# Ericsson Technology Review

#7, August 2026

## Deepfake voice detection in telecom networks

Charting the future of innovation

Ericsson Technology Review — Charting the future of innovation | #7, August 2026

# Deepfake voice detection in telecom networks

**Authors:**
Kevin Kiernan, Stephan Skiba, Magnus Bergstrand

Deepfake voice technologies are fundamentally challenging trust in telecom voice services. As synthetic speech becomes increasingly realistic, traditional reliance on human perception is no longer sufficient to verify authenticity. This shift requires telecom networks to assess voice authenticity in real time. By combining IP Multimedia Subsystem (IMS) based media analysis with Service Management and Orchestration (SMO), network-level deepfake voice detection enables scalable, low-latency trust assurance while supporting new categories of trusted voice services.

ISSN 0014-0171 284 23-3441 | Uen

Ericsson, SE-164 83 Stockholm, Sweden
Phone: +46 10 719 0000



**Voice communication has long been one of the most trusted services that telecom networks provide. Even as digital channels have proliferated, voice calls remain critical in situations requiring immediacy and reliability, including emergency response, enterprise coordination and personal communication. This trust has historically been grounded in the physical nature of voice capture and transmission, which made large-scale impersonation difficult.**

Recent advances in synthetic speech generation are fundamentally changing this premise. Modern artificial intelligence (AI) driven voice generation technologies are capable of producing speech that is coherent, emotionally expressive and closely aligned with the vocal characteristics of specific individuals. As a result, the human ear is no longer a reliable instrument for determining whether a voice is authentic.

This shift has significant implications. Voice services have traditionally relied on an implicit assumption of authenticity. When a call is answered, the listener assumes that the speaker is who they claim to be. As synthetic voice becomes indistinguishable from real speech, this assumption no longer holds. Maintaining trust in voice communication therefore becomes an explicit responsibility of the network, rather than an inherent property of the service.

### Deepfake audio delivered via video services

From a network perspective, synthetic voice in phone calls and synthetic voice embedded in video are both media streams. In either case, the audio is digitally generated and does not originate from a physical capture path. This means the same signal-level techniques used to detect deepfake voice in calls can also be applied when synthetic audio arrives via video services such as enterprise videoconferencing or push-to-talk functionality.

## The rise of synthetic voice as a telecom threat

Synthetic voice generation has evolved through steady advances in machine learning (ML), moving from early text-to-speech systems with limited realism to highly sophisticated neural models capable of capturing subtle acoustic and temporal characteristics. These models can reproduce tone, cadence, emotion and speaker-specific traits with a level of accuracy that was previously unattainable.

## Terms and abbreviations

**AI** – Artificial Intelligence | **API** – Application Programming Interface | **cApp** – Core Application | **dBFS** – Decibels Relative to Full Scale | **I-CSCF** – Interrogating Call Session Control Function | **IMS** – IP Multimedia Subsystem | **MGW** – Media Gateway | **ML** – Machine Learning | **MRF** – Media Resource Function | **O-RAN** – Open Radio Access Network | **PCRF** – Policy and Charging Rules Function | **P-CSCF** – Proxy Call Session Control Function | **RAN** – Radio Access Network | **SAFE** – Synthetic Audio Fraud Evaluation | **SBG** – Secure Border Gateway | **S-CSCF** – Serving Call Session Control Function | **SMO** – Service Management and Orchestration

The increasing accessibility of these technologies has transformed the threat landscape. Voice-cloning tools are now widely available through commercial cloud services and open-source frameworks, enabling both legitimate innovation and malicious use. This accessibility allows attackers to generate convincing impersonations using only short audio samples, significantly lowering the barrier to entry.

Deepfake voice attacks differ fundamentally from traditional telecom fraud. Conventional fraud often relies on exploiting signaling vulnerabilities, protocol weaknesses or identifiable behavioral patterns. These approaches can be mitigated using established network-based detection techniques. In contrast, deepfake voice attacks exploit human perception. By imitating trusted individuals, they create a sense of urgency and authority that can bypass rational scrutiny.

