AI Security Assurance Techniques for AI-Native 6G Networks

Paweł Murdzek, Piotr Szewczyk,
Institute of Telecommunications, Warsaw University of Technology

***Abstract*—The evolution towards 6G networks signifies a fundamental paradigm shift from Key Performance Indicator (KPI) oriented systems to Key Value Indicator (KVI) driven networks and native Artificial Intelligence (AI-Native) [2], [6]. In this architecture, AI is no longer just an optimization tool, but the foundation of every network stack layer, which introduces unprecedented challenges in terms of personal data protection and resilience to attacks [4], [5]. This article analyzes a holistic approach to 6G security, covering the concept of Sovereign AI, privacy-preserving Federated Learning (FL) mechanisms, and the role of Explainable AI (XAI) in building trust [1], [5]. We present a detailed security framework based on the REASON architecture and Zero Trust principles, aimed at mitigating threats such as data poisoning, adversarial attacks, and physical sabotage in integrated terrestrial-satellite environments (TN-NTN) [2], [3].**

I. Introduction and AI-Native Paradigm

6G networks introduce a vision of intelligent infrastructure that can perceive, learn, and adapt autonomously in real-time [6]. Unlike 5G, where AI often played the role of an add-on to specific functions, 6G is designed as a natively intelligent system, where AI manages radio resources, mobility prediction, and security at every level [4], [5].

This transition, however, involves the necessity of processing massive datasets of personal information, such as precise location, movement patterns, and biomedical data, which places compliance with regulations like GDPR at the center of system design [5]. These networks must support **Key Value Indicators (KVIs)**, covering trust, digital inclusivity, and technological sovereignty, which are becoming as important as traditional throughput [2].

II. Personal Data Protection in the 6G Lifecycle

A. Risks Associated with Data Abundance

The ubiquity of sensors and IoT devices in the 6G ecosystem enables capturing data with unprecedented detail. As sources indicate, the integration of high-density antennas combined with precise device localization raises the risk of user re-identification, even if raw data are seemingly anonymized [5].

B. Data Processing in Various Phases

Ensuring security requires data analysis throughout the entire network lifecycle [5]:

- **Design Phase:** Utilization of user preferences and device identifiers (IMEI/MAC) for service tailoring.
- **Network Operations:** Real-time location data and connection logs serving capacity optimization.
- **Resource Allocation:** User profiles and channel status determining Quality of Service (QoS) prioritization.

A key challenge becomes the integration of *privacy-by-design* and *privacy-by-default* principles already at the stage of 6G protocol standardization [5].

III. Threat Model for AI-Native Systems

A. Attacks and Model Manipulations

AI systems embedded in 6G architecture are susceptible to specific forms of manipulation that can compromise network integrity [1]. Sources distinguish three main attack vectors:

- **Data Poisoning:** Malicious updates sent during learning processes, aimed at degrading model performance or creating hidden vulnerabilities (backdoors) [1], [3].
- **Evasion Attacks:** Subtle modifications of input data during inference, deceiving intrusion detection models or beam allocation algorithms [1], [5].
- **Model Theft and Inference Attacks:** Attempts to reconstruct model architecture by analyzing AI responses or extracting sensitive training data from model parameters [4], [5].

B. Security in TN-NTN Infrastructure

Integration of terrestrial networks with non-terrestrial systems (NTN), such as drones and LEO satellites, drastically increases the attack surface. These networks are exposed to Jamming and *Man-in-the-Middle* attacks.

A critical challenge lies in the energy and computational limitations of satellite and UAV platforms, which prevent the use of "heavy" cryptographic mechanisms directly in orbit (onboard processing). This requires delegating security tasks or using lightweight, hardware-assisted solutions. An additional vector are optical links (FSO), susceptible to specific atmospheric interference [3].

IV. REASON Architecture

The **REASON** (Realising Enabling Architectures and Solutions for Open Networks) project proposes a 6G network blueprint based on modularity and interoperability. This architecture is divided into four horizontal layers and two vertical planes [2].

A. Horizontal Layers [2]

- **Physical Infrastructure Layer:** Covers servers, switches, cables, and computational assets (Edge/Cloud), ensuring the foundation for Computing as a Service (CaaS).
- **Network Service Layer:** Defines logical service design, manages data formats and interfaces, ensuring connectivity continuity and API security.
- **Knowledge Layer:** Key AI-native element, integrating **Cognitive** and **AI** planes [6]. This is where AI model lifecycle management takes place, from data acquisition to their retirement.
- **User Application Layer:** Interface for consumers and enterprises, requiring adaptive Quality of Experience (QoE).

