucial roles in enhancing efficiency and performance. Operators are implementing AI solutions to manage complex network functions, including traffic prediction, dynamic resource allocation, anomaly detection, and automated troubleshooting, ensuring more reliable and adaptive connectivity. Recognizing AI's growing influence, standards organizations such as the 3GPP, International Telecommunication Union (ITU), and European Telecommunications Standards Institute (ETSI) are developing frameworks to integrate AI-driven intelligence into network architecture, security protocols, and service orchestration, laying the groundwork for more autonomous and resilient communications systems.

The drivers of AI implementation in 5G include rising data demands, the need for real-time network adaptability, and the pursuit of cost-effective operations, while commercial deployment is fueled by advances in edge computing, ML models, and AI-powered network slicing, shaping the future of next-generation connectivity. Below, we describe the use of AI in key areas of 5G systems and the potential use of AI in 6G networks.

5.1 5G Systems

The 5G system architecture is a significant evolution of previous mobile network architectures designed to support a wide range of applications and use cases, such as enhanced mobile broadband and network slicing, to meet the needs of diverse industries with improved performance, flexibility, and efficiency. The 5G architecture consists of multiple domains, including User Equipment (UE), Radio Access Network (RAN), Backhaul, the Core Network, and the Operations Support System (OSS). 5G Systems, as with mobile networks generally, are typically deployed in highly secured environments isolated from the internet with strong security controls on internal- and external-facing interfaces. The use of AI in these domains is explored further in the following sections.

The fields of AI and telecommunications are extremely broad and complex; the complexity increases manifold when understanding how these fields intersect. An exhaustive approach to enumerate every threat posed by AI/ML when it is used in telecommunications network is impractical because the landscape of both technologies is rapidly evolving, with new use cases, architectures, and vulnerabilities emerging continuously. Therefore, in this report, for each 5G network domain, we first identify a representative set of use cases or core technologies that characterize the use of AI/ML in that domain. These use cases serve as a foundation for assessing security risks specific to that 5G network domain, and to derive mitigation strategies and subsequent FCC recommendations specific to the network domain.

5.1.1 User Equipment

A modern smartphone user interface (UI) provides users with an intuitive and engaging experience. Overall, mobile operating system (OS) developers and handset manufacturers aim to provide an efficient, personalized, and immersive user experience, through a variety of technologies that are ever-evolving and improving. UE raises security considerations that should be factored into both consumer-level and network-level risks.

5.1.1.1 AI Use in Handset UI Integration

AI has become an integral part of the modern smartphone UI, leveraging natural language processing and computer vision to perform tasks more efficiently and intelligently. For example, AI systems have been integrated into handsets to enhance numerous security features such as facial recognition, fingerprint scanning, and anomaly detection to protect against unauthorized access. As further examples, AI models can aid in smart health monitoring[16] and agent-based applications afford real-time, language-based control to open maps and locate what users are looking for.[17]

Recently, handset UIs have been enhanced with LLMs running on cloud-based servers as well as Small Language Models (SLMs) running directly on the handset.[18] SLMs are designed specifically to run in a resource constrained environment such as a mobile phone, they are computationally efficient and cost-effective alternatives to LLMs often trained for specific domains. Both LLMs and SLMs are advanced models that enable more intuitive and natural interactions, allowing users to communicate with their devices in a conversational manner. AI agents can automate tasks, such as sending messages and making calls or even managing calendar events.[19]

5.1.1.2 Risks Associated with AI Handset Integration

These AI integrations can introduce a variety of security risks due to AI's low-level access to both hardware and the OS.[20] AI agents leverage ML to execute tasks on the user's behalf. For instance, after receiving a phone call, the AI can transcribe the conversation, identify required actions, such as sending information to the caller, and then perform tasks such as drafting and sending an email with a requested file. Multiple AI agents can exist on a single device and use a variety of ML models, coordinating to complete tasks. However, without the right guardrails in place, the autonomy of AI agents introduces risks, particularly as they can expand the permissions and access given to AI and L/SLM systems. The

[16] A.V.L.N. Sujith, et al., *Systematic review of smart health monitoring using deep learning and Artificial intelligence,* Neuroscience Informatics 2, no. 3:100028 (2022).
[17] Saikat Basu, Lifehacker, *You Can Now Use Gemini to Navigate with Google Maps Hands-Free*, Apr. 2, 2024, https://lifehacker.com/tech/gemini-google-maps-integration.
[18] *See, e.g.*, Meet Prajapati, Tech Holding, *On Device LLM Processing in Android*, June 24, 2024, https://techholding.co/blog/on-device-llm-processing-android/; *see also* Simone Lini, *The AI race might be about the UI Layer, not the LLMs*, Feb. 3, 2025, https://www.linkedin.com/pulse/ai-race-might-ui-layer-llms-simone-lini-rzc5f.
[19] Hao Wen, et al., *AutoDroid: LLM-powered Task Automation in Android*, ACM Mobicom '24 (Sept. 30—Oct. 4, 2024), https://arxiv.org/pdf/2308.15272 (2024).

security vulnerabilities that may arise from operating LLMs and SLMs on mobile devices closely mirror those found in datacenter-hosted deployments, including risks such as jailbreak attempts, prompt injection, and model extraction. However, certain risks can be unique in the context of UE. For example, UE risks include unauthorized access to applications and data that are locally stored, which is why appropriate access controls are an important aspect of a reasonable security approach, as discussed below.

Next, we present two use cases of SLMs operating on the mobile device. The first use case examines risks inherent in SLMs running on mobile devices, while the second one considers vulnerabilities arising from hardware-level threats.

a. Task Automation Using LLMs

The integration of LLMs with mobile devices provide the language understanding and reasoning capabilities for task preparation, comprehension, and execution by a unified language model leveraging multiple modalities such as text, voice and even video.[21] SLMs have demonstrated unique abilities of instruction following and step-by-step reasoning. Although these abilities do not make them general-purpose task solvers, their programmability is enhanced when coupled with other advances occurring in the field, such as agentic AI and protocols such as Model Context Protocol.[22] As a result, SLMs are increasingly able to execute user-specified tasks through dynamic interactions with smartphones, for example, translating natural language commands into Graphical User Interface (GUI) application operations, thus allowing the SLM to act as a multimodal "exploitation" bridge between an application on the mobile device and the user, based fundamentally on the SLM natural language processing. Under such an arrangement, a user can simply issue a command (a task) -- "*Record the following message and send it to John Csric as an email attachment"*-- causing the SLM to interface with distinct applications (the recording application and the email application) on the mobile phone.

The flexibility derived from the SLM coordinating such tasks must be weighed carefully against the access permissions the model will require to interact with the applications and data stored on the mobile device. Furthermore, this area is evolving rapidly, with new security risks arising when new models and supporting code are released. Indeed, newer L/SLM models may raise novel emulation attack threats, whereby an attacker mimics the behavior of legitimate users, devices, or software to bypass security measures, which are being factored into overall security approaches.

b. Edge LLM Impacts on Device Performance

The continuous operation of large AI models, in the absence of cloud server dependence, could result in increased heat output and higher power consumption of a device's Central Processing Units (CPU). Consequently, this elevated heat can alter the thermal environment, particularly if cooling systems or other electronic components are affected. These thermal changes might indirectly impact digital processing and lead to variations in radio frequency signal behavior that could increase "exploitation

[21] *See, e.g.*, Hao Wen, et al., *AutoDroid: LLM-powered task automation in Android,* ACM MobiCom '24: Proceedings of the 30th Annual International Conference on Mobile Computing and Networking, 543-557 (2024); *see also* Liangtai Sun, et al., *META-GUI: Towards Multi-modal Conversational Agents on Mobile GUI,* Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 6699-6712 (2022*).*

[22] Github, *Introduction: Get started with the Model Context Protocol (MCP)*, https://modelcontextprotocol.io/introduction (last visited June 2, 2025). The Model Context Protocol is an open standard designed to help AI models integrate with external tools, systems, and data sources in a structured way.

surfaces."

5.1.1.3 Mitigation Strategies and Recommendations

The mitigation strategies to protect against the risks associated with LLMs are complex. This complexity results in cloud-hosted LLMs using IT departments to mitigate the risks of LLMs fabricating responses ("hallucinate"),[23] being susceptible to circumvention ("jailbreak"),[24] and divulging privacy of the user's past interaction with the LLM.

Importantly, all stakeholders in the mobile ecosystem have a role to play to mitigate risk, including mobile OS, application developers and providers, app store vendors, OEMs, operators, and end users. For example:

- Mobile OS and application developers should ensure that SLM-powered applications request only essential permissions, meaning denying a request to grant permissions to access the contact list on the phone if the task does not require such access.

- Mobile OS and application developers should enforce containerization by distributing SLMs that operate in a sandbox.[25] Users should be encouraged to prefer the download of such containerized distributions.

- When choosing which SLM to download from an application store, vendors should provide transparent disclosures and warnings to help non-expert users make informed consent.

- Application providers should ensure that the SLM is not unintentionally active when the device is not being actively used, and that the SLM cannot take certain, potentially risky actions until after the device has authenticated the user.

It is infeasible to expect any one stakeholder to independently implement or manage the comprehensive security measures required for these systems. And in considering a comprehensive approach to risk mitigation, it is important to factor in actions and efforts that may undermine security, including end users circumventing app store safeguards. The mitigation strategies, therefore, are inherently constrained by the weakest link.

<u>Recommendations</u>

- The Commission should consider creating consumer guidance addressing the use of SLMs in coordination with smartphone end user devices. This guidance could, for example, be shared through online documentation (e.g., blog posts), or information sessions at local libraries.

- The Commission should collaborate with operators and handset manufacturers, as well as federal partners, to facilitate the development of best practices to ensure that AI-based UI includes appropriate security controls (e.g., access controls, data protection measures, and threat detection and mitigation measures).

---

[23] Lei Huang, et al., *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions,* ACM Transactions on Information Systems, Volume 43, Issue 2, 1-55 (2025).
[24] Alexander Weir, et al., *Jailbroken: how does LLM safety training fail?*, NIPS '23: proceedings of the 37th International Conference on Neural Information Processing Systems, 80079-80110 (2023).
[25] Containerization refers to creating an isolated setup that protects user data, prevents unauthorized access, and minimizes resource usage so the model can function smoothly without slowing down the device or interfering with other apps.

- The Commission should collaborate with handset manufacturers and federal partners to reinforce the importance of considering that "always on" SLMs can impact the CPU, generate excess heat, and possibly alter spectral patterns.

- The Commission should collaborate with industry and federal partners to develop best practices regarding access to LLM models, including potential limitations on access to SLMs before a device has authenticated the user.

- The Commission should collaborate with federal partners to promote the benefits of manufacturers adopting the security measures outlined in the DHS guidelines and the NIST AI Risk Management Framework.

5.1.2 Radio Access Network

The 5G RAN is a critical component of the overall 5G architecture, responsible for connecting mobile devices to the core network, managing radio resources, overseeing data throughput, and optimizing network performance. The rapid evolution of mobile networks from 4G to 5G has transformed the RAN landscape, bringing unprecedented speed, capacity, and flexibility.[26] Along with these advances, there is a growing integration of AI in RAN operations; for example, dynamic spectrum management (DSM) and self-optimizing network functions. To ensure AI in 5G is trustworthy, AI-driven network functions must be reliable, safe, secure, and transparent in their behavior.

3GPP is the leading organization defining global 4G and 5G RAN standards, collaborating across international telecommunications bodies to develop technical specifications.[27] Traditional RAN architectures in 4G and early 5G deployments were largely vendor-specific and vertically integrated, with base stations provided as closed systems, limiting interoperability and flexibility. To address this, the industry moved toward more open and disaggregated RAN models, such as Centralized RAN (C-RAN) and virtualized RAN (vRAN). This evolution culminated in O-RAN, which decouples radio hardware from baseband processing, enabling modular components, described below in Table 5, to communicate over standardized interfaces. O-RAN allows operators to mix and match vendors, fostering competition and innovation while reducing reliance on single-vendor solutions.[28] O-RAN enhances vendor diversity and operational efficiency, but it also introduces security and interoperability challenges due to the increased complexity of multi-vendor networks. O-RAN integrates AI-driven intelligence through the RAN Intelligent Controller (RIC), which is designed to optimize radio resource management and

[26] Mobile Radio Access Network (RAN) standards have progressed through successive generations, with 4G (LTE) and 5G (NR) marking major leaps in capability. 4G LTE, first defined in 3GPP Release 8 (2008), marked a departure from 3G by adopting an all-IP architecture and leveraging advances in digital modulation, such as orthogonal frequency-division multiple access OFDMA, and radio link methods, such as multiple-input and multiple-output (MIMO), for improved speed and latency. LTE-Advanced, introduced in Release 10 (2011), brought enhancements like carrier aggregation and higher-order MIMO, eventually leading to gigabit speeds with LTE-Advanced Pro. The first 5G RAN standard, 5G New Radio (NR), arrived with Release 15 (2018), introducing a flexible air interface, wider bandwidths, including millimeter-wave frequencies, and massive MIMO to support emerging use cases like autonomous systems and industrial automation. Later releases have continued refining 5G with ultra-reliable low-latency communication (URLLC) and massive IoT connectivity.

[27] Through iterative Releases, such as LTE in Release 8 and 5G NR in Release 15, 3GPP has guided key generational advancements in mobile technology. Beyond air interface protocols, it defines entire RAN architectures and standardizes interfaces like S1, X2, NG, and Xn, ensuring interoperability among equipment from different vendors and supporting global roaming. Recognized by ITU-R, 3GPP's specifications form the foundation of IMT-Advanced (4G) and IMT-2020 (5G), fostering innovation, compatibility, and security across networks worldwide.

[28] The O-RAN Alliance, formed in 2018, has been instrumental in defining Open RAN specifications, including the 7-2x fronthaul split for interoperability between RUs and DUs.

interference mitigation.

RAN Component	Function
Centralized Unit (CU)	Handles the higher-layer functions of the RAN, such as control-plane signaling, user-plane data processing, and session management. It is typically located farther from the cell site and can serve multiple DUs. The CU supports extensive data processing capabilities, handling tasks like mobility management, radio resource control, and packet scheduling.
Distributed Unit (DU)	Manages lower-layer functions including real-time data processing, scheduling, and transmission tasks. Located closer to the end user, the DU ensures low-latency, real-time communication. It connects to the RU (which handles radio signal transmission/reception). In a 5G base station (gNodeB), the DU works in tandem with the RU to provide wireless access. The DU's responsibilities include managing UE connections, mobility, scheduling data transmissions, and supporting various radio technologies (e.g., LTE and 5G NR).
Radio Unit (RU)	Responsible for the RF front-end: it manages the transmission and reception of radio waves. The RU converts digital signals from the DU into analog radio signals and transmits them over the air interface to UEs, and vice versa.
RAN Intelligent Controller (RIC)	Provides an intelligent control layer to optimize the RAN in real time using advanced algorithms (including AI/ML). The RIC is split into two components: a Near-Real-Time RIC (handles real-time adjustments to radio resources) and a Non-Real-Time RIC (handles longer-term, higher-level network optimization). The RIC hosts specialized applications -- xApps on the near-real-time RIC for immediate, time-sensitive control tasks (e.g., optimizing radio resource management, handling mobility) and rApps on the non-real-time RIC for strategic, non-time-critical tasks (e.g., traffic forecasting, network planning). These applications enable the RAN to autonomously adapt to changing network conditions, improving capacity, coverage, and user experience while reducing operational costs.