In many cases, a single interaction is sufficient to achieve the attacker's objective. This makes such attacks difficult to detect using systems that depend on repetition or historical data. Calls may originate from legitimate numbers, exhibit normal signaling behavior and leave little trace for conventional analytics to identify.

While endpoint-based countermeasures are emerging, they rely on users making correct decisions under emotional pressure, which is inherently unreliable. This reinforces the need for network-level detection mechanisms that operate independently of user behavior and provide consistent protection at scale.

## Why authenticity belongs in the network

Addressing the challenge of deepfake voice calls requires a shift in perspective. Rather than attempting to determine identity or intent directly, the network can assess whether a voice call exhibits characteristics consistent with a real acoustic environment.

Real-world voice calls are shaped by a combination of physical factors, including environmental acoustics, microphone characteristics, codec processing and network-induced variability. These factors introduce imperfections and variations that are difficult to reproduce accurately in synthetic audio. Even when artificial noise is added, it typically follows constrained statistical patterns that differ from natural acoustic behavior.

Telecom networks are uniquely positioned to observe these characteristics at scale. They process voice traffic across diverse devices, access technologies and environments, providing a broad and consistent dataset for analysis. This enables probabilistic assessment of authenticity based on measurable signal properties rather than subjective perception. A network-level approach also allows detection to be applied uniformly across users and services, avoiding fragmentation and ensuring that trust is maintained consistently across the network.

## Trusted voice as a differentiated service

Network-level authenticity delivers more than threat mitigation: it also enables service differentiation. Voice has long been treated as a basic utility rather than a differentiated service. Network-embedded detection allows operators to introduce trusted voice as a service category built on authenticity assurance. For consumers, this can take the form of optional authenticity indicators that provide reassurance without requiring technical expertise. For enterprises, trusted voice enables assured communication for contact centers and internal coordination, protecting brand integrity and reducing social engineering risk.

Application programming interfaces (APIs) allow enterprise systems to integrate network-level authenticity signals into their workflows.

For emergency services and mission-critical communications, authenticity is not optional. Integrating detection at the network level strengthens confidence that distress calls and command instructions originate from real human sources, preserving operational reliability. Together, these tiers reposition voice as a premium trust service rather than a legacy utility service.

This evolution aligns with the concept of cognitive networks, where AI functions as an embedded sensing capability. Detection occurs where signals originate, within the IMS Core, while orchestration and governance occur within SMO. This separation enables the scalable, adaptive and controlled deployment of intelligence. In this context, trust becomes a measurable and manageable attribute of the network, supporting new service models and long-term value creation.

## IP Multimedia Subsystem as the detection anchor point

IMS provides a natural and standardized location for implementing network-level deepfake voice detection. Defined by the 3GPP (3rd Generation Partnership Project) as the service control framework for voice services in 4G and 5G networks, IMS handles both signaling and media streams for VoLTE (Voice over Long-Term Evolution) and VoNR (Voice over New Radio) calls.

This central role provides several advantages. IMS offers visibility into RTP (Real-time Transport Protocol) media streams under strict latency and reliability constraints, enabling real-time analysis without introducing additional architectural complexity. It also operates within a regulated domain aligned with requirements for emergency services, lawful intercept and service assurance [1].

Detection functions embedded within IMS can extract a range of signal-level and session-level features, including spectral characteristics, temporal dynamics and codec-related artefacts. These features are evaluated using a layered detection pipeline in which lightweight feature extraction is combined with statistical scoring and model-based inference.

ML models trained on authentic and synthetic samples evaluate these features to identify patterns associated with real and generated speech. Importantly, these models operate alongside deterministic checks and heuristics, ensuring that detection remains interpretable and controllable in operational environments.

The use of probabilistic scoring rather than binary classification is critical. Authenticity is not an absolute property but a likelihood that must be interpreted in context.

Network-level authenticity delivers more than threat mitigation: it also enables service differentiation.