B. AI and Cognitive Planes

The AI plane in REASON utilizes an **AI Orchestrator**, which manages model catalog, versioning, and training pipeline automation. The *Cognitive Plane* is responsible for inferring system context, ensuring ethical and regulatory compliance, and detecting so-called **concept drift** (changes in relationships between input and output data), which may signal a *data poisoning* attack [2].

C. mATRIC Controller

REASON introduces an innovative **mATRIC** (Multi-access Technology Real-Time Intelligent Controller) controller, which extends the Near-RT RIC concept from O-RAN [2]. mATRIC enables intelligent control of multiple access technologies (5G, WiFi, LiFi, Optical) in an integrated manner, optimizing radio resources using Deep Reinforcement Learning (DRL) algorithms [2], [6].

V. Sovereign AI and Generative Challenges

A. Sovereign AI Stack

In the face of GenAI model dominance by external providers, the concept of **Sovereign AI** means full control over the "AI stack" – from hardware to data. The report recommends introducing a **"Sovereign Watchdog"** mechanism – an independent, operator-controlled detection module that acts as an auditor for models provided by external vendors ("black boxes"). This allows blocking statistically suspicious decisions without the need to interfere with vendor source code [1].

B. Generative AI (GenAI) Threats

Risks associated with implementing Generative AI in network management must be emphasized. Large Telecom Models (LTM) are prone to **hallucinations**, potentially resulting in erroneous network configurations (Network-as-Code). Adversaries can also use GenAI to create **synthetic network traffic**, indistinguishable to classic IDS systems, and deepfakes in identity verification processes.

VI. Agnostic Attack Detection on AI Mechanisms

The goal of the project is detection of attacks in an agnostic manner, i.e., independent of specific AI model architecture and attack type (attack-agnostic) [5].

A. Autoencoders and Reconstruction Error

One of the most promising methods is the use of **Deep Autoencoders (DAE)**. The system learns to compress and reconstruct "clean" network traffic. In case of an attack (e.g., pilot poisoning in PHY layer), data containing perturbations exhibit a different statistical structure, causing a sharp increase in **reconstruction error**. This allows detecting anomalies without possessing signatures of a specific attack [1].

B. Feature Squeezing and Entropy Analysis

To detect *Evasion* type attacks, the **Feature Squeezing** technique (feature space reduction) is used, e.g., by reducing input data bit depth. Comparison of model prediction on original and "squeezed" data allows revealing "brittle" adversarial perturbations. A complement is **Entropy Analysis** (EBD), detecting chaos introduced into the signal by adversarial attacks – infected samples often exhibit unnatural entropy spikes [2].

C. Detection of Universal Adversarial Perturbations (UAP) in MIMO

In 6G networks, a particular threat are Universal Adversarial Perturbations (UAP) – single noise patterns that disrupt classification of any signal. In MIMO systems, signals are legally spatially correlated. UAP attacks disturb this natural inter-channel correlation. Detectors based on Chebyshev distance can detect these subtle anomalies in real-time, which is crucial for physical layer protection.

D. Explainable AI (XAI) as Verifier

XAI techniques, such as **Shapley values**, act as a semantic verifier. If XAI shows that the model made a decision (e.g., about beam change) based on background noise rather than significant signal features, this is a strong indicator of an adversarial attack [1].

VII. Distributed Detection Architecture (Distributed Framework)

Effective protection requires hierarchical placement of detection mechanisms in O-RAN architecture, adapted to latency requirements:

- **Edge (Near-RT RIC):** Implementation of lightweight methods with low latency (<10ms), such as *Feature Squeezing* or simple autoencoders, to protect the physical layer and verify CSI reports.
- **Core/Regional (Non-RT RIC):** Running "heavy" models (Deep DAE, advanced statistics) for detection of slow *data poisoning* attacks, long-term trend analysis, and sovereignty policy management.
- **Device (UE/IoT):** Simple entropy analysis (EBD) and preliminary data filtration before sending to Federated Learning process.