Table 5 - 5G Radio Access Network Components

In addition to these core components, 5G RAN leverages several advanced radio technologies to improve coverage, capacity, and spectral efficiency:

- **Beamforming**: Focuses radio signals on targeted directions for optimal coverage and capacity, rather than broadcasting uniformly in all directions.
- **Massive MIMO (Multiple Input Multiple Output)**: Enhances the network's capacity and throughput by using arrays of multiple antennas at the base station to transmit and receive more data streams simultaneously.

- **Dynamic Spectrum Sharing (DSS)**: Allows operators to allocate and share spectrum resources dynamically between different generations of mobile networks (e.g., 4G LTE and 5G NR) to improve overall spectrum efficiency.

5.1.2.1 AI Use in 5G RAN

AI's role in the RAN ecosystem is not new as Self-Optimizing or Self-Organizing Networks (SON) technology has incorporated elements of ML for many years. SON is a technology concerned with RAN configuration automation and resource utilization optimization and is typically deployed in the Network Management layer to manage RAN deployments. AI is poised to expand rapidly in RAN implementations, driven by initiatives in industry and standards bodies, including 3GPP, O-RAN Alliance, and ETSI. Government-led modernization efforts, such as the National Spectrum Strategy[29] and the National Spectrum Research and Development Plan (2024),[30] have likewise encouraged incorporating AI in search of spectrum efficiencies and productivity.

This section focuses on two priority use cases of AI in 5G RAN: AI for Energy Savings and AI for DSM/DSS. For each use case, we outline how AI is applied and then examine vulnerabilities, potential attacks, and mitigations, referencing established industry frameworks (e.g., Open Web Application Security Project (OWASP) guidelines and O-RAN Alliance AI security recommendations) where applicable. Applications of AI in the context of DSM/DSS and related uses are further discussed next. The discussions directly address relevant issues closer to the user interface and the impact on wireless access and spectrum management for energy optimization applications.

Energy Savings. Powering RAN infrastructure is energy-intensive, so operators are adopting AI-driven strategies for adaptive energy management. One key application is cell sleep and activation: AI algorithms monitor network demand patterns and can temporarily shut down underutilized cells or equipment during off-peak hours, reactivating them as demand increases. Another AI-driven strategy is beamforming optimization, which dynamically adjusts antenna beam directions and power; by steering signals more efficiently and lowering transmission power, when possible, the system can reduce energy consumption while maintaining service quality. Additionally, AI optimizes power control, continually tweaking the transmit power of cells to ensure adequate coverage and capacity with minimal waste.

Energy management techniques introduce new security concerns as they are potentially vulnerable to malicious intervention—for example, an attacker could conduct data poisoning on the AI's inputs (e.g., feeding false network load or sensor data) to trick the system into erroneously deactivating critical cells at busy times or keeping redundant cells active, leading to service disruptions or wasted energy. Generally, direct AI-focused attacks on RAN energy-saving features are less likely than broader network attacks, since they would require targeting specific AI models; nonetheless, the possibility exists and must be guarded against in any AI-managed energy optimization system.

DSM/DSS. Efficient use of radio spectrum is another crucial area where AI is applied in 5G RAN. AI-driven DSM allows the network to intelligently allocate frequencies in real time; for example, in the Citizens Broadband Radio Service (CBRS) band where commercial 5G systems share spectrum with government incumbents and with each other. AI/ML models can analyze interference and usage patterns to dynamically assign spectrum channels to cells or users, optimizing utilization while avoiding conflicts.

---

[29] *See* NTIA, *National Spectrum Strategy Implementation Plan*, Mar. 12, 2024, https://www.ntia.gov/sites/default/files/publications/national-spectrum-strategy-implementation-plan.pdf (seeking recommendations for potential investment based on assessment of smart spectrum management technologies including artificial intelligence and machine learning).

[30] Wireless Spectrum Research and Development Interagency Working Group, Networking and Information Technology Research and Development Subcommittee of the National Science and Technology Council, *National Spectrum Research and Development Plan*, Oct. 2024, at 13, https://www.nitrd.gov/pubs/National-Spectrum-RD-Plan-2024.pdf.

This real-time spectrum allocation helps maximize network throughput and efficiency by making the most of available spectrum resources.

Another important aspect of AI-driven DSM/DSS is incumbent protection. The AI must ensure that primary users of spectrum -- military or government communications in shared bands -- are not adversely affected when secondary users -- commercial operators -- access the band. This involves sensing or predicting when and where incumbents are active and adjusting wireless transmissions accordingly, maintaining fairness and preventing critical communication disruptions. Additionally, AI facilitates Multiple Radio Access Technology (multi-RAT) coexistence – orchestrating multi-RAT (e.g., 5G, LTE, and Wi-Fi operating in overlapping frequencies) so that they can coexist with minimal interference. By coordinating across technologies, AI can help the RAN optimize overall throughput while respecting each technology's interference constraints.

Having introduced these use cases, the next section examines specific AI-related vulnerabilities in the 5G RAN, focusing on how they impact the energy savings and spectrum management scenarios discussed above.

5.1.2.2 Risks Associated with AI in 5G RAN

Our analysis categorizes AI-related vulnerabilities in the 5G RAN into two broad groups. The first category involves attacks that leverage AI (using AI as a tool to enhance attacks on the network). These include threats such as AI-assisted phishing campaigns, realistic voice or image impersonation of operators/administrators, or automated exploit discovery using AI. Such attacks are typically general cyber-attack vectors that could lead to credential theft or infiltration of an operator's network environment; they are serious, but notably they do not specifically target the AI control loops within the RAN.

The second category consists of attacks targeting the AI systems in the RAN. These are more direct threats to the AI components themselves; for example, data poisoning, adversarial inputs, model evasion techniques, or "model-in-the-middle" manipulations. These attacks are particularly critical in the RAN context because they focus on exploiting the AI-driven elements of the RAN (such as the O-RAN RIC and its xApps/rApps) that govern key functions like resource allocation and energy management. Due to the large range of existing RAN related technologies that exist (many of which are proprietary) this report uses Open RAN (O-RAN) in places simply to highlight examples of the kinds of threats to be considered within the RAN.

Below, we expand on several key AI vulnerabilities that are especially relevant to the 5G RAN energy saving and DSM use cases:

Data Poisoning/Model Manipulation. In a RAN context, data poisoning occurs when an attacker injects malicious or misleading data (such as false telemetry readings or manipulated sensor inputs) into the AI's training or operational data stream, thus corrupting either the models themselves or the output from the models, respectively. This can corrupt the model's understanding of network conditions, causing it to make harmful decisions. For instance, in the energy-saving scenario, poisoned data could cause the AI system model to incorrectly predict low traffic and deactivate essential cells, wasting energy or degrading Quality of Service (QoS) for users. In a DSM/DSS scenario, falsified spectrum occupancy data might mislead the AI to improperly assign spectrum in shared bands, potentially creating interference or violating incumbent protections by transmitting when and where it shouldn't.

Adversarial Inputs. Adversarial input attacks involve crafting specialized input signals or traffic patterns that exploit the AI model's vulnerabilities and force misclassification of conditions.[31] In terms of energy management, an adversary could generate traffic patterns that confuse an AI model overseeing beamforming or power control, causing it to either under-provision (leading to capacity shortfalls) or over-provision resources (wasting energy). In spectrum management, adversarial signals might trick an AI's interference detector into overlooking unauthorized transmissions or erroneously blocking legitimate ones. The result can be network disruptions or improper enforcement of spectrum sharing policies.

Model Evasion and Reverse Engineering. These attacks occur when an adversary probes an AI system (via repeated queries or observations of its decisions) to learn about its internal model and decision boundaries. By reverse-engineering aspects of the model, the attacker can then craft inputs that systematically evade detection or exploit the model's blind spots. In the energy saving use case, if attackers discern how the AI decides to activate or deactivate cells, they could time their activities to avoid triggering energy-saving mode, forcing the network to stay in a high-power state or to deliberately trigger it at the wrong times. In a DSM/DSS context, understanding the model's spectrum allocation strategy could allow an attacker to predict and occupy frequencies that the AI believes are free, thus gaining unauthorized spectrum access or causing interference while avoiding the AI's detection mechanisms. This leads to unauthorized spectrum usage, network interference, and potential violation of regulations.

System Control and Automation Dependency. This vulnerability refers to the heavy reliance on automated, AI-driven control loops in an O-RAN architecture. If an attacker gains unauthorized access to the RIC or its applications (xApps for near-real-time control and rApps for non-real-time operations), they can manipulate or disrupt these control loops. Such access could be obtained via exploiting software vulnerabilities or insufficient access controls. The consequences are severe: in the energy domain, an intruder could send incorrect control signals or disable certain automation routines, resulting in large-scale misconfiguration of cell power states and significant energy waste. In the spectrum domain, an attacker controlling the RIC could improperly reassign frequencies or disable incumbent protection mechanisms, immediately causing network congestion, harmful interference, or regulatory violations. Essentially, the automation that normally improves efficiency becomes a single point of failure if compromised.

Organizational or Process Vulnerabilities. Not all vulnerabilities are technical; some stem from how AI systems are managed. Without proper human oversight, rigorous testing, and incident response processes, the deployment of AI in RAN can introduce risk. For example, if a new AI model or an update to an xApp/rApp is rushed into production without sufficient quality assurance, it might contain errors. A misconfigured model for energy management could inadvertently keep thousands of cells in an active state unnecessarily, causing massive energy wastage across the network.

In the spectrum context, insufficient testing of a DSM algorithm could lead to uncoordinated frequency shifts or channel selections that conflict with policy constraints. Poor change management or lack of a rollback plan can exacerbate these issues. Thus, gaps in processes – for example, inadequate validation of AI models or lack of a clear incident response plan for AI-related failures—can turn into security and reliability vulnerabilities in their own right.

[31] *See, e.g.,* Brian Kim, Yi Shi, et al., *Adversarial Attacks against Deep Learning Based Power Control in Wireless Communications*, Oct. 2021, https://arxiv.org/pdf/2109.08139; Yalin E. Sagduyu, Yi Shi, et al., *Adversarial Deep Learning for Over-the-Air Spectrum Poisoning Attacks*, Nov. 2019, https://arxiv.org/pdf/1911.00500; Zikin Liu, Changming Xu, et al., *Exploring Practical Vulnerabilities of Machine Learning-based Wireless Systems*, April 2023, https://deepakv.web.illinois.edu/assets/papers/rafa_nsdi_23.pdf.

xApps, rApps, and RIC Vulnerabilities

In the O-RAN architecture, the separation of concerns between xApps (extensible applications on the near-real-time RIC) and rApps (radio applications on the non-real-time RIC) creates distinct attack surfaces. xApps, which handle time-sensitive tasks such as scheduling, handovers, and beamforming adjustments, could be targeted with fast-acting adversarial attacks. For instance, if an attacker introduces an adversarial perturbation into the data stream that a scheduling xApp uses, it might cause immediate degradation in network performance - dropped calls, increased latency, or coverage holes. Because xApps influence the network almost instantaneously, any compromise can have immediate and noticeable impact. On the other hand, rApps handle longer-term functions, such as policy enforcement, performance analytics, or capacity and spectrum planning. An attack on an rApp might be less immediately obvious but can introduce slow-building issues; for example, feeding a planning rApp subtly poisoned data over time could lead it to make suboptimal strategic decisions, such as misallocating spectrum for weeks or misplanning network capacity, which accumulate into significant inefficiencies or regulatory non-compliance. In short, xApps are vulnerable to real-time disruption and rApps to strategic manipulation, and both layers need robust protection.

5.1.2.3 Mitigation Strategies and Recommendations

To address the above vulnerabilities, stakeholders should adopt a multi-layered security strategy grounded in zero trust,[32] combining technical safeguards, strong governance, and thorough testing and oversight. The following recommendations draw on guidance from the O-RAN Alliance, ETSI, and industry best practices:

Technical Safeguards. Technical safeguards for AI systems focus on securing the entire pipeline and ensuring robust adversarial defenses.

- **Secure AI Pipeline**: Ensure the integrity of data and models throughout the AI lifecycle. This includes implementing data integrity checks and filters to detect anomalies in real-time telemetry (preventing data poisoning attempts) and maintaining strict model and test/training data provenance. All AI models should be cryptographically signed and version-controlled so that only verified, approved models are deployed. New or updated functionality (e.g., xApps and rApps) should be thoroughly tested in isolated sandbox environments before live rollout to catch any malicious or errant behavior.

- **Adversarial Robustness Testing**: Subject AI models to rigorous adversarial testing and hardening. This can involve adversarial training, where the model is trained on examples of potential attacks to improve resilience, as well as stress testing and fuzz testing of the AI-driven control loops.[33] Continuous validation in production is also important. Monitor the AI's outputs for signs of concept drift or unusual decisions that might indicate a subtle attack. By proactively probing the models with simulated attacks and edge-case scenarios, operators can identify and fix vulnerabilities before they are exploited.

- **Encryption and Access Control Mechanisms**: Protect the RAN's AI systems and interfaces through strong encryption and access management. All communications with the RAN, including internal RAN subsystems (e.g., with RIC and xApps/rApps), should be encrypted in transit, and models should be encrypted at rest to prevent unauthorized extraction or reverse

[32] *See* Scott Rose. Oliver Borchert, et al., NIST Special Publication, *Zero Trust Architecture*, Aug. 2020, https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-207.pdf.
[33] Fuzz testing (or fuzzing) is a type of software testing that involves providing invalid, unexpected, or random data inputs to a program to identify vulnerabilities, crashes, or unexpected behaviors. The goal is to detect bugs, memory leaks, security flaws, and other issues that might not be easily found through traditional testing methods.

engineering. Access to the RIC's management interface and the deployment of xApps/rApps must be tightly controlled under the principle of least privilege: each component or user gets the minimum access necessary and no more. Employing a zero trust security model is advisable, wherein every request or data feed into the AI system is authenticated and verified, even if it originates from within the network. This ensures that malicious actors cannot easily insert themselves into the control loop or tamper with AI inputs/outputs.

Governance and Policy Frameworks. The governance and policy frameworks for AI-driven systems in RAN operations must ensure alignment with industry standards and robust operational oversight.

- **Alignment with Industry Standards**: Communications Service Providers and network operators should adhere to applicable, recognized security standards and frameworks specific to AI in telecommunications. For example, the O-RAN Alliance's security specifications include requirements and best practices for securing the RIC and AI-driven RAN functions. Likewise, frameworks from bodies such as ETSI (e.g., ETSI ENI for autonomic network management) and relevant 3GPP technical reports outline secure integration of AI in network architectures. Aligning internal policies with these industry guidelines helps ensure that the AI systems in RAN are designed and operated with security in mind from the start.[34]

- **Operational Oversight**: It is important to maintain human oversight of critical AI decisions, especially in the early stages of deployment or for high-stakes scenarios. Operators should consider implementing a "human-in-the-loop" approach whereby certain actions proposed by an AI (for instance, shutting down a large cluster of cells for energy saving or opening up a new spectrum channel in a dynamic sharing scenario) require human review or approval. Even when fully automated, there should be real-time dashboards and alerts for engineers to monitor AI behavior. Additionally, enforce strict change management for AI models; any update to an xApp/rApp or its algorithms should go through a review process, and the ability to quickly roll back to a previous stable version must be in place in case anomalies are detected post-deployment.

