AI SURGE AND ITS IMPLICATIONS FOR 6G

V 1.0

ngmn.org

WE MAKE BETTER CONNECTIONS

AI SURGE AND ITS IMPLICATIONS FOR 6G

by NGMN Alliance

Version:	1.0
Date:	19 February 2026
Document Type:	Public
Programme:	6G
Approved by / Date:	NGMN Board, 16 February 2026

The information contained in this document represents the current view held by NGMN Alliance e.V. on the issues discussed as of the date of publication. This document is provided "as is" with no warranties whatsoever including any warranty of merchantability, non-infringement, or fitness for any particular purpose. All liability (including liability for infringement of any property rights) relating to the use of information in this document is disclaimed. No license, express or implied, to any intellectual property rights are granted herein. This document is distributed for informational purposes only and is subject to change without notice. Readers should not design products based on this document.

CONTENTS

EXECUTIVE SUMMARY

This is a pivotal moment in the telecommunications industry, propelled by the unprecedented AI surge and the beginning of 6G standardisation. AI is advancing at a rapid pace and will remain a dominant force, reshaping society far beyond the 6G era.

This document consolidates NGMN's perspectives on how AI will likely impact 6G standardisation, providing guidance for ongoing 6G studies. This study examines three key dimensions: **(1) impact of AI traffic on networks, (2) network for AI, and (3) AI for network and implications for 6G architecture evolution.**

Impact of AI Traffic on Networks

The rapid proliferation of AI applications – particularly those with autonomous, task-driven capabilities - introduces significant uncertainty into future network demand. While the precise impact of these AI-driven workloads on traffic patterns is difficult to predict, several factors could materially alter today's assumptions:

- Multi-modal AI applications: Services requiring real-time video exchange may drive substantial traffic growth and shift traditional traffic patterns.
- AI-enabled devices and use cases: Consumer applications (e.g. AR glasses) and enterprise scenarios (e.g. autonomous vehicles) could require frequent upload of images and video after local processing, increasing uplink demand and challenging current downlink-heavy network designs.
- Geographic and device density: AI-intensive areas and device clusters may experience sharp, localised surges, creating increasingly uneven traffic patterns.

Given these uncertainties, network design must prioritise flexibility. Standards Development Organisations should explore mechanisms that allow semi-permanent adjustments in uplink/downlink ratio without requiring major standard revisions, as well as solutions to enhance the uplink coverage. This adaptability will be critical to accommodate evolving AI-driven requirements across diverse devices, networks and regions.

Network for AI

6G should go beyond providing connectivity services to deliver new AI enabled services and capabilities (e.g. new data exposure), by designing networks that are more intelligent, flexible, and trustworthy.

Key design enablers include:

- Flexible (e.g. token-based) charging models reflecting real resource use.
- Dynamic and intelligent networking for AI agents collaboration.
- Support for explicit QoS and computing demand from an AI-based application, to facilitate meeting the required QoS at minimum cost and environmental impact.
- Enhanced QoS and adaptive policy control to support traffic routing achieving seamless connectivity.
- Unified data and model frameworks across devices and domains.
- Secure trust, authentication and authorisation mechanisms for AI agents' digital identity.

AI for Network and Implications for 6G Architecture Evolution

AI is expected to be an important network capability for 6G networks, enabling more efficient usage of network resources, network automation, intent-based management and intelligent orchestration. AI could be applicable to all domains and different layers of the network, including the operation and maintenance.

NGMN expects that 6G will be AI-ready, and the 5G Service-Based Architecture (SBA) will be considered as the starting point towards 6G architecture.

Challenges and considerations for adopting AI:

- Adoption of AI capabilities should allow agents and large language models (LLM) to be deployed in a way that avoids unnecessary impact on the existing architecture. This should not restrict the possible integration of AI-related features embedded within network functions (NFs).
- AI interfaces (e.g., A2A, MCP) will complement existing and future APIs, ensuring readiness for the traffic volumes and capabilities required by emerging AI services.
- Multi-vendor interoperability frameworks are needed to ensure secure, scalable, and open ecosystems.
- Deployment strategies must align with cost and sustainability goals, and validation of real-world performance gains is essential.
- Continued support for non-AI alternatives if these alternatives are necessary to ensure reliability, flexibility and openness.
- Coordinated UE–network operation is needed, i.e., to efficiently execute AI models in both two-sided and one-sided models.