TABLE I
HIERARCHICAL DETECTION IN O-RAN

Level	Method and Application	Reaction Time
Edge (Near-RT RIC)	**Lightweight models (Feature Squeezing, Autoencoders):** Protection of physical layer and CSI report verification.	< 10 ms
Core/Regional (Non-RT RIC)	**Heavy models (Deep DAE, Statistics):** Detection of *data poisoning* attacks, trend analysis, and sovereignty policy.	> 1 s
Device (UE/IoT)	**Entropy analysis (EBD):** Preliminary data filtration for Federated Learning.	Real-time

VIII. DISTRIBUTED DETECTION ARCHITECTURE (DISTRIBUTED FRAMEWORK)

Effective protection requires hierarchical placement of detection mechanisms in O-RAN architecture. Table I presents an integration matrix adapted to latency requirements.

It is also worth emphasizing that **blockchain** technology can constitute a "Trust Layer", integrating detection results from the above levels. It ensures log immutability and decision transparency in this distributed environment.

IX. TN-NTN SECURITY AND ZERO TRUST

The vertical REASON security layer enforces **Zero Trust Architecture (ZTA)** principles, where every AI microservice must be continuously verified [2]. AI supports ZTA through predictive anomaly analysis in satellite traffic and automatic decision-making on isolation (*black-holing*) of infected nodes [3], [4].

X. SEMANTIC COMMUNICATION AND RIS: OPPORTUNITIES AND THREATS

Modern physical layer technologies support security but also introduce new threat vectors [6]:

- **Semantic Communication:** Although it reduces data amount (sending intents), it is susceptible to **semantic attacks**, involving manipulation of message meaning without violating its syntactic correctness [4].
- **Intelligent Surfaces (RIS):** Allow directing beams away from eavesdroppers, however there is a risk of **RIS hijacking** – seizing control over RIS controller and intentionally redirecting signal to unauthorized receiver [6].

XI. SUMMARY AND FUTURE DIRECTIONS

Building a secure AI-native 6G ecosystem requires synergy of algorithmic, architectural, and regulatory innovations [3], [4]. The future of security lies in integrating innovative paradigms:

- **Bio-inspired AI Agents:** Immune systems at network edge, learning to recognize "foreign" patterns without prior signature base.
- **Quantum Semantic Communication:** Utilizing quantum phenomena for physical security of information meaning, which may immunize the system against classic adversarial attacks.
- **Simulations:** Verification of methods using frameworks such as **NVIDIA Sionna** combined with attack libraries (e.g., ART) to test effectiveness of proposed autoencoder methods.

Integration of **blockchain** technology can additionally ensure AI decision log immutability and sovereign identity management in the decentralized future network [3], [6].

REFERENCES

[1] Swarna Bindu Chetty, David Grace, Simon Saunders, Paul Harris, Eirini Eleni Tsiropoulou, Tony Quek, and Hamed Ahmadi. Sovereign ai for 6g: Towards the future of ai-native networks. 2025.

[2] Konstantinos Katsaros, Ioannis Mavromatis, Kostantinos Antonakoglou, Saptarshi Ghosh, Dritan Kaleshi, Toktam Mahmoodi, Hamid Asgari, Anastasios Karousos, Iman Tavakkolnia, Hossein Safi, Harald Hass, Constantinos Vrontos, Amin Emami, Juan Marcelo Parra-Ullauri, Shadi Moazzeni, and Dimitra Simeonidou. Ai-native multi-access future networks—the reason architecture. *IEEE Access*, 12:178586–178622, 2024.

[3] Sasa Maric, Rasil Baidar, Robert Abbas, and Sam Reisenfeld. System security framework for 5g advanced /6g iot integrated terrestrial network-non-terrestrial network (tn-ntn) with ai-enabled cloud security. 2025.

[4] Akheel Mohammed, Zubair Ahmed Mohammed, Naveed Uddin Mohammed, Shravan Kumar Gunda, Mohammed Azmath Ansari, and Mohd Abdul Raheem. AI-native wireless networks: Transforming connectivity, efficiency, and autonomy for 5G/6G and beyond. *International Journal of Computer Science & Information Technology (IJCSIT)*, 17(5), October 2025.

[5] Keivan Navaie. Personal data protection in ai-native 6g systems. 2024.

[6] Fabian Chukwudi Ogenyi, Chinyere Nneoma Ugwu, and Okechukwu Paul-Chima Ugwu. A comprehensive review of AI-native 6G: integrating semantic communications, reconfigurable intelligent surfaces, and edge intelligence for next-generation connectivity. *Frontiers in Communications and Networks*, 6:1655410, sep 2025.