- **Risk Management and Incident Response**: Treat AI systems as critical infrastructure from a risk management perspective. This means regularly conducting threat modeling exercises focused on AI/ML aspects of the RAN (identifying what new threats emerge from using AI, how they could manifest, and what controls exist or need improvement). Based on these assessments, develop and refine incident response playbooks for AI-related incidents. For example, if an AI model is suspected of being compromised or behaving erratically, the response plan might include steps to isolate or shut down that model, reverting to manual control or a backup system, purging any suspect data inputs, and performing a forensic analysis on the AI model and its data to understand the breach. Having predefined procedures for scenarios like "AI model output is compromised" will greatly reduce response time and impact when an incident occurs.

Testing and Assurance. Testing and assurance for AI-based systems involves several critical areas.

- **Use Case-Specific Testing**: Beyond generic AI testing, validate AI behavior in the context of specific RAN use cases. For energy savings features, test that algorithms for cell sleeping or power scaling do not inadvertently degrade network QoS or, importantly, do not interfere with

[34] *See* ETSI, *Experiential Networked Intelligence (ENI) - Use Cases*, ETSI GS ENI 001/005, 2023, https://standards.globalspec.com/std/14609445/gs-eni-005; 3GPP, *Study on Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN*, TR 38.743, https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=4286.

critical services like emergency calls (e.g., ensure a cell serving an emergency call is not powered down by the AI). For DMS/DSS, test AI models in controlled lab environments that mimic real-world spectrum conditions, including simulated interference patterns, presence of incumbent signals, and multi-RAT scenarios operating together, to ensure the AI makes correct decisions and gracefully handles edge cases before it is trusted in the field.

- **Performance Benchmarks**: Establish baseline performance metrics for the RAN before and after introducing AI to quantitatively measure the AI's impact and detect anomalies. Key metrics might include energy consumption, spectral efficiency, network throughput, latency, and drop rates. Continuous monitoring should be deployed to watch these metrics in near-real time. The system can use automated counters and alarms to flag if, for instance, energy usage suddenly spikes or spectrum utilization drops unexpectedly, which might signal that an AI model is malfunctioning or under attack. Through benchmarking and ongoing monitoring, operators can quickly spot when AI behavior deviates from expected norms and investigate immediately.

- **Regulatory Compliance**: Ensure that AI-driven decisions stay within regulatory bounds at all times. This is especially critical for spectrum management in shared bands (e.g., CBRS) where regulations enforce protection for incumbents. The RAN must be configured so that no AI action can violate FCC or National Telecommunications and Information Administration (NTIA) rules (e.g., the system should prevent an AI from assigning frequencies that are off-limits or transmitting at powers higher than allowed). Regular audits of the AI's decisions against compliance checklists are prudent. Consideration should also be given to how AI can support regulatory compliance through enhanced explainability, potentially leveraging approaches such as neuro-symbolic logic.[35]

Taken together, these strategies form a robust, multi-layered defense for AI in the 5G RAN. By tailoring the above recommendations to their specific network context and use cases, operators can confidently harness AI to achieve a more efficient, flexible, and secure next-generation wireless ecosystem.

5.1.3 Backhaul

The 5G backhaul architecture is a critical component of the overall 5G mobile network, providing connectivity between the RAN and the core network. It serves as the intermediary for high-speed data transport, ensuring minimal latency, high throughput, and robust reliability required for the ultra-connected, low latency demands of 5G.

Component	Function
Physical Layer	The physical infrastructure includes fiber optic cables, microwave links, and millimeter-wave connections. Fiber remains the primary backbone for high-capacity transport, while wireless links provide more flexible and cost-effective solutions in areas where fiber installation is challenging or impractical.
Core Network Connection	The backhaul links connect the RAN to the core network, enabling data processing, routing, and service delivery.
Aggregation Networks	These are networks where multiple backhaul connections from different base stations converge. The aggregation points combine traffic before sending it to the core network.

[35] *See* ATIS, *Advancing Generative AI Implementation in Telecommunications Networks*, Nov. 2024, pp. 15-16, https://atis.org/resources/advancing-generative-ai-implementation-in-telecommunications-networks/.

Component	Function
Microwave and Millimeter-Wave Links	These high-frequency wireless communication systems provide high-capacity links over medium-to-long distances, offering low latency and high throughput.
Edge Computing Nodes	These nodes sit closer to the users, providing distributed data processing. They often need to communicate with the backhaul for real-time applications.
Network Slicing	5G supports the concept of network slicing, where a virtualized and customized network is created for specific use cases. Backhaul in such networks must dynamically manage different slices to ensure efficient and flexible traffic delivery.

Table 6 – Key Components of 5G Backhaul

5.1.3.1 AI Use in 5G Backhaul

AI is transforming how backhaul networks are managed, optimized, and evolved. Their capabilities can improve several areas within the backhaul infrastructure.

Traffic Optimization and Resource Allocation. AI algorithms can enhance traffic optimization and resource allocation in 5G networks by predicting traffic patterns based on historical data, user behavior, and current network conditions. These predictions allow for proactive adjustments, ensuring efficient traffic routing, minimizing congestion, and boosting overall throughput. Additionally, AI systems can dynamically allocate bandwidth in real-time based on demand, ensuring optimal use of available resources. This flexibility is especially important in 5G, where demand can fluctuate significantly, enabling efficient and responsive network management in highly dynamic environments.

Network Optimization. AI-driven SONs allow for the autonomous configuration and optimization of network parameters, reducing the need for human intervention and lowering operational costs. These algorithms can automatically adjust backhaul links to enhance performance and ensure efficient operation. Additionally, AI plays a role in latency management by monitoring network latency in real-time. It can predict and correct potential issues before they lead to significant delays, and by detecting anomalous latency patterns, AI systems can proactively reroute traffic to alternative paths, maintaining optimal performance and ensuring a seamless user experience.

Predictive Maintenance and Fault Detection. AI can enhance predictive maintenance and fault detection in 5G backhaul networks by analyzing both historical and real-time data from network components like routers, switches, and links. Through anomaly detection, AI can identify abnormal patterns, such as performance degradation, which may indicate potential hardware failures or network congestion. Additionally, AI-driven predictive fault management can forecast the likelihood of hardware failure, enabling preventative maintenance before issues arise, ultimately reducing downtime and improving overall network reliability.

Load Balancing and Traffic Prioritization. AI can automatically balance the load across backhaul links and optimize the routing of high-priority traffic, such as low-latency or mission-critical services, to ensure optimal user experience. By continuously analyzing current network conditions, AI ensures that traffic prioritization policies are dynamically enforced, maintaining performance and meeting the demands of critical services while preventing congestion and ensuring efficient network utilization.

Optimization of Hybrid Backhaul Networks. In 5G networks, backhaul often combines fiber and wireless technologies, and AI can play a role in optimizing this hybrid infrastructure. By intelligently selecting

the most optimal backhaul path based on factors such as network load, weather conditions (especially for wireless links), and distance, AI helps enhance both cost-efficiency and performance. This dynamic optimization ensures that the backhaul network can adapt to varying conditions, delivering reliable and efficient connectivity.

5.1.3.2 Risks Associated with AI in Backhaul

As 5G networks become more advanced, they also become more susceptible to a range of sophisticated cyberattacks, with AI playing a role in carrying out these attacks. Denial of Service (DoS) attacks, Man-in-the-Middle (MITM) attacks, malware and ransomware, misconfiguration vulnerabilities, and insider threats all pose significant risks to the security and reliability of the network. Malicious actors are increasingly leveraging AI to automate and optimize their attack strategies, making traditional security measures less effective and complicating the task of network defense.

- **Denial of Service Attacks**: DoS attacks overwhelm network resources by flooding them with excessive traffic, disrupting communication between 5G backhaul components such as gNodeBs (GNB) and the core network. The high-capacity nature of 5G networks makes them attractive targets for such attacks, which can lead to significant service degradation or outages. Malicious actors can leverage AI to orchestrate more sophisticated DoS attacks. AI algorithms can analyze network traffic patterns to identify optimal attack strategies, making the malicious traffic blend seamlessly with legitimate traffic, thereby evading traditional detection mechanisms. Additionally, AI can dynamically adjust attack parameters in real-time to counteract mitigation efforts.[36]

- **Man-in-the-Middle Attacks**: MITM attacks involve an adversary intercepting or altering communication between two parties without their knowledge. In the context of 5G backhaul, this could mean intercepting data between DUs and the core network, leading to data breaches or unauthorized command execution. AI can assist attackers in real-time analysis of intercepted data, enabling them to extract valuable information swiftly. Additionally, AI can help in automating the process of injecting malicious data into the communication stream, making MITM attacks more efficient and harder to detect.

- **Malware and Ransomware**: Malware and ransomware attacks involve malicious software infiltrating the network to steal data, disrupt operations, or encrypt critical information for ransom. In 5G backhaul networks, such attacks can compromise virtualized network functions, leading to widespread service disruptions. AI coding assistants lower the bar for low-skilled adversaries who wish to develop malware but would not otherwise have the requisite knowledge. Attackers can employ AI to develop malware capable of evading traditional detection methods by altering its behavior or appearance. AI can also enable malware to make autonomous decisions, such as selecting the most valuable data to encrypt or exfiltrate, increasing the attack's effectiveness.

- **Configuration and Misconfiguration Vulnerabilities**: Misconfigurations in network settings, such as open ports, weak authentication mechanisms, or incorrect routing rules, can create security gaps in the 5G backhaul infrastructure. These vulnerabilities can be exploited by attackers to gain unauthorized access or disrupt network operations. Malicious actors can use AI to automate the discovery of misconfigurations across large and complex network environments. AI can rapidly scan for vulnerabilities and identify the most exploitable ones, enabling attackers to launch targeted attacks with minimal effort.

[36] Chafika Benza et al., *AI for Beyond 5G Networks: A Cyber-Security Defense or Offense Enabler*, https://arxiv.org/pdf/2201.02730.

- **Insider Threats**: Every network provider has challenges with insider threats in multiple areas of their network. Insider Threats involve individuals within the organization, such as employees or contractors, who misuse their access to compromise network security. In the 5G backhaul context, insiders may intentionally or unintentionally cause data breaches, disrupt services, or facilitate external attacks. AI can help insiders automate processes like data exfiltration, lateral movement, and privilege escalation while evading traditional security controls. For example, ML algorithms can analyze system logs to identify gaps in monitoring, enabling attackers to avoid triggering alerts. Additionally, insiders could use AI to manipulate or tamper with system configurations or bypass security policies without raising suspicion. This combination of AI's adaptive capabilities and the insider's privileged access makes insider threats more dangerous and harder to detect.[37] Lastly, it is common for the credentials of an authorized administrator to be compromised through phishing attacks and then used to gain access into various parts of the network. These attacks look like an insider attack because they are coming from within, but the attacker is not an insider.

5.1.3.3 Mitigation Strategies and Recommendations

Denial of Service Attacks.

- **Traffic Analysis and Near Real-Time Anomaly Detection:** AI-driven anomaly detection models are particularly effective at identifying malicious traffic amidst legitimate traffic flows. Traditional DoS mitigation systems rely on static thresholds or predefined signatures, which are insufficient in a 5G environment where traffic volumes and patterns fluctuate dynamically. These algorithms establish baseline network behavior over time using unsupervised learning techniques and detect deviations such as sudden traffic spikes or abnormal patterns; for example, AI systems can distinguish between legitimate user-driven traffic and bot-generated traffic aimed at overwhelming the network. This real-time detection enables operators to throttle or block suspicious traffic effectively, minimizing disruptions. AI's precision reduces false positives that could otherwise disrupt legitimate services.

- **Predictive Threat Intelligence**: Using time-series analysis and historical data, AI can predict traffic surges and differentiate between legitimate user spikes (e.g., due to new 5G services) and malicious traffic. This capability enables telecom providers to anticipate potential threats and take proactive measures, such as pre-allocating resources or configuring defenses for high-risk periods. For instance, predictive models may identify patterns suggesting an impending attack, allowing security teams to deploy countermeasures before the attack materializes. This forward-looking approach enhances network resilience and minimizes downtime during attacks.[38]

- **Automated Mitigation**: AI-driven automation accelerates the response to Distributed Denial of Service/Telephony Denial of Service attacks by dynamically rerouting or filtering malicious traffic. Automated systems, such as AI-enabled traffic scrubbing centers, examine data packets in real-time, allowing legitimate traffic to flow uninterrupted while blocking harmful requests. This minimizes the manual intervention required during an attack and ensures that critical services remain available to users. AI-based solutions like Deep Packet Inspection (DPI) paired with ML models can analyze packets to identify whether traffic originates from known malicious sources or botnets. Over time, these systems learn and adapt to new attack techniques,

[37] *See, e.g.*, CISA, *Potential Threat Vectors to 5G Infrastructure*, 2021, https://www.cisa.gov/sites/default/files/publications/potential-threat-vectors-5G-infrastructure_508_v2_0%20%281%29.pdf; Chafika Benza et al., *AI for Beyond 5G Networks: A Cyber-Security Defense or Offense Enabler*, https://arxiv.org/pdf/2201.02730.
[38] Alex Pavlovic, *How AI/ML Can Thwart DDoS Attacks*, Dec. 20, 2022, https://www.darkreading.com/cyberattacks-data-breaches/how-ai-ml-can-thwart-ddos-attacks.

improving their effectiveness in mitigating evolving threats.[39]

MITM Attacks.

- **Real-Time Traffic Anomaly**: MITM attacks often exploit weak encryption, misconfigured backhaul components, or vulnerabilities in network protocols such as Transport Layer Security (TLS). These attacks can manipulate communication channels, altering latency, packet structure, or encryption handshakes to remain stealthy. AI and ML can use ML based behavioral analytic models to analyze network traffic in real time to baseline normal communication patterns. Sudden deviations, such as unusual packet latencies, unrecognized certificates, or unexpected rerouting of traffic, can trigger alerts. AI driven time series models, such as Long Short-Term Memory (LSTM) networks, can detect temporal anomalies in traffic flow that may signal unauthorized interception or packet alteration. Lastly, AI enhances deep packet inspection tools by analyzing packet contents and structure to detect injected malicious code, unauthorized alterations, or unexpected payload sizes. An example of AI-powered anomaly detection can identify changes to TLS handshakes or certificate anomalies, signaling an active MITM attack. An earlier CSRIC report found that securing traffic at the protocol layer is vital to preventing such threats.[40]

Malware and Ransomware.

- **Anomaly Detection:** By analyzing vast amounts of network traffic data, AI algorithms can identify anomalous patterns indicative of malicious activity, enabling early detection of threats that traditional signature-based methods might miss. For instance, AI-based ransomware detection frameworks utilize behavioral analysis to identify ransomware activities in real-time, allowing for prompt intervention to prevent data encryption and exfiltration.

- **Adaptive Security Measures:** AI can facilitate the development of adaptive security measures that evolve alongside emerging threats. In the context of 5G networks, where the integration of AIoT (Artificial Intelligence of Things) devices increases the attack surface, AI-driven security solutions can dynamically adjust to new vulnerabilities. For example, AI-powered intrusion detection systems in Multi-access Edge Computing (MEC) environments can detect and mitigate malware and ransomware attacks by analyzing device behavior and network traffic in real-time, ensuring the resilience of 5G backhaul infrastructures.[41]

Configuration and Misconfiguration Vulnerabilities.