Recommended Standardisation Focus Areas

- Standardised architecture, protocols, and interfaces enabling efficient end-to-end support of AI functionalities, integrated across all domains (RAN, Core, Transport) and all network layers, including devices.
- Standards that support explicit network QoS and computing demand from an AI-based application, to facilitate meeting the required QoS at minimum cost and environmental impact.
- Standards that allow adaptability to support changing traffic patterns, accommodating uncertainty in the impact of evolving AI use cases.
- Evolution of the existing (5G SBA) network architecture should be justified by value driven AI use cases and service scenarios, ensuring alignment with societal and business needs.
- 6G standards that support agent-to-agent and agent-to-network communications.
- Functional and performance requirements for AI capabilities across the 6G system.
- Establishment of interoperability and trust frameworks to enable secure, multi-vendor, and multi-agent deployments and operations (including models retraining, fine tuning).
- Emphasis on the reuse, adoption, or enhancement of “AI interface” from telco and non-telco worlds where appropriate and mainstream. (e.g. (A2A) Agent-to-Agent or (MCP) Model Context Protocol).

01 INTRODUCTION

The rapid evolution of large-scale AI models is driving a paradigm shift toward an “AI-native” era. The proliferation of large language and multi-modal models is enabling the emergence of AI agents—autonomous, collaborative, and self-learning entities that may outnumber human users in upcoming years. This shift toward pervasive, agent-driven ecosystems will fundamentally reshape industries, services, and everyday life.

To support this transformation, networks may need to progressively introduce AI features for intent-driven programmability, autonomous operation, and dynamic compute distribution across central and edge domains. This evolution aims to deliver differentiated connectivity, high reliability, energy efficiency, and simplified operation, positioning 6G as the best network for AI and a foundation for AI-based applications, management, and innovation.

As 6G standardisation enters a critical phase, the growth of AI and AI agents presents both opportunities and challenges for mobile network operators (MNOs). NGMN has outlined key 6G objectives and architectural design principles emphasising innovation across networks, AI, computing, sensing, modularity, operational simplicity, sustainability, trustworthiness, cloud nativeness, network-as-a-service, automation, smooth migration, and a disaggregated multi-vendor approach. These principles aim to guide the evolution of networks that are efficient, sustainable, cost-effective, and socially beneficial [1][2][3][4][5][6][7].

To address the implications of AI on future network design and ensure alignment with NGMN’s objectives, this document examines three dimensions from an operator’s perspective and highlights recommended standardisation focus areas to support industry alignment:

- Impact of AI traffic on networks
- Network for AI
- AI for network and implications for network architecture evolution

02 IMPACTS OF AI TRAFFIC ON NETWORKS

2.1 TRAFFIC GROWTH

Today, mobile data consumption is dominated by video applications, accounting for 70-75% of total traffic [8][9]. A handful of social media and streaming services contribute more than 50% of this demand.

Although AI applications have grown exponentially, their current impact on mobile network traffic remains modest – with primary interactions being text-based [10]. This could change as AI services proliferate, but predicting the scale of impact remains highly speculative due to several factors:

- **Optimisation of AI Models**

  AI models are being optimised using techniques such as quantisation, pruning and reduced token sizes to enable efficient high-performance inference directly on device. [11]

- **Local Processing**

  More complex AI models are expected to run natively on device as chipsets evolve with larger and more capable Neural Processing Units (NPU), faster on-chip memory and cache, increased RAM allocation for AI workloads and tighter integration of hardware with AI frameworks and runtime engines.

- **Unclear Adoption Curve**

  End-user adoption curve: it remains uncertain which new services provide true additional value for end-users, impacting services adoption, traffic curves and commercial models.

- **Regulatory and Privacy Constraints**

  Several challenges would need to be resolved, for data-heavy AI features, such as automatic image or video capture via AR glasses.

Against this uncertainty, the potential impact of AI on traffic growth needs to be considered in the following aspects:

- **Substitution of Current Demand**

  Multi-modal AI applications are likely to proliferate capturing more user attention, with smartphones likely remaining a primary interface. However, it is expected that most video traffic from these applications will replace existing user behaviour – such as watching social media video feeds – rather than creating truly incremental demand.

- **Potential Rise of Wearables**

  AR glasses and similar interfaces could dramatically increase traffic if they continuously interact with cloud-based AI applications using video or images. This traffic would be considered incremental, rather than substitutional, but adoption hinges on overcoming privacy and security concerns as discussed above.