By producing confidence metrics, the system enables flexible decision-making and integration with higher-level policy frameworks.

## Authentic versus synthetic voice characteristics

While often imperceptible to human listeners, the differences between authentic and synthetic voices become clear in spectrogram analysis, which forms a key basis for network-level detection. **Figure 1** shows that authentic voice calls, at left, exhibit dynamic patterns across time and frequency domains, while synthetic calls, at right, have much greater statistical stability.

Authentic voice calls are characterized by:

- Irregular background noise – uneven, scattered textures and fluctuating energy patterns distributed unpredictably across the spectrogram
- Unstable harmonics – natural variations in harmonic structure over time
- Non-uniform timing – irregular speech rhythm and timing between phonetic components.

The differences between authentic and synthetic voices become clear in spectrogram analysis.

By contrast, even the most perceptually convincing synthetic calls exhibit clean pauses, overly consistent and smooth harmonics across time, uniform timing of speech elements and consistent rhythm. This distinction enables a shift from perception-based evaluation to objective measurement. By focusing on how signals behave rather than what they convey, networks can identify inconsistencies that indicate synthetic origin.

## Network-level deepfake voice detection mechanisms

Effective deepfake voice detection combines multiple independent signals observed within the IMS network, including background noise, session behavior, source integrity and mobility behavior. Each contributes a distinct perspective on call authenticity.

Background noise analysis evaluates the variability of environmental audio. Authentic calls exhibit continuously changing acoustic conditions, while synthetic audio often lacks this natural variation or shows artificially generated patterns.

Session behavior analysis considers call duration and interaction patterns. Human communication tends to be irregular, whereas automated campaigns may exhibit consistent timing structures across multiple sessions.

Source integrity analysis examines signaling attributes such as numbering ranges, device identifiers and geographic consistency. These provide additional context for identifying anomalies related to spoofing or centralized call generation.

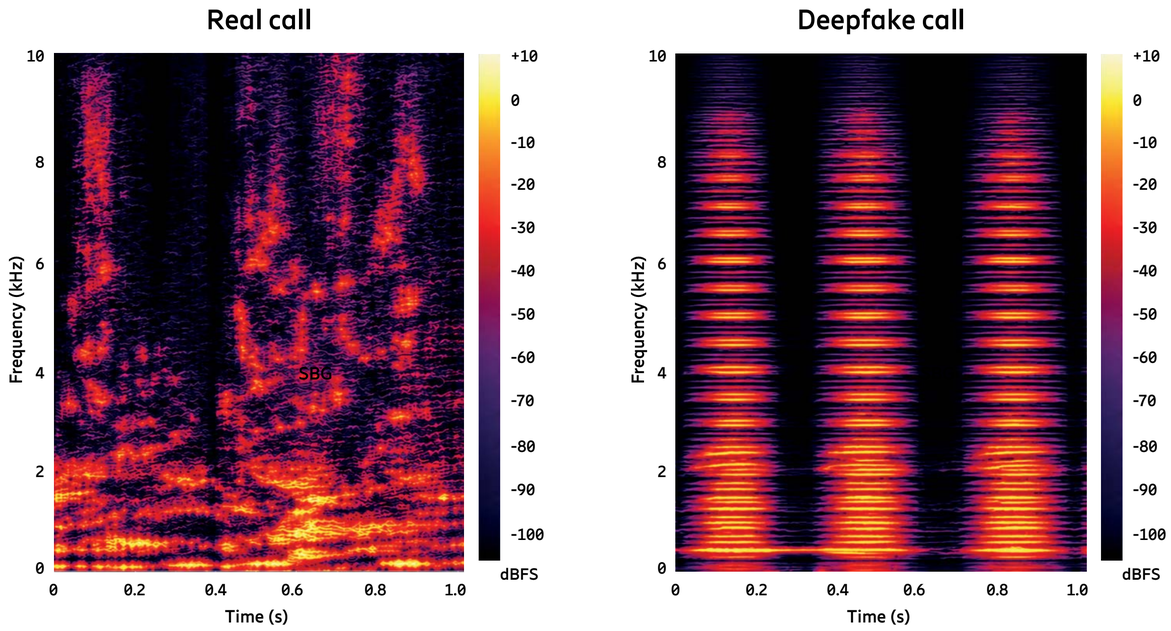


**Figure 1:** Spectrogram comparison of authentic and synthetic voice calls highlighting acoustic authenticity differences