- **Analysis of Network Configurations:** AI algorithms can analyze network configurations to identify anomalies or deviations from established baselines, allowing for the prompt identification of potential vulnerabilities. For instance, AI-driven systems can monitor network parameters and automatically adjust settings to optimize performance and security, reducing the risk of human error.[42] Furthermore, AI and ML facilitate predictive maintenance by analyzing

[39] Andrew Wooden, *Nokia Builds Europe's 'biggest anti-DDoS solution' for an IXP Environment*, Sept. 16, 2024, https://www.telecoms.com/security/nokia-builds-europe-s-biggest-anti-ddos-solution-for-an-ixp-environment.
[40] CSRIC VIII, *HTTP/2 Vulnerabilities and Mitigations* (2022), https://www.fcc.gov/sites/default/files/CSRIC8-Report-SecurityVulnerabilitiesMitigationsHTTP2-0623.pdf.
[41] *See* S. M. Cheng, B. K. Hong, et al.,, *Attack Detection and Mitigation in MEC-Enabled 5G Networks for AIoT*, IEEE Internet of Things Magazine, vol. 5, no. 3, pp. 76-81, Sept. 2022, https://ieeexplore.ieee.org/document/9945850; Lucy Colback, *Technology and Cyber Crime: How to Keep Out the Bad Guys*, Financial Times, July 3, 2024, https://www.ft.com/content/8a79ab25-c902-4110-bcb8-be2fd422f6bf.
[42] *See, e.g.*, Oral Mohan, *5G Network AI Models: Threats and Mitigations*, Nov. 15, 2024,

historical data to forecast potential configuration issues before they impact network performance. This proactive approach enables network operators to address vulnerabilities preemptively, ensuring the resilience and reliability of 5G backhaul infrastructures. By automating routine configuration tasks and continuously monitoring discrepancies, AI reduce the likelihood of misconfigurations that could be exploited by malicious actors.[43]

- Traffic analysis and near real-time anomaly detection should be considered to identify attackers performing remote port scanning with intention to identify vulnerabilities.

Insider Threats.

- **Behavior Deviation Detection:** Insider threats pose a critical vulnerability to the cellular backhaul infrastructure, where compromise can lead to data interception, service disruption, or unauthorized access deep into the core network. In the 5G environment, backhaul networks are more complex and distributed, making traditional monitoring methods less effective. AI offers transformative capabilities to detect and mitigate insider threats within backhaul systems. By establishing dynamic baselines for normal activities, such as expected traffic flows, access patterns to routers, switches, and edge computing nodes, AI systems can rapidly detect deviations that suggest malicious insider behavior or misuse of privileged credentials. This enables early identification of potential threats like unauthorized configuration changes, data exfiltration, or anomalous command executions targeting critical transport links.

- **Behavioral Analytics:** AI further enhances security by integrating behavioral analytics across multiple backhaul components, correlating access logs, traffic anomalies, and device interactions in real-time. In high-capacity hybrid backhaul environments, where fiber and wireless links coexist, AI can prioritize and automate responses to suspected insider threats, such as isolating compromised network slices or rerouting traffic away from suspicious nodes. By applying predictive risk models, telecom operators can identify individuals or assets with elevated insider risk before exploitation occurs. Leveraging AI-driven monitoring and mitigation strategies is essential for maintaining the integrity, reliability, and security of cellular backhaul networks that underpin 5G critical infrastructure.[44]

AI models are central to realizing the full potential of 5G backhaul networks by enhancing efficiency, scalability, and adaptability. However, their deployment introduces new challenges, particularly around security, model robustness, and the consequences of failure. By adopting strong mitigation strategies, such as continuous monitoring, data validation, and adversarial robustness, operators can ensure that AI-powered backhaul systems are secure, reliable, and effective in supporting the demands of 5G networks.

---

https://blog.checkpoint.com/artificial-intelligence/5g-network-ai-models-threats-and-mitigations/; Ericsson, *The AI Standard for 5G RAN: what it is, why it's needed, and how to get there*, Nov. 7, 2023, https://www.ericsson.com/en/blog/2023/11/ai-ml-5g-ran-3gpp.

[43] *See, e.g.*, Mohammed Nasser Al-Mhiqani *et al*, *A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations*, Applied Sciences 10, no. 15, 2020, https://www.mdpi.com/2076-3417/10/15/5208; 5G Americas, *How Generative AI Could Impact Network Planning, RAN Configuration, and Spectrum Management*, Mar. 2024, https://www.5gamericas.org/how-generative-ai-could-impact-network-planning-ran-configuration-and-spectrum-management/.

[44] Mohammed Nasser Al-Mhiqani *et al*, *A Review of Insider Threat Detection: Classification, Machine Learning Techniques, Datasets, Open Challenges, and Recommendations*, Applied Sciences 10, no. 15, 2020, https://www.mdpi.com/2076-3417/10/15/5208.

5.1.4 5G Core Network

The 5G Core Network (5GC or core) is an architecture defined by 3GPP that serves as the backbone of 5G communication systems.[45] Unlike its predecessors, the 5GC is designed to be a fully cloud-native,[46] service-based architecture, meaning it may be built using virtualized software functions.[47] It is composed of multiple Network Functions (NFs), some of which are optional, each exposing via standardized Service Based Interfaces (SBI) a defined set of services or capabilities. The adoption of SBI interfaces enables authenticated and authorized NFs to consume the services and data exposed by other NFs in the 5GC. Furthermore, the SBI allows new NFs to be defined and adopted into the 5GC over time without impact on already existing NFs.

In terms of AI functionality in the Core, the Network Data Analytics Function (NWDAF) was initially defined to support rule-based analytics, but later enhancements saw the introduction of AI capabilities, namely training and inference. In the core, the NWDAF is the AI platform which, thanks to the adoption of SBI, can consume data from other NFs and provide its training and inference services to other NFs in the 5G Core. Besides the NWDAF no other core NF has standardized AI capabilities. And while it cannot be ruled out that other core NFs are not using AI to some degree, such information is not in the public domain. As such, the remainder of this section focuses on the NWDAF with the observation that use cases for the adoption of NWDAF with AI capabilities are currently being studied by operators; however, no live deployments were identified for inclusion in this Report. The inclusion of NWDAF is exemplary and not intended to be taken as a normative example.

The 5G Core architecture is depicted in Figure 4 and described in Table 7 below which also shows the 5G RAN & 5G UEs.

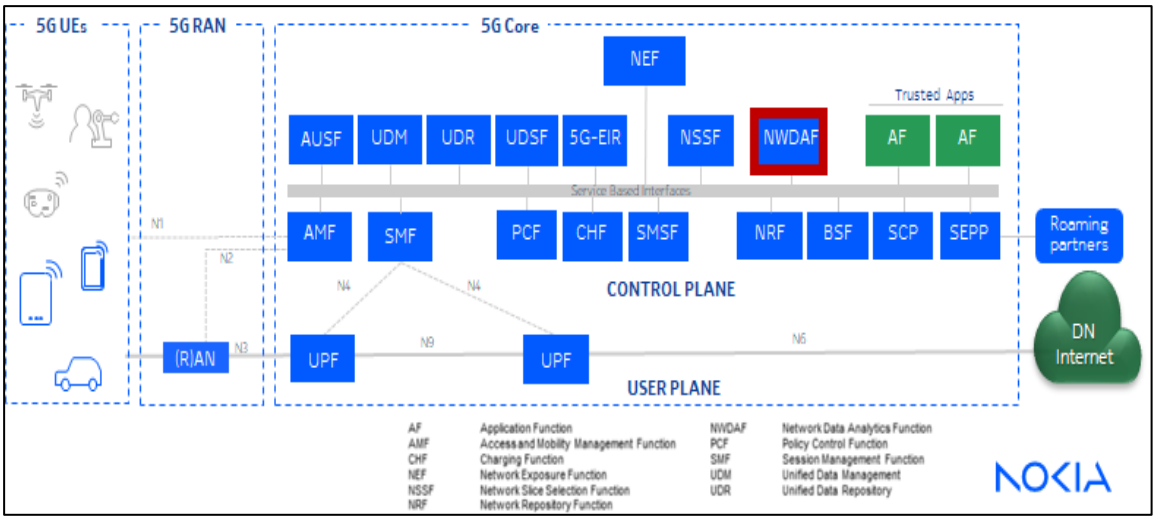

Figure 4 – 5G Core Architecture

[45] *See* Harri Holma, Harri, Antti Toskala, et al.,. *5G Technology: 3GPP Evolution To 5G*, Advanced. Second edition, Chichester, England: John Wiley & Sons Ltd, 2024, at 70-75; *5G Core Network Architecture – A Beginner's Guide to Next Generation Connectivity*, Niral Networks, Mar. 14, 2024, https://niralnetworks.com/5g-core-network-architecture/?form=MG0AV3; *3G4G Blog, Tutorial: Service Based Architecture (SBA) for 5G Core*, Feb. 9, 2018, https://blog.3g4g.co.uk/2018/02/tutorial-service-based-architecture-sba.html.

[46] Cloud native is an approach to building, deploying, and managing applications that fully leverage cloud computing environments. *See, e.g.*, AWS, *What is Cloud Native?*, https://aws.amazon.com/what-is/cloud-native/?form=MG0AV3; Microsoft, *What is Cloud Native?*, https://learn.microsoft.com/en-us/dotnet/architecture/cloud-native/definition?form=MG0AV3.

[47] NetworkBuildz, *5G Core Network Architecture: Detailed Guide*, Jan. 25, 2023, https://networkbuildz.com/5g-core-network-architecture/?form=MG0AV3.

Network Function	Description
Access and Mobility Management Function (AMF) (Mandatory)	A key control plane entity responsible for handling UE registration, connection management, and mobility.[48] It interfaces to the 5G RAN and manages authentication and supports handovers between 5G RAN network cells to maintain connectivity as users move.
Session Management Function (SMF) (Mandatory)	Handles the establishment, modification, and termination of user sessions.[49] It manages IP address allocation, policy enforcement, and communication with the data plane to ensure efficient traffic routing and resource usage.
User Plane Function (UPF) (Mandatory)	Responsible for routing and forwarding user data traffic between devices and external networks. It implements packet inspection, traffic shaping, and QoS enforcement based on policies from the Policy Control Function.
Policy Control Function (PCF) (Optional)	Responsible for defining and enforcing network policies that govern resource allocation, QoS, and user access based on subscription and network conditions.
NWDAF (Optional)	An optional NF in 5G networks that collects, analyzes, and provides insights from network data to optimize performance and enhance user experience. It leverages advanced analytics and ML to predict network behaviors, detect anomalies, and support informed decision-making processes.

Table 7 – 5G Core Network Functions

5.1.4.1 AI Use in the 5G Core

The initial release of 3GPP 5G standards, Release 15, defined the NWDAF to support basic analytic capabilities. In Release 17 (2022), 3GPP enhanced the NWDAF to support AI capabilities.[50] The NWDAF generates two types of analytics data, namely statistics data and prediction (inference) data, for specific use cases. One example of many use cases is UE Mobility analytics whereby past and future (predicted) data of the UE location can be utilized by the network to improve end-user experience and optimize network resource utilization. The NWDAF exposes a set of 3GPP standardized APIs via which analytics data for specific use cases can be requested by other network functions. The role of the NWDAF is to generate and provide the requested analytics data; it does not undertake any actions based on the generated analytics data nor does it know how that data eventually will be used. The NWDAF may provide the same analytics data to multiple different network functions with each deciding independently from the NWDAF how it will use the received analytics data.

- **Model Training**. The NWDAF utilizes multiple different statistical/predictive models for different use cases and even in the case of the same use case different models may be used; for example, using different algorithms. 3GPP standards do not define the models used in NWDAF and as such are vendor/operator specific. Typically, a model will follow a ML lifecycle whereby it is designed, trained, and tested in a development environment first and, after passing quality

[48] Telecom Trainer, *What is the Role of the Access and Mobility Management Function in the 5G Core Network*, Jan. 15, 2024, https://www.telecomtrainer.com/what-is-the-role-of-the-access-and-mobility-management-function-amf-in-the-5g-core-network/.
[49] *What is the 5G Session Management Function (SMF)?,* Dec. 29, 2022, https://techcommunity.microsoft.com/blog/azureforoperatorsblog/what-is-the-5g-session-management-function-smf/3693852.
[50] Xingqin Lin, *Artificial Intelligence in 3GPP 5G-Advanced: A Survey, (AI Survey)* https://arxiv.org/pdf/2305.05092.

assurance checks, delivered into the production environment to run on the NWDAF in the operator's network.

In the case of predictive ML models, the NWDAF can further train/retain the model using network data (referred to as Online Training) and check for its accuracy against ground truth data, which it can obtain from the network at the future predicted point in time. If the model accuracy fails to improve, then it can be brought back into the development environment where it can undergo various improvement steps or even be replaced altogether. In the development environment, the source of the training/testing data may vary based on the use case and customer to which the NWDAF model will be deployed. In the production environment, the NWDAF will subscribe to multiple different data sources depending on the specific analytics use case in order to obtain the necessary data to train/retrain an ML model. Examples of such data sources are the network NF, such as AMF and SMF, but also Operations, Administration, and Maintenance (OAM) systems. Enhancements to 3GPP Releases also made it possible for multiple NWDAFs in a network to collaborate in training a model via so called Horizonal and Vertical Federated Learning mechanisms.[51]

- **Model Inference**. The NWDAF can receive analytics requests from various NFs (e.g., AMF, SMF) for prediction data related to a specific use case targeting a future time period and other parameters such as geographical location for example. If the NWDAF does not have a model fully trained to service the analytics request, for example, the model has not been trained on historical data for the time period requested (e.g., 6 a.m. -7 a.m.), it will trigger training of the available model using the appropriate data sources for that use case. Once it has a trained model, it will run inference based on the request it received and provide the requested prediction data. It is the responsibility of the network function that requested the analytics prediction data to decide how to use that data; NWDAF has no role to play in how the analytics data it provides are ultimately used.

AI Use Case Model. One of the many analytics service use cases standardized by 3GPP in the NWDAF is UE Mobility analytics for both statistics and predictions. Being able to predict the mobility patterns of subscribers -- such as the expected number of subscribers in a particular geographical location during a future timeframe and how many are expected to move to new location and during what time period -- is information that networks can use to improve end user experience and optimize network resource utilization. Many network functions/entities in the network may be interested in obtaining UE Mobility prediction data in order to optimize the resources they control.

To obtain UE Mobility prediction data, a network function first needs to discover a NWDAF that supports that specific use case and geographical area(s). It then requests the discovered NWDAF, which then either selects a new ML model or retrains an existing ML model that supports UE Mobility predictions use case. The request to the NWDAF would typically include information such as the identity of the subscriber(s), the time period, and the area that is of interest to the requester. The NWDAF would then run (perform inference) using its selected ML model for UE Mobility for this request and return the prediction data results to the requestor. The requestor may or may not then take the received NWDAF prediction data into account when executing subsequent actions. Such actions may include adjusting network processing resources to address bandwidth needs, selecting and reselecting network resources geographically closer to the subscriber to improve latency, and adjusting network parameter configurations to help reduce the overall network signaling overhead. Network

[51] Horizontal and Vertical Federated Learning are two different approaches to distributed machine learning that allow multiple parties to train models collaboratively without directly sharing their data. Horizontal Federated Learning (HFL), also known as sample-based federated learning, is used when different organizations or entities have similar types of data but from different user bases. Vertical Federated Learning (VFL), also known as feature-based federated learning, applies when different organizations have overlapping users but distinct data features.

entities may utilize the NWDAF predictions data in different ways. For example, the AMF may use UE Mobility prediction data to optimize its paging strategy and hence conserve the usage of scarce radio resources, whereas the SMF may use that same prediction data to select an alternative user plane resource that is closer to the UE such that latency is reduced resulting in improved end user experience.