- **Enterprise and Other Applications**

  Autonomous drones, connected cars, humanoid robots / cobots and industrial AI use cases could add significant traffic—provided technological and regulatory hurdles are cleared.

- **Uplink Trends**

  Current uplink demand is moderate, but future use cases such as AI agents could reverse this trend [10].

  AI agents with advanced perception and reasoning capabilities may reside on smartphones or wearables, continuously gathering data and interacting autonomously - potentially generating far more data than humans, subject to battery capacity and computational power of the device.

  However, this shift is uncertain, as many AI agents could instead operate in the cloud, performing inference and delivering recommendations to the user.

Future scenarios differ greatly in both likelihood and scale of impact. Use cases that drive truly incremental video traffic beyond today's demand will exert the greatest pressure on networks. While some scenarios present significant potential for increased demand, they must be weighed against their likelihood when setting priorities for network evolution.

This uncertainty makes flexibility a cornerstone of 6G standardisation – ensuring the network can adapt seamlessly to diverse and unpredictable requirements.

2.2 SHIFT IN NETWORK REQUIREMENTS

The rise of AI may introduce fundamental changes in both the form and direction of traffic:

- **Machine-oriented Media**

  Traditional networks primarily carry human-perceivable content (text, images, audio, video). In contrast, agent-to-agent communication may involve exchanging models, feature vectors, latent representations, and other forms of information optimised for machines rather than humans.

- **Uplink-heavy Behaviour**

  While today's traffic is mostly downlink-dominated, many AI-enabled use cases are assumed to reverse this pattern. For instance, AR glasses with AI may require continuous uplink transmission of environmental images, and AI-inferenced autonomous vehicles may upload real-time video and sensor data more often, in contrast to traditional connectivity patterns.

  6G networks should support these use cases with sufficient flexibility to increase uplink traffic as a major design driver for 6G networks. For example, increased uplink (UL) slot occurrences that maximise the UL transmission opportunities to manage the increased UL traffic expected with new services.

  Some of the proposals that are being discussed in industry and under review in 3GPP are around the definition of flexible and dynamic downlink (DL) / UL patterns, for example, Full-Duplex or Sub-band Full Duplex operation. Enhancing UL coverage is also a desirable feature.

- **Regional and Sectoral Variability**

  The impact of AI traffic will differ across regions and industries. Urban centers are likely to experience more AI traffic surges than rural or remote areas. Certain industries such as manufacturing, transportation, healthcare, and smart cities may generate higher volumes of AI traffic. AI-intensive areas and device clusters may experience sharp, localised surges, creating uneven traffic patterns.

03 NETWORK FOR AI

AI-driven applications impose new requirements on 6G networks, encompassing not only improved connectivity performance but also new capabilities beyond connectivity.

3.1 PERFORMANCE VS. BUSINESS VALUE

For performance improvements related to traditional connectivity, it is essential to validate the necessity of any network enhancements from a business value perspective in order to avoid unnecessary investment and resource waste. While network optimisation can improve user experience to some extent, not all scenarios require extreme performance gains, as the existing services offered may not be directly impacted by these network enhancements. For example, humans generally have a relatively high tolerance for latency in audio and video conversational interactions through Internet —compared with face-to-face communication, users can typically accept an additional delay of the order of a few milliseconds without a significant impact on experience.

6G aims to enhance network performance, especially in relation to network capacity and latency due to the emergence of new services, such as AI applications. However, in today's text-based conversational generative AI services, the dominant factor affecting response time is not the network latency but the processing delay of computationally intensive AI models that require specialised and high-performance infrastructure. In this case, the bottleneck lies in the AI services and computing infrastructure rather than the network. Assuming these bottlenecks will be resolved in the future, some services will require tighter network performance control. For instance, in the case of conversational AI services for real-time immersive experience through XR devices (AR, VR and others) the network throughput and latency requirements will become more stringent, so that the network Quality of Service (QoS) will need to be tightly managed to ensure good user experience. In general, it is therefore important to identify the impact of network performance on the user experience of an AI service.

Beyond ensuring adequate connectivity performance, the true value of 6G for AI-based services lies in delivering the required capabilities to efficiently support these new services.

3.2 CAPABILITIES BEYOND CONNECTIVITY

To support AI and AI agents effectively, 6G should integrate capabilities such as dynamic networking, advanced QoS, distributed computing, trust management, and intelligent orchestration.