Mobility behavior analysis evaluates radio conditions over time. Genuine mobile users exhibit dynamic radio behavior, including handovers and signal fluctuations. Calls that remain unnaturally stable may indicate non-physical or emulated endpoints.

These mechanisms are not applied in isolation but form part of a coordinated detection pipeline. Cross-signal correlation integrates outputs from multiple domains, producing a unified authenticity assessment. When several indicators align, detection confidence increases significantly while reducing the risk of false positives. All of the mechanisms are applicable for existing 4G and 5G networks, as well as for future 6G networks where it is likely deepfake fraud will become even more advanced.

## Conceptual architecture for network-level detection

Voice services require low latency and high reliability, placing strict constraints on real-time processing. To ensure deepfake voice detection can operate at scale without degrading service performance, our conceptual

architecture, shown in **Figure 2**, combines IMS-based synthetic voice detection with SMO. While commonly associated with the open radio access network (O-RAN), SMO is not confined to the radio access network (RAN) domain. Its capabilities extend across core and service layers, where it can consume analytics, apply policies and coordinate network-wide actions.

Within the IMS Core, detection functions perform passive analysis. They extract features and evaluate signals without modifying media streams or interrupting sessions, which ensures that the user experience remains unaffected while enabling continuous monitoring. This design allows detection capabilities to be introduced incrementally and transparently, without requiring changes to user devices or service behavior.

The role of SMO in this architecture is to aggregate detection insights from multiple sessions, identify patterns such as coordinated fraud campaigns and enforce consistent response policies across the network. This includes adjusting detection sensitivity and enabling coordinated mitigation actions. By operating above the real-time execution layer, SMO allows the network to evolve from isolated call-level decisions to system-level intelligence and adaptive protection.

The potential capabilities of the SMO core application (cApp) Synthetic Audio Fraud Evaluation (SAFE) include:

- flagging suspicious calls before completion
- triggering post-call verification workflows for enterprises
- generating trust metrics for assurance dashboards
- triggering biometric challenges in financial transaction workflows via network APIs
- cross-domain mobility insights.

By separating real-time detection from policy enforcement, the conceptual architecture in Figure 2 ensures scalability and flexibility. IMS performs time-critical signal analysis, while SMO enables governance, auditability and human oversight.

## Use case examples

The following three use case examples illustrate how our conceptual architecture works in practice. Together, they show how network-level intelligence can evolve from protecting individual users to enabling trusted communication and supporting enterprise decision-making. This progression highlights the expanding role of telecom networks as providers of trust, context and decision support across both consumer and enterprise domains.

**Example 1 – detecting an impersonation attempt**

In the first example, which is visualized in **Figure 3**, an elderly subscriber receives a voice call from what appears to be their grandchild, urgently requesting financial help. The voice is convincing, emotionally charged and contextually plausible. From the user's perspective, there is little reason to doubt the authenticity of the call.

As the session is established and media begins to flow through the IMS, the network performs in-line assessment of the audio stream and session characteristics. While the speech content appears natural, the system detects inconsistencies in the underlying media behavior. Background audio lacks the variability typically associated with a real environment and exhibits patterns that remain unusually stable over time.

In parallel, session-level indicators begin to diverge from expected norms. The call originates from a source with atypical characteristics, and the session shows highly consistent timing behavior. Additionally, mobility-related signals remain static, suggesting the absence of a physically moving endpoint.

Taken individually, these signals may not be conclusive. However, when combined, they form a coherent pattern indicative of synthetic media generation or centrally orchestrated activity. Based on this assessment, the IMS classifies the call as suspicious and triggers a real-time response. The subscriber receives a subtle in-call warning indicating that the call may not be authentic. This intervention occurs without interrupting the session, allowing the user to reassess the situation.