5.1.4.2 Risks Associated with AI in the 5G Core

At the time of this Report's issuance, NWDAF for analytics prediction use cases were still in the early stages of evaluation by operators, and no live deployments were identified or disclosed for inclusion in the Report. As such, the following is a general analysis based on assumptions and cybersecurity practices.

Again, taking UE Mobility analytics as a use case example, if the integrity of the NWDAF capabilities is compromised during any of its main phases, such as Machine Learning Operations (MLOps), Online Training and/or Inference, it could potentially result in the NWDAF providing compromised and incorrect prediction data. If compromised prediction data is acted upon without proper checking of the intended outcomes of those actions, it could potentially lead to negative outcomes, including degradation of subscriber experience and/or network resource utilization. For example, if the AMF receives compromised prediction data from NWDAF which it uses as input to its paging strategy it could result in the network paging for the UE in the wrong geographical areas initially and hence waste scarce network resources and increase the UE connection setup time. A further example, if the SMF receives compromised prediction data from NWDAF, then the SMF could end up selecting an alternative, but less optimal UPF which may result in increased user data latency and poor end user experience.

5.1.4.3 Mitigation Strategies and Recommendations

Network entities using NWDAF prediction data would do so for a reason: they intend to gain some benefit from using that data otherwise they will not use it. Therefore, entities using NWDAF prediction data can be expected to continuously monitor performance metrics to measure improvements or degradation of outcomes and can hence detect if the NWDAF prediction data is beneficial or not and, where deemed not useful, suspend or discontinue its use.

For the UE Mobility use case example mentioned above, the AMF and SMF can be expected to collect and monitor relevant metrics from the network to measure performance gains or losses and in the case of the latter take corrective actions, one of which may be to suspend or discontinue the use of NWDAF predictions and flag this for further investigation.

The NWDAF can also perform accuracy monitoring of its ML Model prediction data since it has access to the ground truth data; for example, it can collect actual network data for the subscriber(s) and geographical area and time period for which the prediction was requested and then compare it to its prediction data. Where there is a gap, the NWDAF can trigger accuracy improvement actions such as ML Model retraining, select a different ML model with alternative algorithms, and/or discontinue its use.

In general, the generic mitigations listed in Appendix B: Generic Threats & Mitigations to the AI Lifecycle, can be considered applicable to the NWDAF and its development and production environments.

- **Access Controls**: Access control security mechanisms can help prevent unauthorized access to NWDAF-related training and test data in both development and production environments and access to NWDAF analytics services and data in the production environment. Access control security mechanisms can also help prevent unauthorized access to the discovery of, subscription to, and usage of data and services from other network entities -- for example, to prevent a rogue network entity, including a rogue NWDAF, from gaining access to network function services and data. Access control security mechanisms should support mutual authentication and

authorization and be used to ensure that only mutually authenticated and authorized entities can discover and gain access to services and data sources. Network deployments with NWDAF should ensure that such access controls are supported and enforced throughout the network. Applicable standardized mechanisms defined by 3GPP should be leveraged where possible.

- **Data Protection**: Encryption and integrity protection of data can help prevent the usage and manipulation of that data. Data sources at rest and in transit should be encrypted and their integrity protected as appropriate. Data sources in transit includes all data (e.g., training/test data, models, model parameters) transferred between the NWDAF development and production environments. It also includes data transferred between NWDAFs themselves in the production environment (as there can be multiple NWDAFs deployed) and between NWDAFs and other NFs (e.g., AMF, SMF) in the production environment. Applicable standardized mechanisms defined by 3GPP should be leveraged where possible. Included in this consideration is the handling of subscriber specific data.

Additional security and reliability controls for consideration and application where deemed necessary and/or feasible are listed below.

- **NWDAF Model Accuracy Checking**: The NWDAF checks the accuracy of its prediction data against ground truth data and trigger model retraining and/or redesign if observed results are not acceptable.

- **Continuous Monitoring & Prediction Data Usage Control**: Users of NWDAF prediction data have controls in place to make decisions to continue/discontinue/suspend use of NWDAF prediction data. While this will be implementation-specific, a simple example of this is continuously monitoring whether the prediction data is resulting in improved outcomes for a targeted use; if satisfactory outcomes are not observed, then suspend/discontinue its usage.

- **Overload Protection**: Data Sources also have overload protection mechanisms in place to prevent resources being overloaded/exhausted. For example, an ill-behaving/compromised NWDAF issuing large amounts of subscription requests requiring event notifications be issued frequently (e.g., every 10 seconds as opposed to every 10 minutes). This is not specific to NWDAF; any ill-behaving/compromised network function could also inflict the same harm on another network functions.

- **Anomaly Detection**: Leverage Anomaly detection mechanisms to detect abnormal behavior. For example, an NWDAF might be consuming abnormally high amounts of computing resources for training, or network bandwidth utilization may be abnormally high on certain interfaces. These anomalies could be indicators of a compromised NWDAF, but these may not be exclusive to NWDAF.

5.1.5 Operations Support Systems

The OSS in a telecommunications network is used to provision and manage the network and network-based services and constitutes the "management plane" for the entire network. OSS plays a role in automating and orchestrating various network operations, including tasks such as resource management, configuration management, performance monitoring, capacity management, fault management, and security management.[52] In 5G networks, OSS encompasses a variety of applications designed to enable

[52] OSS is often integrated with the Business Support System (BSS), but they are separate and disparate parts of the network. The BSS is used for billing, managing revenues, and collecting usage data for the billing of services, while the OSS is used to manage the actual network. For example, in service provisioning, when a customer

and streamline multiple critical functions, broadly categorized into three areas: planned network operations, unplanned network operations, and performance optimization.

Planned network operations involve the deployment and life-cycle management of network functions. OSS applications in this area are responsible for planning, designing, and implementing network infrastructure and services. This includes deploying network elements, provisioning resources, configuring network elements, scaling resources, restoring resources and ensuring that new services are seamlessly integrated into the existing network. Additionally, OSS manages the entire life cycle of network functions, from initial deployment to updates and eventual decommissioning. Effective management in this area ensures that the network can adapt to evolving demands and technological advancements while maintaining service continuity.

Conversely, unplanned network operations focus on the detection and remediation of faults and disruptions. OSS applications play a crucial role in monitoring the network for any anomalies or issues that may impact service quality. When a fault is detected, OSS helps isolate the problem, identify its root cause, and initiate corrective actions to restore normal operations as quickly as possible. This proactive approach minimizes downtime and ensures that customers experience minimal disruptions. By automating fault management processes, OSS enhances the efficiency and responsiveness of network operations.

Performance optimization encompasses the monitoring and implementation of performance-enhancing configuration changes. As described in Table 8, OSS applications continuously track key performance indicators to assess the health and efficiency of the network. Advanced analytics and artificial intelligence are leveraged to identify patterns, predict potential issues, and recommend configuration changes that can improve network performance. By implementing these enhancements, OSS ensures that the network operates at its optimal level, delivering high-quality service to customers. This proactive and data-driven approach helps network operators stay ahead of potential problems and adapt to changing conditions in real-time.

Application	Function
Service order Management and orchestration	Manages and coordinates customer service requests from initiation to fulfillment. It ensures that all necessary network resources and processes are properly aligned to deliver the requested service efficiently. This application automates service provisioning, ensuring timely and accurate execution of orders, thereby enhancing customer satisfaction and operational efficiency.
Inventory and topology management	Maintains an accurate record of all network assets and their configurations. It maps the physical and logical layout of the network, ensuring efficient resource allocation and easy troubleshooting. This application helps network operators track the status and location of network elements, facilitating optimal network performance and reliability.
Fault management	Responsible for detecting, isolating, and resolving network issues promptly. It ensures that any faults or disruptions in the network are identified quickly, minimizing service downtime and maintaining optimal performance. This application helps operators proactively address potential problems, ensuring a reliable and resilient network for customers.

requests a new service, the BSS handles the order management, billing, and customer information. It then communicates with the OSS to configure the network elements and ensure the service is provisioned correctly. After reviewing the BSS functionality, CSRIC tentatively concluded that there were no credible ways in which the use of AI within the BSS could cause network disruptions.

Application	Function
Monitoring and reporting	Continuously tracks the performance and health of the network, ensuring real-time visibility into its operations. It generates detailed reports that help operators analyze trends, detect anomalies, and make informed decisions. This application is essential for maintaining optimal network performance, quickly addressing issues, and enhancing overall service quality.
Performance management and analytics/AI	Harnesses advanced data analytics and artificial intelligence to optimize network performance, predict potential issues, and make informed decisions. It continuously monitors key performance indicators, identifying trends and predicting potential issues before they impact services. This proactive approach enables operators to make data-driven decisions, ensuring efficient, reliable, and scalable network operations.

Table 8 – Key OSS Applications

5.1.5.1 AI Use in OSS

The telecommunications industry strives to achieve autonomous networks, which can adapt their size and capacity based on real-time traffic patterns. By transitioning to a virtualized network where all functions are software-based rather than hardware-based, this goal becomes attainable. AI plays a crucial role in dynamically managing network resources, predicting when expansion is necessary by analyzing historical data and current traffic patterns. In such scenarios, the OSS AI system can initiate new instances of required functions to alleviate bottlenecks. Conversely, AI can also determine when it is appropriate to shrink the network as traffic levels decrease, ordering the removal of certain instances of the network elements that are not needed, thereby optimizing resource utilization.

Having a network that can be scaled to meet increasing traffic demands by creating more instances of the required services and routing additional traffic to these new instances is crucial. This approach eliminates the need for network operators to "over engineer" their networks by incorporating redundant components, excessive capacity, or overly complex configurations in an anticipation of potential need. For instance, a network designed with far more bandwidth than the anticipated maximum demand, or layers of unnecessary hardware and software that do not directly contribute to performance or reliability, could be considered over-engineered. Consequently, the integration of OSS AI represents a significant cost reduction in terms of appliances, licensing, and power consumption for which the industry historically had to plan.

Given that software must run on some form of hardware (servers), the ability to add and remove software functions on dedicated hardware is essential. Orchestration, directed by AI tools, plays a pivotal role in this process. AI determines when network functions need to be reduced, terminating some instances from servers, and when other functions should be added, creating a dynamic environment. Inventory plays a critical role in modern network management, especially in highly scalable and dynamic architectures. As networks expand or contract based on demand, keeping track of available resources, such as servers, virtual instances, and software components, ensures efficient operation, cost management, and security. With AI-driven orchestration handling network scaling, dynamic inventory management becomes essential for tracking which assets are active, retired, or needed. Without an up-to-date inventory, it would be impossible to monitor, optimize, and protect network components. AI helps automate this process, providing real-time visibility into infrastructure while supporting virtualization, which allows services to adapt seamlessly without unnecessary hardware investments.

Beyond life-cycle management and scaling, AI can also facilitate autonomous operational decisions based on network observability. Logs from multiple network functions can be analyzed to identify issues

or degradation thresholds that may require reporting or, in some cases, necessitate so-called automated action for remediation (AI closed-loop), which involves detecting problems, such as network congestion, hardware failures, or security breaches, and taking corrective actions automatically to restore normal operations. Automated remediation actions include traffic rerouting, service restarts, resource scaling, and security responses. Closed loops help reduce the window of degradation and minimize Mean-Time-To-Restore (MTTR), but they require thorough testing and validation, along with a period in an intermediate human-assisted state to mitigate any unforeseen issues. This ongoing consideration becomes increasingly relevant as greater levels of automation and autonomy are introduced into the network management plane.

AI enhances the efficiency and adaptability of OSS in the RAN, particularly as O-RAN services and equipment emerge. It supports key tasks like spectrum management, capacity forecasting, interference reduction, and energy efficiency by dynamically adjusting RAN radio and baseband configurations. AI-driven automation, traditionally handled by SON platforms, is now extending to the O-RAN RIC platform, which hosts rApps designed to optimize RAN parameters for specific use cases. These rApps, powered by AI, not only improve performance but also lower the total cost of operation. To maximize OSS benefits, it is crucial to ensure data integrity, maintain transparency in AI decision-making, and rigorously test AI-generated outcomes across various scenarios.

Potential Use Cases.

- **Model Training**. AI/ML algorithms utilize historical and real-time network data to improve predictive models. Operators constantly collect network data, key performance indicators (KPIs), traffic patterns, and network element logs to refine AI-driven decisions on anomaly detection, optimization, and fault management.

- **Power and Configuration Management**. AI can dynamically adjust power levels, antenna tilt, and spectrum allocation within the RAN based on traffic demand, environmental factors, and network congestion. This improves energy efficiency and optimizes service quality while reducing operational costs.

- **Network Scaling**. AI-driven orchestration in OSS enables adaptive scaling of virtualized network functions, ensuring resources align with demand. As described above, the AI workload can be used to scale the network elements up and down. As traffic increases, the AI model will provide a signal to the OSS to start new virtualized network element instances to support the additional workload. As the traffic decreases, some of these virtual network element instances will be terminated. This helps with power consumption and using the resources in an optimal way.

5.1.5.2 Risks Associated with AI in OSS

It is unlikely that AI could be used to directly attack a network OSS since initial access to the network is required. The telecommunications network management plane is isolated from external (internet) connections, meaning that all the inputs and outputs of the OSS are contained within the operator's network and are isolated from the outside world. AI/ML makes it faster and easier for attackers to identify critical vulnerabilities and gain access to the network. Under a zero trust framework, operators must assume breach and build additional layers of security to secure OSS. An attacker must first gain access to the network management plane before launching any attacks. This scenario is similar to traditional, non-AI threat surfaces.

Dealing with the Potential Use Cases.

- **Data Poisoning/Leakage:** Key unintended consequences from the use of AI in Network OSS include data leakage or poisoning and unexpected outcomes from learning or prediction-based

workloads. Data poisoning will happen during the training phase of the AI/ML models. This requires the attacker to gain access to the operator's network or the ISV's environment and be able to manipulate existing historical network data that would be used specifically for this training. Such poisoning will lead to inaccurate prediction and leads to degradation of network performance and improper utilization of network assets.

- **False Traffic Generation:** For instance, AI can be used to perform an attack on the network by creating fake traffic targeting a given network element. An attacker might aim to improperly alter transmitted power and antenna settings at specific sites where AI workloads are used for dynamic RAN power and configuration management. This can be done by faking traffic at the given site which will force the AI/ML system to add resources while it is not really needed. This causes unrequired depletion of computing resources in the network.

- **Model Manipulation:** If an AI-based workload within the OSS is compromised, an attacker could manipulate the underlying model and workload, leading to network degradation and other adverse effects by improperly allocating resources where they are not needed or not allocating resources where they are needed. This type of attack would be similar to a hacker that is able to access an OSS that does not utilize AI/ML. Essentially, adding an AI workload to the OSS increases its risk surface, as would adding any other software module.

5.1.5.3 Mitigation Strategies and Recommendations

Several critical areas must be addressed to prevent unintended consequences from the use of AI in OSS.