- **New Charging Models**

  Charging rules for mobile AI services and applications should reflect their specific demands on network resources. For instance, a token-based charging model could be investigated, where tokens correspond to fine-grained units of resource consumption, such as bandwidth, latency guarantees, or edge computing capacity. This approach facilitates flexible, transparent, and scalable transactions among users, AI agents, service providers, and network operators, promoting fair cost allocation while incentivising efficient resource usage.

- **Dynamic and Intelligent Networking**

  Future networks are expected to support dynamic and intelligent collaboration among physical AI agents by enabling the on-demand creation of intent-driven private networks. These networks may be short-lived and mission-specific, supporting scenarios such as collaborative humanoids/robots, drone swarms, robotic dog swarms, autonomous vehicle fleets, and industrial embodied AI agents. Compared with static grouping models, such ephemeral network groups are expected to support dynamic joining and leaving of agents, adapt to changing service requirements and environmental conditions, and minimise manual provisioning and operational overhead, while dynamically adjusting membership, connectivity, and performance characteristics based on task objectives, agent mobility and proximity, real-time service requirements, and trust and authorisation policies.

- **Enhanced QoS Mechanisms**

  AI services are expanding beyond text to become multi-modal and it is expected that the communication of different types of AI-related content will need different treatment. 5G and previous generations have supported QoS and network slicing mechanisms to support traffic differentiation, but there is no means to have clear end user and network-based policies that enable the routing of traffic into the most suitable connection (e.g. into the corresponding slice or QoS flow). 6G should target a much better use of existing QoS and/or slicing mechanisms and enable advanced policy control with finer granularity for priority handling and multi-modal information handling and synchronisation. To achieve this, improving collaboration with Over-the-Top (OTT) applications and device manufacturers is needed to face the future demand in the best way. Furthermore, in case the network involves itself in AI service tasks in addition to offering connectivity, methods will be needed to assure the performance of the AI tasks undertaken, through task level monitoring, measurement and prediction.

- **Edge Computing**

  AI and AI agent services depend heavily on computing speed, particularly for low-latency inference. Edge devices / and user equipment are limited in computing power, while centralised cloud processing introduces latency and bottlenecks, which requires careful assessments depending on geo-location. 6G should support distributed edge computing to enable real-time processing, collaborative intelligence among agents, and efficient resource utilisation close to the data source.

- **Unified and Distributed Data Framework**

  Achieving "Intelligence everywhere" requires both data and compute resources to be available ubiquitously. This implies transparent data sharing across different domains, which requires architecture enhancements supported by new protocols and/or interfaces. AI agents and applications need to share data, models, inferences, and intermediate results across heterogeneous devices. Without a unified framework, data may remain siloed and inconsistent. 6G should introduce an end-to-end data framework to enable efficient and flexible data, model, and inference sharing, management, processing, and storage across UE, RAN, Core Network functions (NFs) and application functions (AFs).

- **Trust and Authentication**

  AI agents acting on behalf of customers require mutual authentication with networks. Strong encryption and integrity checks are essential for sensitive prompts and personal data. Trust frameworks are necessary for agent-to-agent communication to identify and block malicious AI content. Compliance and lawful interception capabilities must be in place to meet regulatory obligations.

- **Dynamic and Intelligent Resource Allocation**

  Adaptive scheduling is needed to handle bursty AI traffic, prioritising latency-sensitive prompts and inferences while efficiently utilising shared resources. Orchestration between edge and cloud AI models enables dynamic workload distribution, optimising performance, scalability, and resource efficiency, while dynamic scaling of network functions can help improve energy efficiency.

- **Resilience and Reliability**

  Mission-critical AI applications—such as those in healthcare or autonomous control—require continuous availability and failover mechanisms to maintain user trust.

- **AI Traffic Optimisation and AI Agent Interaction**

  6G should support fine-grained traffic analytics to distinguish between model updates, inference requests, and agent communications, enabling optimised management. The network should also support an AI agents interaction framework that facilitates seamless interaction between AI agents, the network and third-party applications.

04 AI FOR NETWORK AND IMPLICATIONS FOR 6G ARCHITECTURE EVOLUTION

AI is not only a consumer of network resources but also a core enabler of network evolution. AI will not just help to improve performance of new networks but also to enable new services and use cases that were not possible with previous generations, such as digital twins and sensing.