High-level detection insights are simultaneously exposed to the SMO layer, where they can be aggregated with similar events across the network. This enables the identification of emerging fraud patterns and supports coordinated policy responses beyond the individual call. At the same time, the network records the event as part of a broader set of observations, enabling continued monitoring and coordinated response if similar patterns emerge elsewhere.

**Example 2 – branded calls with network-assured authenticity**

A subscriber receives an incoming call from what appears to be a well-known bank. The call includes a branded caller ID

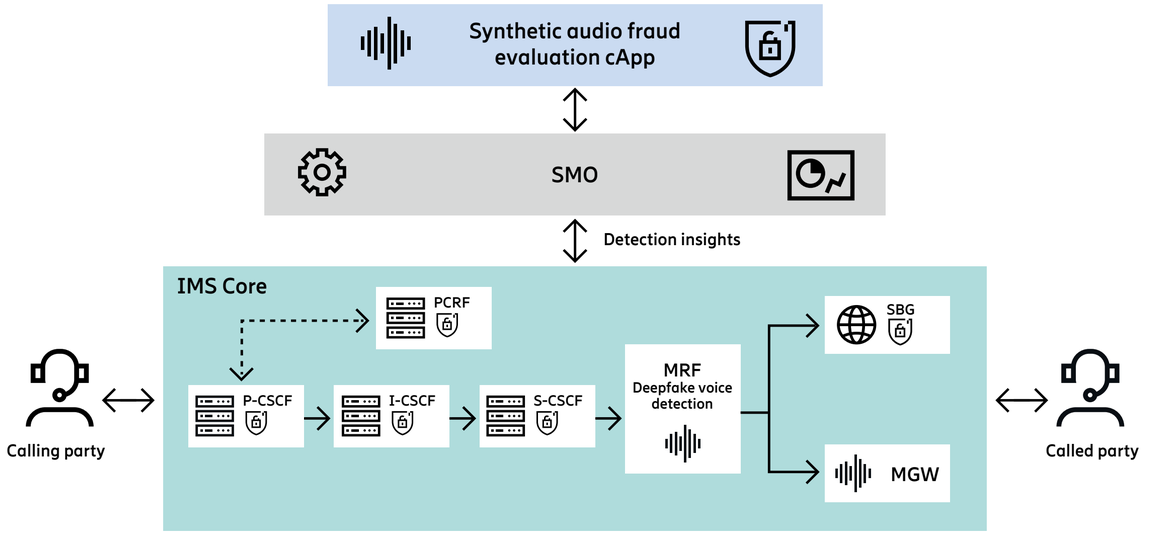


**Figure 2:** Conceptual architecture showing how IMS-based synthetic voice detection can be orchestrated via SMO

and a short message informing the user of unusual activity on their account. The voice is synthetic but clear, consistent and aligned with the institution's communication style. As the call is established through the IMS, the network performs the same in-line assessment used for fraud detection. Media and session characteristics are evaluated, and the system identifies that the voice content is AI-generated. However, unlike suspicious calls, the session is accompanied by verified service indicators. The originating entity is recognized as an authorized enterprise, with validated identity, known service behavior and consistent network-level characteristics.

The network distinguishes this scenario from impersonation attempts by correlating multiple factors:

1. The source is authenticated and linked to a registered service provider.
2. The call setup and signaling patterns align with expected enterprise communication profiles.
3. The use of synthetic voice is consistent with previously observed, approved service behavior.

Rather than flagging the call as suspicious, the IMS classifies it as trusted, branded communication. The subscriber may receive a subtle visual or audio indicator confirming that the call originates from a verified source, even if AI-generated content is used. This approach enables a clear distinction between malicious impersonation, where synthetic media attempts to deceive, and legitimate branded communication, where synthetic media is used transparently and under network assurance. By making this distinction at the network level, operators can support emerging enterprise use cases for AI-generated communication while maintaining user trust and protection.