- **Access to the Network**: The need to secure network access is a known security risk. The importance of ensuring that only legitimate users access the network is an important factor in preventing bad actors from accessing the management plane of the network. Standard zero trust security measures needs to be followed, and overlapping security measures based on assumed breach should be implemented based on a risk-based determination.

- **Safeguarding Data**: Any data used to train, fine-tune, and query the AI model must be protected whether on site or in the cloud, and provenance of training data must be tracked.

- **Output Testing and Validation**: Quality assurance and testing/validation techniques used for traditional orchestration and automation scripts must be used for the AI-assisted versions of these scripts. For traditional ML use cases, input-output testing should be conducted in a variety of simulated and real-world conditions before production deployment. Automated test execution should be independently tested from the automated test creation, ensuring thorough validation of AI-generated automation/orchestration scripts.

- **Periodic Internal Assessments**: Operators should periodically reassess AI output to ensure it adapt to changing scenarios and inputs. As AI algorithms modify RAN parameters based on usage, it is important to verify that the AI's output shifts as expected with usage changes.

- **Monitor Unexpected Network Behavior**: Because AI models react to data and variable traffic patterns, an attack might be disguised as unexpected traffic, either localized or distributed. This traffic can cause the AI model to allocate resources in a way that negatively impacts normal operations. A human-in-the-loop approach should be used. Human observers must validate any excessive, abnormal traffic that does not match historical patterns to avoid misallocation of

resources, in addition other anomaly detection techniques can be used to alert the network operator of unexpected behavior.

5.2 6G Networks

6G represents the next generation of wireless communication technology that promises faster speeds, lower latency, and enhanced connectivity to support innovative applications and services beyond the capabilities of 5G. The 6G network concept, now in its initial stages of design for anticipated launch in 2030, fundamentally draws on the 3GPP mobile broadband standards and specifications. Its operational concept also derives from the O-RAN Alliance standards and the ITU-R International Mobile Telecommunications-2020 (IMT-2020) guidelines for 5G networks, devices, and services. The deployment and findings associated with 5G networks will continue to drive 6G's evolution, including relevant considerations for AI-Native design in view of potential AI threats and attack strategies that may arise as AI itself further evolves over the next several years. 3GPP standards have already begun to address the incorporation of AI for the envisioned 6G architecture.

However, 6G network concepts and the envisioned architecture including the use of AI-Native constructs are still in their early stages of development. An overview of the top-three use cases involving AI-Native unique to the 6G space, along with an identification of potential threat vectors and vulnerabilities that could arise and mitigation strategies that should be considered are provided in Appendix C. The analysis of the relevant 6G network issues using AI/ML were also coordinated with CSRIC IX Working Group 3 on 5G Security and Reliability as part of a liaison activity and are incorporated to the extent practical in the discussion.

6 Recommendations

The role of a federal advisory committee is to provide expert guidance on complex technical and policy matters. Throughout this report, CSRIC has described threats posed by AI/ML to the security and reliability of communications networks and presented options for industry to promote sound policies and practices that support network security and resilience. We summarize those below and recommend that the FCC work in collaboration with federal partners, industry and other stakeholders to develop and disseminate best practices to clarify and strengthen how AI is used to support and optimize communications networks.

The rapid evolution of AI within telecommunications demands a proactive and comprehensive strategy to safeguard networks against emerging threats. To ensure resilient, secure, and efficient operations, operators should consider adopting a zero trust approach, assuming breach and must implement a series of targeted mitigating measures that span the full network across the entire network spectrum. These guidelines emphasize robust access controls to restrict unauthorized data access, stringent data protection measures, including encryption and integrity safeguards, as appropriate. Additionally, deploying continuous monitoring and anomaly detection systems, enforcing rigorous output testing and validation procedures, and conducting periodic model assessments are critical to mitigating risks associated with AI-driven functionalities. Furthermore, support of overload protection mechanisms, the adoption of standardized security protocols, and the promotion of AI education and awareness training, all of which collectively establish a robust defense against potential vulnerabilities across all technological layers.

<u>Zero Trust Approach</u>

Under a zero trust approach, operators and network vendors should consider maintaining strong access control measures across all layers of telecommunications networks, including OSS, RAN, and backhaul, to prevent unauthorized access to sensitive data and services. Specifically, rather than assuming trust based on weak assurances like network location, operators and vendors should analyze all aspects of access requests, including identity, endpoint, network, and resource and apply threat intelligence and

analytics to assess the context of each request. Additionally, least privileged access ensures that permissions are only granted to accomplish specific tasks from the appropriate environment and on appropriate devices. Where appropriate, information that can be connected to subscriber's identity or location.

Monitoring and Anomaly Detection

Deploy continuous monitoring systems across the 5G core, OSS, RAN, and backhaul to identify and respond to abnormal behaviors or unexpected traffic patterns. Advanced anomaly detection mechanisms should be used to flag compromised systems (e.g., an ill-behaving NWDAF or misallocation of resources by AI models) and alert network operators promptly. These measures ensure the network's operational integrity remains intact.

Output Testing and Validation

Conduct rigorous testing and validation of AI-generated outputs across all network functions, including predictive analytics and automation scripts. Quality assurance protocols, simulated environments, and real-world conditions should be utilized to confirm the reliability of AI-driven orchestration and decision-making processes before deployment.

Periodic AI System and Model Assessments

Continuously reassess AI Systems and AI/ML models, ensuring they adapt to evolving network conditions and threats. This is particularly critical for OSS systems modifying RAN parameters and NWDAF predictions. Retraining models based on updated datasets and reevaluating their model accuracy ensures that AI/ML systems operate effectively.

Overload Protection Mechanisms

Introduce safeguards to prevent resource exhaustion due to ill-behaved or compromised network functions. These mechanisms are vital for NWDAF subscriptions and equally important for RAN and backhaul resource allocations. Overload protection ensures that malicious traffic or unexpected surges do not disrupt regular network operations.

Access Controls

Maintain strong access control measures across all layers of telecommunications networks, including OSS, RAN, and backhaul, to prevent unauthorized access to sensitive data and services. Access control security mechanisms play a vital role in preventing unauthorized access to sensitive data and services within the telecommunications industry. These controls are critical in both development and production environments, safeguarding data such as training and test datasets, analytics services, and other network-related information. By implementing mutual authentication and authorization protocols, operators ensure that only trusted entities can discover, access, and utilize network services and data sources, thereby mitigating risks posed by rogue entities. To maintain consistency and interoperability, operators should adopt applicable standardized mechanisms, such as those defined by 3GPP, and enforce these access controls across all network components. Such robust measures help protect both the management and operational planes of the network, ensuring secure and reliable functionality throughout.

Data Protection

Adhere to appropriate data protection, integrity and provenance protection measures to safeguard data from unauthorized usage and manipulation within the telecommunications industry. All data, whether

stored ("at rest") or being transferred ("in transit"), should be protected to ensure its confidentiality, and integrity protection mechanisms should be applied to prevent tampering or unauthorized alterations. Data in transit includes critical elements, such as training and test datasets, machine learning models, and model parameters, which are exchanged between development and production environments, as well as data shared between multiple deployed systems and other network functions within the production environment. Standardized encryption and integrity protection methods, such as those defined by 3GPP, should be adopted wherever possible to maintain consistency and secure data across all network operations. Consider double encryption (e.g., encrypting at both the application service layer and the platform layer) for especially sensitive data. These practices provide a robust framework for mitigating risks and preserving the reliability and security of telecommunications systems.

Standardized Security Protocols

Adopt applicable security protocols standardized by 3GPP for interoperability. These protocols can address vulnerabilities across NWDAF, OSS systems, and other AI-driven functionalities in RAN and backhaul, enhancing resilience against emerging threats.

In particular the adoption of NIST standardized Post Quantum Cryptography (PQC) algorithms and standardized security protocols supporting PQC (e.g., TLS, IPsec, etc.) is encouraged.[53]

Strengthening Supply Chain Security for AI-Enabled Telecommunications

To safeguard the telecommunications supply chain from emerging threats and ensure the responsible deployment of AI technologies, the Commission, in coordination with manufacturers and industry telecom operators, should collaborate on best practices for rigorous monitoring of the multivendor ecosystem, adhering to AI governance policies, and ensuring transparency in component sourcing. For example, as AI integration in UE handsets expands, it is critical to track hardware and software provenance through bills of materials (BOMs) that include inventory lists, certifications, and security benchmarks. Beyond BOMs , stakeholders should consider establishing a root-of-trust framework that validates the authenticity of AI-enabled devices at multiple levels, reducing exposure to counterfeit components, manipulated firmware, or unauthorized access risks. Implementing secure data pedigree tracking will help maintain integrity across the AI lifecycle, ensuring models and datasets adhere to strict security and reliability standards. Collaboration with industry stakeholders must extend to continuous risk assessments, vendor audits, and enforcement of secure software supply chain principles, mitigating vulnerabilities before they disrupt critical communications infrastructure. By promoting standardized security protocols and proactive coordination, the Commission can reinforce industry resilience against AI-related exploitation, ensuring a trusted, transparent, and protected telecommunications ecosystem.

Sharing of Data and Models

CSRIC observed that there are limited real world examples of AI being deployed. Promoting access to training and test data, and sample models for key telecommunications AI use cases to foster a robust ecosystem of model providers for telco scenarios. Standardized interfaces, datasets and schemas will help vendors create standardized AI solutions for the telecommunications network, and accelerate operators' ability to evaluate, contrast, and securely deploy AI systems.

AI Education

Promoting comprehensive AI education and awareness training, especially related to the use of AI in

[53] *See, e.g.*, NIST, *Status Report on the Fourth Round of the NIST Post-Quantum Cryptography Standardization Process*, Mar. 2025, https://csrc.nist.gov/pubs/ir/8545/final.

telecommunications. Training should include considerations for evaluating and adopting AI-based technologies in a secure and responsible manner. Appendix A contains a good list of resources as a starting point, and a forthcoming report from CSRIC will address "Recommended Best Practices for the FCC and Industry on the Ethical and Practical Use of Artificial Intelligence/Machine Learning."

7 Conclusions

The integration of AI into telecommunications networks presents both unprecedented opportunities and serious security challenges. Rapid AI adoption introduces complex risks that demand proactive mitigation. To ensure the reliability and integrity of AI-powered networks, industry stakeholders must implement stringent safeguards, including robust risk assessment protocols, continuous monitoring systems, and reinforced data protection measures. Industry's commitment to standardized security frameworks and ongoing consumer education will be crucial in protecting against adversarial interference and operational vulnerabilities. At the same time, the Commission has a vital role in encouraging best practices, ensuring AI implementations align with security requirements, and promoting policies that safeguard both industry and public interests. As network operators ramp up AI use in 5G networks and prepare for the design and deployment of 6G networks, continued collaboration between industry and the Commission will be essential.

Appendix A: Overview of Key Frameworks for Understanding Threats and Risk Mitigation in AI Systems

- Based on the notions of security, resilience, and robustness of ML systems from the National Institute of Standards and Technology (NIST) AI Risk Management Framework,[54] NIST further developed a taxonomy for adversarial machine learning (AML) risk assessment. They considered the following five dimensions: (i) AI system type (Predictive or Generative AI), (ii) stage of the ML lifecycle process when the attack is mounted, (iii) attacker goals and objectives, (iv) attacker capabilities, and (v) attacker knowledge of the learning process and the AI system.[55] Using these attributes, they provide a detailed taxonomy for adversarial machine learning attacks and potential remedies.

- Open Worldwide Application Security Project (OWASP), a nonprofit foundation that works to improve security of software, 2023 release of Top Ten Machine Learning Security issues in ML systems.[56] In 2025, OWASP updated the list specifically targeting the development and deployment of Large Language Models.[57] Similar to their work on cybersecurity, OWASP identifies potential vulnerabilities at a very detailed level with examples of attack scenarios and mitigation strategies.

- MITRE's ATLAS Matrix provides a matrix of tactics, techniques, mitigations and case studies for various AI related attacks.[58]

- MIT AI Risk Repository is a massive compendium of AI risk-related papers in the academic literature.[59] It has three parts:

    - A Database with 1000+ risks extracted from 56 existing frameworks and classifications of AI risks.
    - A Causal Taxonomy that relates how, when, and why these risks occur.
    - A Domain Taxonomy classifies these risks into 7 domains (e.g., "Misinformation") and 23 subdomains (e.g., "False or misleading information")

- IBM AI Risk Atlas provides a framework for understanding the broad range of risks underlying the use of LLMs in the enterprise.[60] It covers both predictive and generative AI applications. Risks capture three distinct aspects: (i) Model Inputs during training (e.g. biases in the training data) or during use (e.g., sensitive data given during prompt engineering), (ii) Model Outputs (e.g., hallucination) and (iii) other considerations (e.g., legal). IBM's risk framework also includes three

---

[54] NIST. *NIST AI Risk Management Framework*, https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.
[55] A. Vassilev, et al., *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST Trustworthy and Responsible AI NIST AI 100-2e2023, https://doi.org/10.6028/NIST.AI.100-2e2023.
[56] *OWASP Machine Learning Security Top 10*, 2023, https://owasp.org/www-project-machine-learning-security-top-10/.
[57] *Id*.
[58] MITRE, *ATLAS Matrix*, https://atlas.mitre.org/matrices/ATLAS.
[59] MIT, *AI Risk Repository*, https://airisk.mit.edu/.
[60] IBM, *AI Risk Atlas*, Feb. 7, 2025, https://www.ibm.com/docs/en/watsonx/saas?topic=ai-risk-atlas.

indicators to capture the origin of the risk: (1) Traditional - known risk from prior or earlier forms of AI systems; (2) Amplified - known risk but now intensified because of intrinsic characteristics of LLMs, most notably their inherent generative capabilities; and (3) New - emerging risk intrinsic to LLMs and their generative capabilities.

- Organization for Economic Cooperation and Development (OECD) created the AI Risk Ontology (AIRO) "for expressing risk of AI systems based on the requirements of the AI Act, ISO/IEC 23894 on AI risk management and ISO 31000 series of standards. AIRO assists stakeholders in determining "high-risk" AI systems, maintaining and documenting risk information, performing impact assessments, and achieving conformity with AI regulations."[61]

[61] OECD, *AI Risk Ontology*, Dec. 5, 2024, https://delaramglp.github.io/airo/.

Appendix B: Generic Threats & Mitigations to the AI Lifecycle

AI Lifecycle Training Phase: Threats & Mitigations

The training phase of AI/ML models is particularly critical, as it lays the foundation for their functionality and accuracy. However, this phase is vulnerable to various security threats, such as data poisoning attacks, where adversaries introduce malicious data to compromise the model's performance. Similarly, the testing phase faces risks like evasion attacks, where manipulated input data is used to deceive the model. Addressing these threats requires a multi-layered approach, including robust access controls, encryption, continuous monitoring, and auditing processes. By incorporating these measures, organizations can enhance the security and resilience of AI/ML systems throughout their lifecycle.

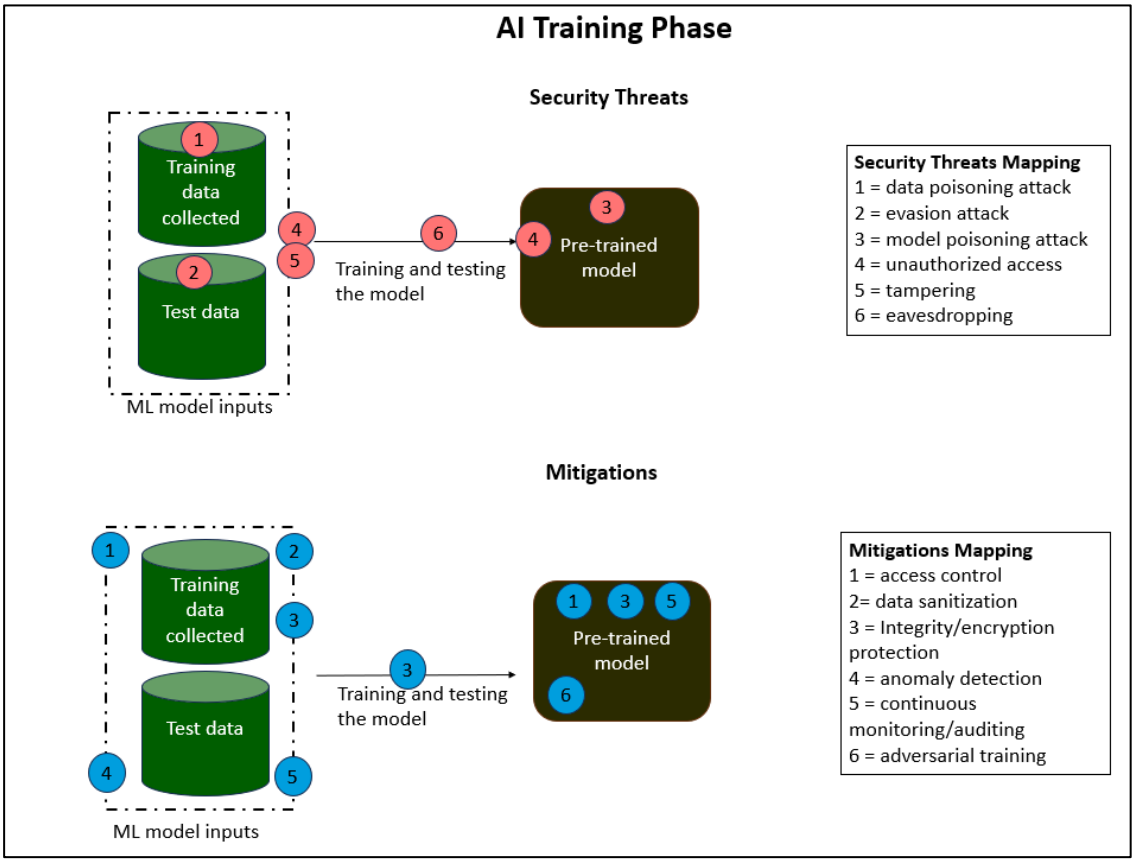

Figure 7 – AI/ML Training Phase Threats and Mitigations
Source: Nokia

As seen in Figure 7, in the context of machine learning, collecting training data presents the risk of a data poisoning attack. A data poisoning attack involves an adversary injecting deceptive or malicious data into the training dataset, which can degrade the model's performance or lead it to make incorrect decisions. To mitigate this threat, implementing access control measures is crucial. Access control ensures that only authorized personnel can modify or contribute to the training dataset. This involves authentication mechanisms such as multi-factor authentication (MFA), which requires multiple methods of verifying a user's identity, and role-based access control (RBAC), which grants permissions based on a user's role within the organization.

When it comes to maintaining test data, the primary security concern is an evasion attack. During an evasion attack, an adversary manipulates input data to deceive the model into making incorrect predictions. To protect against this, anomaly detection systems should be employed. Anomaly detection algorithms monitor the input data for unusual patterns or behaviors that could signify an attack. Additionally, continuous monitoring and auditing of the test data are essential. Continuous monitoring involves regularly checking the data for signs of tampering or manipulation, while audits are systematic reviews of data and processes to ensure compliance with security standards and identify potential security gaps.

In the phase of training and testing the machine learning model, other threats include unauthorized access, tampering, and eavesdropping. Unauthorized access refers to an individual gaining access to the system without permission, while tampering involves altering the data or models to cause incorrect functioning. Eavesdropping is an unauthorized interception of and listening to a data transmission. To mitigate these threats, integrity and encryption protection, both data at rest and in transit, must be implemented. Encryption ensures that data is encoded and only accessible to those with the proper decryption key, protecting it from unauthorized access during transmission or storage. Integrity checks involve verifying that data has not been altered, maintaining the trustworthiness and accuracy of the data and model.

Once the model is pre-trained, it remains vulnerable to unauthorized access, tampering, and eavesdropping. Securing a pre-trained model using encryption can secure it from unauthorized access and alterations. Moreover, continuous monitoring and auditing remain essential to detect any attempts at tampering or unauthorized access promptly.

Lastly, the pre-trained model is also susceptible to model poisoning attacks, where adversaries introduce vulnerabilities during the training process. To mitigate this, strict access control must be enforced, ensuring only authorized personnel can modify or access the model. Additionally, continuous monitoring and auditing, along with integrity and encryption protection, ensure the model remains secure and any unauthorized changes are promptly identified and addressed. Adversarial training can be used when training the model in which the model is trained on adversarial examples to make the model more robust against evasion attacks,

AI Lifecycle Inference Phase: Threats & Mitigations

The inference phase in AI/ML systems is a critical stage where input data and model outputs must be safeguarded against potential security threats. This phase is particularly susceptible to risks such as unauthorized access, tampering, and prompt injection attacks, which can compromise the integrity and reliability of the system. To address these vulnerabilities, implementing robust access control mechanisms, integrity checks, and encryption protocols becomes essential. Additionally, continuous monitoring and auditing play a pivotal role in detecting unusual activities or anomalies in real-time. By fostering a culture of AI/ML security awareness and education among all personnel, organizations can effectively mitigate these threats and ensure the safe and reliable operation of AI/ML systems during inference.

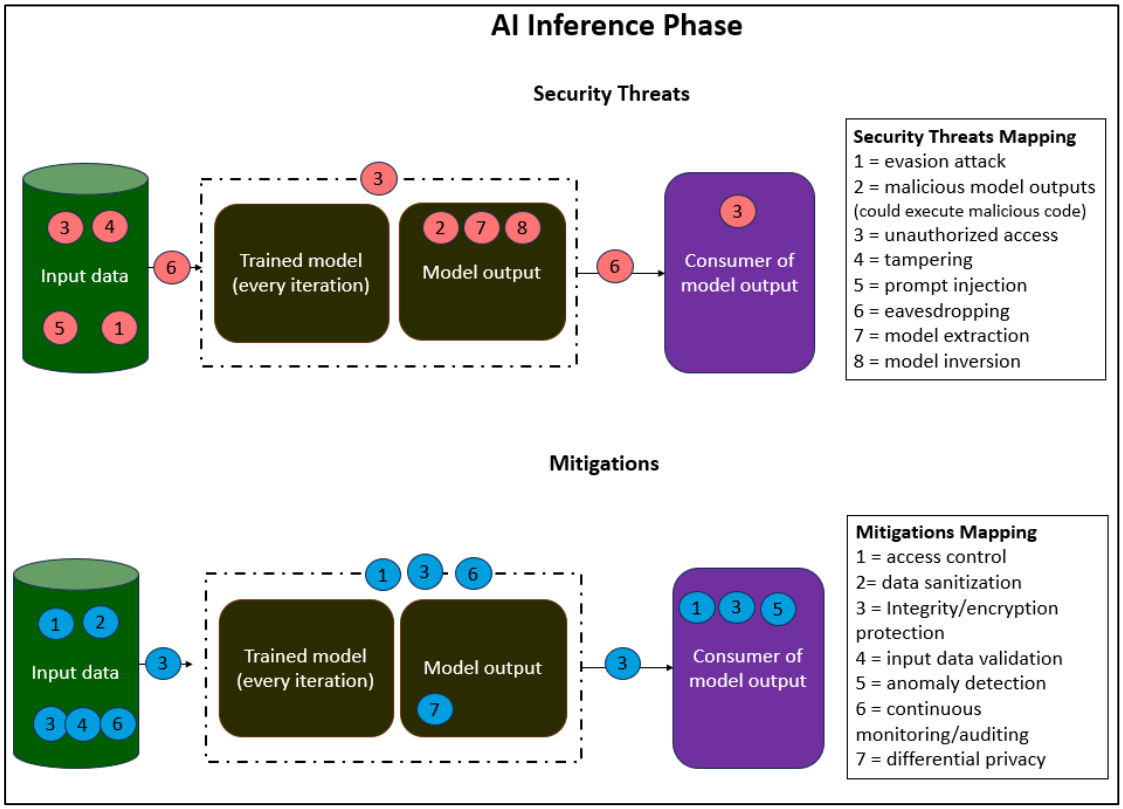

Figure 8 – AI/ML Inference Phase Threats and Mitigations
Source: Nokia

As seen in Figure 8, during the inference phase, input data is susceptible to unauthorized access, tampering, and prompt injection attacks. Unauthorized access refers to individuals gaining access to the system without proper authorization, which can compromise the integrity and confidentiality of the input data. Tampering involves the manipulation of input data, leading to incorrect model predictions or outputs. Prompt injection is a threat where malicious prompts or inputs are injected to influence the model's responses. In addition, a malicious adversary could attempt to steal the model or reverse engineer the functionality of the model through model extraction and model inversion attacks. In a model extraction attack, a model is stolen, and in a model inversion attack, through querying the model, an adversary is able to gain insight and information into the inner workings of the model.

To mitigate these threats, several security measures should be put in place. Access control is crucial to ensure that only authorized users can access and change the input data, which can be achieved through multi-factor authentication (MFA) and role-based access control (RBAC). Data sanitization techniques help in cleaning and validating input data to remove any malicious or corrupt entries. Integrity and encryption protection safeguard the data by encoding it and checking for unauthorized modifications. Input data validation involves verifying the input data to ensure its accuracy and authenticity. Anomaly detection mechanisms can be put in place by the consumer using another AI algorithm to detect any anomalies from the model output. Differential privacy techniques can also be used to protect the model output to avoid model inversion or model extraction attacks. Finally, continuous monitoring and auditing are essential to detect any unusual activities or anomalies in the input data, allowing for swift detection and response to potential threats.

The pre-trained model and its output also face risks such as unauthorized access and tampering. To address these, access control measures should be enforced to restrict access to the trained model and its

outputs to authorized personnel only. Integrity and encryption protection ensures that both the trained model and its outputs are secure from unauthorized alterations and eavesdropping. Additionally, continuous monitoring and auditing help in identifying and addressing any unauthorized access attempts or modifications, maintaining the integrity of the model and its outputs.

One specific threat to the model output is the execution of malicious code, known as malicious model outputs. This threat involves manipulating the model outputs to include malicious code that can harm end-users or systems. To mitigate this risk, access control is necessary to ensure that only authorized individuals can modify or access the model outputs. Integrity and encryption protection secure the outputs by encoding them and maintaining their integrity, preventing unauthorized modifications. Continuous monitoring and auditing help detect and address any attempts to manipulate the model outputs.

Lastly, the consumer of the model's output is also exposed to risks like unauthorized access. Access control measures should be in place to ensure that only authorized consumers can access the model outputs. Integrity and encryption protection safeguard the outputs, ensuring they remain secure and unaltered during transmission and storage.

By implementing these comprehensive security measures, the model remains secure and reliable, protecting against various threats during the inference phase. Throughout all these phases, emphasizing AI/ML education and security awareness training for all personnel involved is crucial. This helps create a culture of security, ensuring that everyone is aware of the importance of maintaining data integrity and protecting against threats. Regularly updating this training keeps team members informed about the latest security threats and best practices for mitigating them.

Appendix C: AI Use in 6G Network Technology

AI capabilities for training and inference have already been defined for a few specific network entities such as the NWDAF in the Core Network, which analyzes data to improve network performance and user experience. A similar but expanded concept is envisioned for 6G systems. To offer a baseline, as depicted in Figure 9, the insertion of the AI and NF component (Figure 9b) is aimed at supporting MLOps performance. The equivalent NWDAF and supporting functions in 6G will be the central target of AI threat-attack-mitigation strategies. Figure 9 further depicts the extended **AI-Native functionality** anticipated in 6G. The implication is that, rather than strictly limiting AI to a few network entities, such as the NWDAF, operators will leverage AI use throughout the network in various network entities across multiple domains (e.g., RAN, Core, and Operations blocks). In addition, operators will most likely support some level of AI capabilities in all network entities are anticipated to optionally support some level of AI capabilities; for example, training and/or inference, to support ISAC, holographic communications, and digital twin functionalities and utilities.

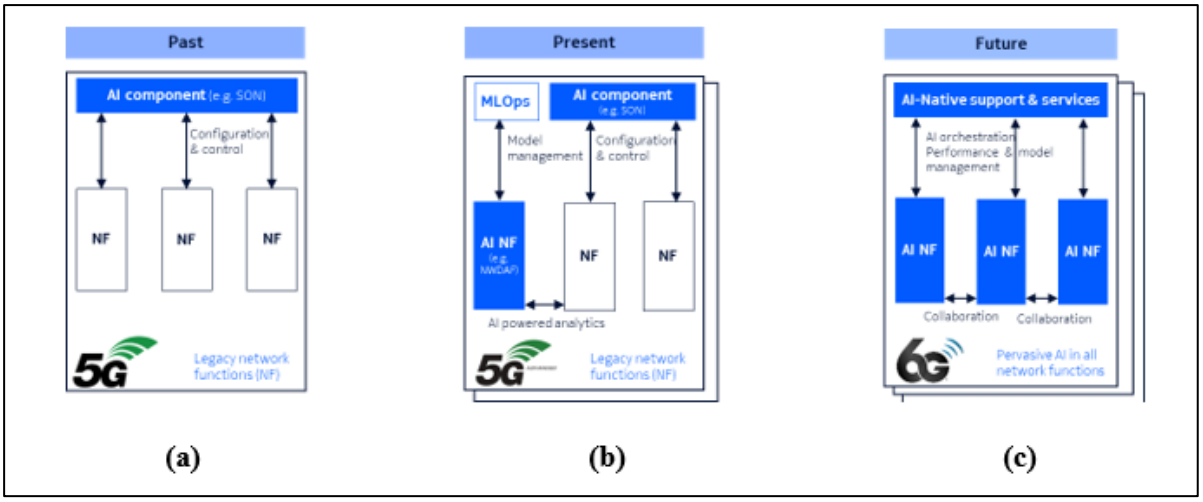

Figure 9 -- (a) Past , (b) present AI capability evolutions in 5G & 5G Advanced, and (c) AI capability evolution in 6G
Source: Nokia

Use Cases. AI will advance virtualization in 6G by extending software-defined networking to additional network elements, including radio units and antennas, and enable AI-driven optimization of radio hardware to enhance spectrum efficiency, reduce energy consumption via adaptive power control and real-time resource allocation, and help with cost reduction. A representative subset of high-priority use cases for 6G networks using AI technologies includes Integrated Sensing and Communications (ISAC), Holographic Communications, and Digital Twins.

- *Integrated Sensing and Communications.* ISAC is a technology that combines wireless communication and sensing capabilities into a unified system, allowing networks to transmit data while simultaneously detecting and analyzing the surrounding environment. 6G ISAC will merge RF sensing and wireless communication for dynamic spectrum access, object detection, and environmental awareness. However, ISAC introduces new attack vectors, including RF spoofing, adversarial perturbations, and sensor data poisoning, which can compromise AI-driven decision-making. Tight security protocols will be essential to protect sensitive information and maintain the integrity of both the sensing and communication functions within 6G networks.