Classification outcomes and service indicators can simultaneously be exposed to the SMO layer, where they are correlated with enterprise onboarding data, service policies and historical behavior. This enables consistent governance of branded communications across the network and supports scalable trust frameworks without impacting real-time call handling.

**Example 3 – enterprise risk validation via telecom exposure (CAMARA)**

An enterprise customer initiates a high-value transfer through a banking application, requesting funds to be sent to an overseas account. The transaction deviates from the user's typical behavior and triggers internal fraud checks within the bank. As part of this assessment, the bank queries the telecom operator using standardized APIs aligned with industry initiatives such as the GSMA (GSM Association) Open Gateway and the CAMARA framework. Using the customer's mobile number, the bank requests real-time network context indicators relevant to the social engineering risk.

The telecom network evaluates and returns signals such as:

- whether the subscriber is currently engaged in an active voice call
- whether the call originated from outside the subscriber's usual context
- whether the call duration exceeds a threshold consistent with prolonged manipulation.

If all indicators are positive, the situation is classified as high risk, suggesting a potential ongoing scam or coercion scenario. Optionally, additional analysis may be applied within the network, such as in-call voice assessment, to further strengthen confidence in the risk evaluation.

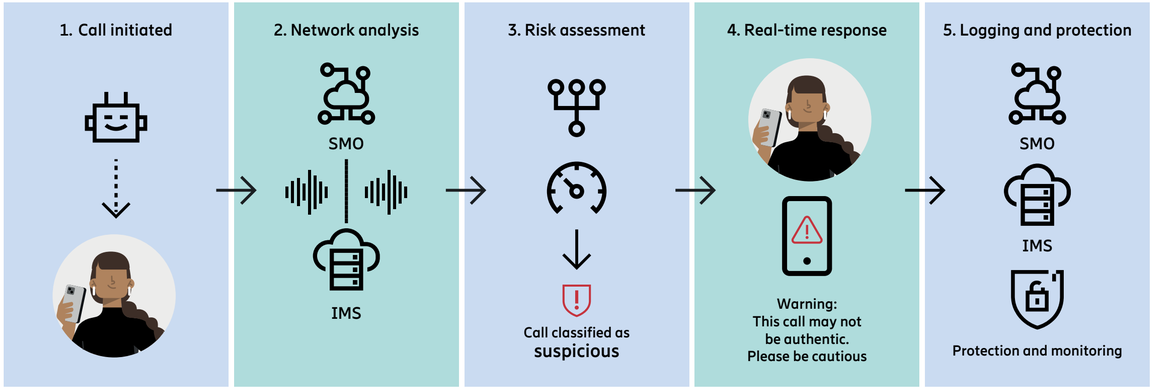


**Figure 3:** Visualization of how network-level deepfake voice detection works to detect an impersonation attempt

This step is not mandatory but can provide an additional layer of validation. Based on the combined risk signals, the bank proceeds with caution: the transaction is registered but placed on hold pending verification by the customer.

This approach demonstrates how telecom-derived context, exposed via standardized interfaces, can act as an independent trust signal for enterprise systems. It enables proactive fraud prevention without requiring direct control of the transaction by the network, while remaining extensible to more advanced media-based detection techniques.

SMO can aggregate such exposure requests and outcomes across multiple enterprises and sessions, enabling cross-domain correlation, policy refinement and adaptive risk thresholds. This allows operators to continuously improve detection effectiveness, support enterprise-specific policies and maintain consistent governance of exposed network capabilities at scale.

## Deployment and performance considerations

A common concern is that deepfake detection must operate on every call, potentially introducing latency or requiring heavy AI processing at scale. In practice, the concept is designed to avoid per-call computational overhead by combining lightweight, in-line checks within the IMS with selective, higher-level analysis in the SMO domain.

Within the IMS, detection is implemented using computationally efficient heuristics and signal consistency checks rather than full-model inference on every media stream. Examples include background noise variability checks, session timing analysis and basic anomaly scoring. These operations are well-suited to real-time environments and can be executed within existing media-handling functions with minimal impact on latency.