  As we push further into 6G, we see more AI incorporated into the physical layer, bringing data

closer to the user. As the shift grows from network-centric data towards AI-driven SONs and localized AI frameworks for autonomous spectrum management and real-time analysis, as shown below in Figure 10, it brings with it the need for stronger security and privacy measures at the user level. While data at higher layers may be more structured and tolerate higher processing delays, the raw, fine-grained data processed closer to the user demands heightened scrutiny. Ensuring that security considerations are as robust if not more so than in the core, will be critical as the transition toward AI-Native networks in 6G takes place.

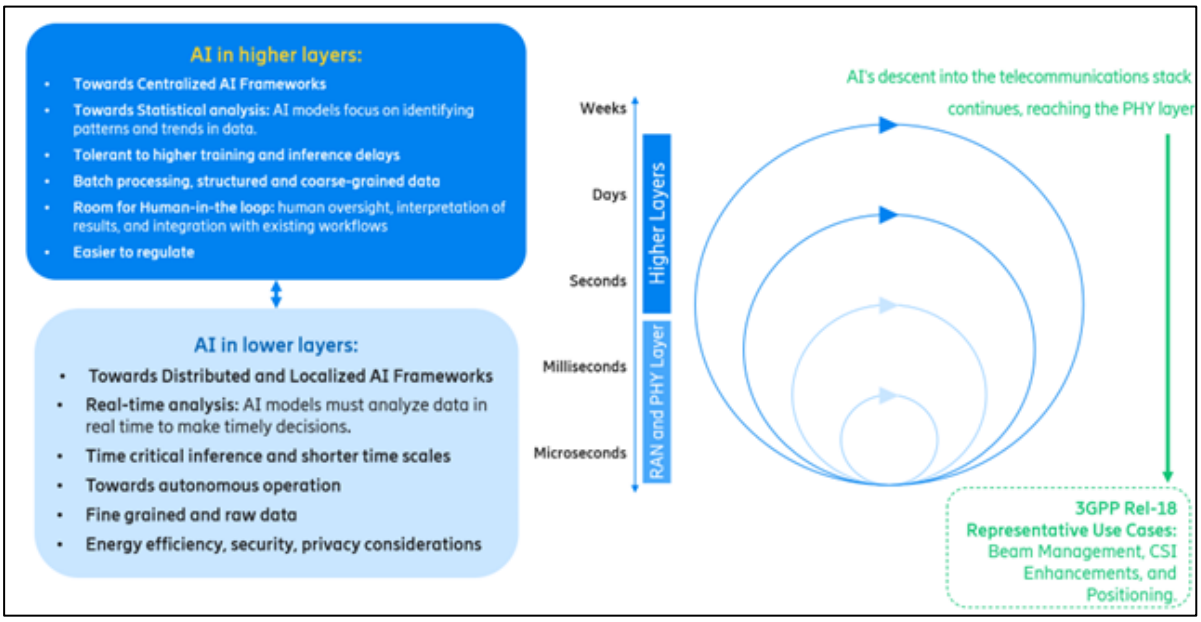

Figure 10 -- AI Across the Network Stack.
Source: Ericsson

- *Holographic Communications.* Holographic communication is an advanced technology that enables real-time, immersive 3D interactions, allowing people to communicate as if they were physically present in the same space. 6G networks will integrate AI and advanced communication technologies to enable hyperconnectivity and immersive experiences by bringing over-the-air data closer to the user. These networks will leverage millimeter-wave (mmWave) and terahertz (THz) spectrum to achieve ultra-high data rates and low latency. In the envisioned AI-Native 6G design, AI will be embedded directly into networking equipment from design inception. AI will play a crucial role in designing the 6G air interface, optimizing waveform generation, beamforming, and dynamic spectrum allocation to provide for an overall enhanced user experience. In addition to enabling autonomous 6G networks to optimize themselves in real-time -- enhancing performance, reliability, and energy efficiency, and facilitating predictive maintenance and intelligent resource allocation -- AI will set the stage for a highly adaptive communication ecosystem to provide for personalized user experiences. Some advanced forms of Generative AI, deep/reinforcement learning, or other AI variations will likely be used in 6G networks for this purpose and are yet to be defined. This creates an opportunity for adversarial breaches using AI data poisoning methods to confound/disrupt the user experience. This will also require addressing guardrails for self-learning processes in diverse environmental applications or scenarios.

- *Digital Twins.* Digital twins refers to virtual replicas of physical network components, systems, or environments that enable real-time simulation, analysis, and optimization. These digital models allow operators to predict network behavior, test configurations, and enhance

performance without disrupting live systems. 6G networks will enhance the use of digital twins as part of an emulation and "hardware-in-the-loop" environment by enabling real-time, AI-driven simulations of physical systems. Improved connectivity, sensing capabilities, and massive data collection will allow more accurate predictions and optimization across industries, including smart cities, industrial automation, and healthcare. This will be possible due to 6G networks' enhanced connectivity, stronger sensing capability, and capacity to collect massive amounts of data. The new functionality that AI will introduce includes communication plus sensing, communication plus computing, and communication plus autonomy through AI-enabled support for ISAC elements. Generative AI, deep reinforcement learning, and LLMs will be used in concert to ingest, reason and infer upon large data sets to compare ideal operation and functionalities (through a set of predefined metrics) via simulation and comparing to *in situ* real-time network performance to make any necessary adjustments to optimize operations. This same methodology can be used to detect both AI- and non-AI-induced anomalies and take adaptive corrective actions.

Digital twins can be used to train network entities to collaborate, meaning that information could be shared among them by leveraging AI capabilities. For example, one network entity could request another network entity to train a model on its behalf (cross-training), or a network entity could act as a data producer for another network entity that trains a model and so on. AI is also expected to be leveraged in a more distributed way with multiple network entities collaborating locally and centrally. For example, Federated Learning (FL) could be enhanced in 6G where distributed network entities locally train a model using local data sources. Local digital twin models are subsequently aggregated by a central network entity for iterative or cyclical improved learning.

Risks Associated with AI in 6G Networks

Failures in 6G AI design and implementation could give rise to vulnerabilities downstream of deployment, without sufficiently exercising steps early on to preclude deficiencies. Additionally, attacks targeting 6G AI-Native systems are flagged, including attacks using AI that may be reasonably anticipated along with mitigation measures that should be considered to limit vulnerabilities.

Virtually everything we know about cyber threats and attack strategies in the context of 5G systems will apply also to 6G networks, but there remain unknowns. While not intending to "over signal" with respect to the expected severity of looming AI-based threat-attack scenarios for 6G, the current immaturity of 6G technologies and the uncertainties associated with the specific nature of future AI threats are significant concerns. The focus of this section is on the relevant AI considerations in the context of high-priority use cases relevant to 6G evolution and deployment.

The 6G AI-Native architecture will enable intent-based networking, autonomous Medium Access Control scheduling, and predictive analytics for proactive fault management. However, as AI integrates deeper into network control and orchestration, security challenges such as adversarial AI, data poisoning, and model inversion attacks must be addressed to ensure network resilience.

Further, advancements in FL and integrated cloud services will also be pivotal in 6G, allowing networks to train AI models on distributed edge data while preserving privacy. Integrated into NWDAF, FL enables real-time adaptation without exposing raw data. However, FL models are vulnerable to gradient inversion attacks, model poisoning, and adversarial updates. To mitigate these risks, secure aggregation, homomorphic encryption, and differential privacy mechanisms should be integrated as appropriate in order to safeguard model integrity.

Possible types of risks and mitigation responses across the 6G network stack based on the use cases of concern include:

- Adversarial breaches using AI data poisoning methods to confound/disrupt user experience.
- Introducing bad actor AI agents via the multivendor ecosystem that disrupts normal functions and desired operation or produces undesired decisions and unintended outcomes.
- Model evasion attacks.
- Model exfiltration (e.g., extract energy efficiency optimization models).
- Real time data manipulation generates false reports.
- Configuration attacks on network settings.

AI vs. AI competing strategies where different AI systems or AI-driven approaches are pitted against each other either in direct competition or as part of a broader strategic challenge, may deplete network resources, causing desired attention mechanisms to stray.

Table 9 provides a high-level summary of a threat model/risk analysis for how use of AI can affect the 6G network communications profile and performance for the three main use cases cited.

Risk Category/ Use Case	ISAC	Holographic Communications	Digital Twins
AI Attacks	Deep reinforcement learning leveraging AI-assisted pattern of behavior models are leveraged to enhance/automate remote cyber-physical attacks (replays, MITM, DOS/DDOS, spoofing or impersonation, etc.) on critical communications infrastructure to infiltrate the internal network, exfiltrate data, and gain network control. Common AI-driven threats include automated vulnerability exploitation (where AI identifies and exploits security gaps at scale) for phishing or misinformation campaigns, and AI-powered reconnaissance (where ML models analyze network behavior to detect vulnerabilities before launching attacks).	Generative AI can be leveraged to enhance and automate cyber-physical attacks on critical communications interfaces negatively affecting user experience or forcing bad data-to-decisions processes that confound users and operators. Common AI-driven threats include deepfake-assisted social engineering (creating hyper-realistic impersonations for phishing or misinformation campaigns), and AI-powered reconnaissance that exploits patterns of behavior.	Generative AI and deep reinforcement learning mechanisms can be used to confound or poison digital models and indirectly impact real-time network operation or functionalities. Common AI-driven threats include automated vulnerability exploitation through learning data corruption that can further exploit security gaps at scale in real networks, including imposter or deepfake-assisted social engineering, as well as AI-powered reconnaissance using ML to analyze network behavior to detect vulnerabilities before

Risk Category/ Use Case	ISAC	Holographic Communications	Digital Twins
			launching attacks.
Attacks Targeting AI-Native Systems	Both AI- and non-AI-enabled attacks further assisted by pattern of behavior models can be used to target critical AI-Native communications infrastructure and exploit vulnerabilities to enhance/automate remote cyber-physical attacks (replays, MITM, DOS/DDOS, spoofing or impersonation, etc.) to infiltrate the internal network, exfiltrate data, and gain network control.	Both Generative AI- and non-AI-enabled attacks further assisted by pattern of behavior models can be used to enhance and automate cyber-physical attacks on critical communications interfaces negatively affecting user experience or forcing bad data-to-decisions processes that can be used to confound users/operators. Model evasion attacks, including adversarial inputs deceiving AI classifiers, data poisoning (attackers injecting malicious samples to degrade model performance), and model inversion attacks (adversaries reconstructing sensitive training data from exposed AI models). Backdoor attacks can introduce hidden behaviors into AI systems, making them vulnerable to targeted exploitation. Barrage and replay attacks can be automated to overmatch AI-Native functions to the point where vulnerabilities are detected to exploit models for use cases that are insufficiently trained.	Both Generative AI- and non-AI-enabled attacks can be used to poison training data and telemetry data collected for recursive digital twin model improvements that may further affect real 6G network system operation causing cascading or precipitous network degradation. Adversarial inputs used to deceive AI classifiers, including data poisoning and model inversion attacks (adversaries reconstructing sensitive training data from exposed AI simulation models).
AI-Native Design/ Implementation Failures	Insufficient use cases and corresponding training models are used to develop AI-Native constructs across critical 6G communications	Use cases and corresponding training models are insufficient in AI-Native design to account for potential AI- and non-AI cyber-attack	Non-vetted or corrupted training data are used to train first- or subsequent-generation digital models used in

Risk Category/ Use Case	ISAC	Holographic Communications	Digital Twins
	infrastructure network segments. Deficiencies or inadequacies in the planning, structure, implementation, execution, or maintenance of AI-Native tools or system leading to malfunctions or other unintended consequences that affect critical communications infrastructure operations. Failures in AI design and implementation can lead to unintended disruptions in critical communications infrastructure where such failure modes include Autonomy Risk, Brittleness, Fracture, and Inscrutability. Applying unvetted or source-corrupted training data.	vectors and attack surfaces used in developing AI-Native constructs across critical 6G communications infrastructure, including over-the-air/user interfaces. Deficiencies or inadequacies in the planning, structure, implementation, execution, or maintenance of critical communications infrastructure at/near user interfaces. Using unvetted or source-corrupted training data.	emulation systems intended to optimize 6G network performance. Deficiencies or inadequacies in the planning, structure, implementation, execution, or maintenance of AI simulation training data and/or AI-Native tools or system leading to malfunctions or other unintended consequences that affect critical communications infrastructure operations. Using unvetted or source-corrupted training data.

Table 9 - Threat Model/Risk Analysis Categories vs. 6G Network Use Cases

Considerations for 6G

The considerations below are tempered by the uncertainties in the 6G network concept designs and the envisioned level of AI use for the selected scenarios. By the time 6G is deployed, AI-Native will be pervasive in its architectural design, and external network threats will have evolved to a level of sophistication where some forms of advanced AI attacks will likely be used. The following are suggested for further consideration in the context of the intended use of AI in 6G networks to ensure operational efficiency, security, and resiliency for the above use cases as first steps in thwarting downstream AI-powered and non-AI cyber threats:

- Adaptive AI Anomaly Detection
- Adapting Adversarial Machine Learning (AML) mechanisms
- Radio Frequency Machine Learning Operations (RFMLOps) approaches:
    - Consider applying optimized-AI based multi-objective optimization for threat detection/classification.
    - Consider incorporating a unified, multitask learning framework to analyze and classify external threats in real-time based on current RFMLOps research and development

specifically for this purpose applied to 6G networks.

- Examine the role of open-source generative adversarial networks (GANs) for unsupervised synthesis of raw-waveform audio, as opposed to image-like spectrograms. These ML algorithms learn to synthesize raw waveform audio by analyzing numerous examples of real audio and can be extended to handle raw RF within an RFMLOps framework. The use of such open-source tools can support AI design, enhance security by mitigating inherent AI vulnerabilities, and strengthen defenses against AI-based external threats.

- Distributed AI-Native design for data protection (validating training data sources), model security (periodic retraining), operational controls (semi-autonomous parameter limiting on rate/power), and architectural guardrails on hierarchy of functions, local validation, and failover mechanisms.

- Leverage ML systems for discerning authorized from unauthorized network users and preventing unauthorized access.

- Examine the use of AI-enabled dynamic spectrum evasion techniques to further mitigate spectrum exploits.

- Research supervised learning for signal traffic classification or unsupervised methods for identifying novel anomalies, including AI-spectrum management solutions designed for access denial

- Consider incorporating AI techniques with post-quantum security designs and use predictive AI techniques to execute proactive defenses, including against threats that may be AI-based (i.e., build in "anticipatory" AI-based features in the 6G AI-Native design to predict threat vectors and attack strategies). In particular, consulting existing literature on preparing 5G for post-quantum security may be useful.[62]

- Consider use of cloud-native architectures requiring zero trust models, secure API gateways, and hardware root-of-trust mechanisms to mitigate threats arising from virtualization and containerized deployments to address supply chain risks, potential for backdoor vulnerabilities, and configuration drift across diverse components.

[62] *See* ATIS, *Preparing 5G for the Quantum Era: An Analysis of 3GPP Architecture and the Transition to Quantum-Resistant Cryptography,* 2025, https://atis.org/resources/preparing-5g-for-the-quantum-era-an-analysis-of-3gpp-architecture-and-the-transition-to-quantum-resistant-cryptography/ (describing a phased cryptographic migration strategy within the 5G network architecture).