More advanced analysis, including pattern correlation across sessions or refinement of detection thresholds, is handled outside the real-time path by SMO. This allows the system to shift heavier processing away from the critical call path, reserving it for aggregated data, where scale can be managed more efficiently.

The architecture also supports selective escalation. Only calls that exhibit suspicious characteristics are subject to deeper inspection, ensuring that most traffic is processed with minimal overhead. This reduces unnecessary load while maintaining detection effectiveness.

From a scalability perspective, the approach distributes responsibility across domains: the IMS handles real-time, per-call decisions, while SMO handles cross-call analysis and policy optimization. This separation ensures that detection remains responsive while scaling to network-wide volumes. Rather than applying heavy AI uniformly, the system uses a tiered processing model, allowing operators to balance performance, accuracy and resource utilization in line with deployment requirements.

Finally, it is important to acknowledge that deepfake audio detection operates within an adversarial environment where techniques continuously evolve. As attackers improve their ability to replicate realistic background noise, detection mechanisms must advance in tandem. The system is therefore designed for ongoing refinement, enabling the identification of subtle artefacts and inconsistencies inherent in AI-generated audio. This adaptive approach ensures sustained effectiveness as synthetic audio generation techniques become more sophisticated.

## Conclusion

Synthetic voice represents a fundamental shift in the threat landscape for telecom networks. By exploiting human perception rather than signaling behavior, deepfake voice attacks challenge traditional detection methods and undermine trust in voice communication. Network-level deepfake voice detection provides a scalable and effective response. By combining signal analysis, behavioral indicators and network context, telecom systems can identify inconsistencies that reveal synthetic media, even when the content appears authentic.

Embedding detection within the IMS (IP Multimedia Subsystem) and orchestrating responses through Service Management and Orchestration (SMO) enables real-time assessment at scale while preserving performance. Our proposed architecture supports integration with enterprise systems, exposure of trust signals and the development of new service capabilities.

As synthetic media continues to evolve, trust will become a defining dimension of network value. Deepfake voice detection is therefore not only a defensive capability, but a foundational element of future telecom networks, enabling secure communication, enterprise integration and trusted digital interactions in the era of artificial intelligence (AI).

# The authors


**Kevin Kiernan** joined Ericsson in 2007 and is a master engineer working with Ericsson Network Manager. His work focuses on large-scale network management systems, automation and the resilience of carrier-grade platforms. He holds an M. Sc. in software engineering from the Technological University of Shannon, in Athlone, Ireland. Kiernan contributes to Ericsson's innovation program and is the author of several patent filings related to AI-driven network management and telecom system optimization.

**Stephan Skiba** joined Ericsson in 1994 and is a portfolio manager within Ericsson Core Networks, based in Stockholm. His work focuses on strategy and roadmap development for data-driven automation and autonomous networks, with an emphasis on enabling intelligent, scalable telecom platforms. He holds a Master's degree in Applied Mathematics from the Technical University of Darmstadt, Germany. Skiba has contributed to the evolution of value-added messaging services (SMS, MMS) and the virtualisation of IMS, and has represented Ericsson at multiple Mobile World Congress events.

**Magnus Bergstrand** joined Ericsson in 1990 and is a strategic product manager within Ericsson Core Networks.  He has extensive experience in product management for telecommunications networks and has worked with Circuit Switched networks (fixed and mobile) and with voice services in 4G and 5G. In his current role, he focuses on Service Exposure, AI-driven spam and fraud protection and AI-powered voice services. He holds an M.Sc in Electrical Engineering from Lund Institute of Technology, Sweden.

## References

1. 3GPP TS 23.228, IP Multimedia Subsystem (IMS); Stage 2 ↗

## Further reading

- Ericsson, Discover the potential of network APIs ↗
- Ericsson, Telecom security ↗
- Ericsson, 6G networks ↗
- CAMARA Project ↗
- GSMA Open Gateway ↗