However, 6G is not just about AI. Some important lessons learned from 5G show that network evolution should focus also on aspects such as network simplification and energy efficiency. These two aspects may contradict AI requirements to some extent. AI requires the introduction of new network entities and interfaces which lead to architectural changes, adding complexity to network evolution. Additionally, AI engines typically require more computational resources, leading to some increase in energy consumption.

Therefore, with regard to 6G deployments it is important to recognize that AI workloads should be deployed where they are most efficient—across network domains, layers, and physical sites from central clouds to the edge and even end devices – and whenever they add some value in terms of network performance and user experience, hence looking for a good trade-off between business value, network complexity, energy consumption and cost.

In 6G networks, AI is proposed to be deeply integrated into the various layers and domains of network: RAN, transport, core, and management and orchestration. Depending on the level of integration, AI could bring more benefits or could pose more challenges, hence it needs a careful evaluation of what the requirements are at each domain.

4.1 NETWORK MANAGEMENT LAYER

Network management is a predominant layer responsible for overseeing all network assets, and the actions it takes can significantly improve network performance and operational efficiency given that it controls the entire network, making it possible to adapt to service requirements and scenario constraints. For this reason, in this layer, AI should not merely be considered as an auxiliary tool, but rather as a foundational capability that enables autonomy and intelligence shifting from rule-based networks towards autonomous operation.

With this integration of AI, 6G networks can evolve from passive response to proactive decision-making, supporting intent-driven management, automation and intelligent orchestration.

- **Intent-driven Management**

  As networks grow more complex, automation becomes essential to achieve near-autonomous operations. Human oversight will remain, but operators will express high-level intents rather than prescribing specific actions. However, special care is needed for intents with potential conflicting targets, e.g. performances and energy savings. With high degrees of automation and use of AI agents, the attack surface of networks and the associated risk become greater. Failsafe mechanisms should be enabled in order to both allow the interruption of multiple agents as well as operating the network in agent-less mode.

- **Automation**

  AI-enabled automation is used to improve resource planning, anomaly detection, and self-healing, minimising human intervention and reducing operational costs.

- **Intelligent Orchestration**

  AI and AI agents enable cross-domain orchestration of network, computing, and storage resources ensuring efficient utilisation and adaptive service delivery.

- **Energy Optimisation**

  AI has shown benefits for energy saving in 5G and is expected to deliver further gains in 6G.

4.2 CORE NETWORK

The core network domain can benefit from AI adoption in several areas:

- **Network Exposure**

  As AI is increasingly adopted in third-party software components (e.g. adopting AI agents), the way in which they interact with the network is also changing. This requires that network APIs must evolve to meet these new requirements from third parties. It is expected that AI agents will consume tools, so it is required to evolve the exposure layer to manage these new requirements efficiently.

- **Operation and Optimisation**

  AI tools such as network digital twin can be useful for different use cases such as root cause analysis, predictive maintenance, capacity planning, and to assess the impact of new features' activation or changes in the architecture. AI integration will also be useful for the use of core resources and optimise procedures (e.g. paging, mobility management, and deep packet inspection (DPI)). However, many of these capabilities may be more related to implementation aspects rather than architectural requirements.

- **Architectural Evolution**

  Core network is also required to evolve. Details on network architecture are discussed in section 4.5.

4.3 RADIO ACCESS NETWORK

RAN will also benefit from AI integration as this can help to make more efficient use of radio resources and improve air interface management. However, not all RAN layers or functions are expected to benefit equally from AI support, hence it is recommended to apply AI selectively, focusing on areas where it will deliver clear value and the benefits (i.e. performance improvements relative to computational cost) are notable.

AI is suitable for RAN in those cases where large data volumes need to be processed and/or need to be resolved, such as for instance, at layer L2 (MAC-layer). By contrast, functions that already operate close to the optimal limits with well-defined, standardised models, or that involve primarily linear problems—such as channel coding, HARQ procedures or basic synchronisation—are not expected to see substantial performance gains from AI-based algorithms. Nevertheless, AI-based implementations of such functions are not precluded and may still be considered to improve flexibility, adaptability or implementation efficiency.

Due to the nature of RAN, implementing AI models at the base station requires these models to meet strict latency requirements and performance so as not to impact RAN functions negatively. This basically implies that AI inference must be executed locally at the edge to enable real-time operation.

RAN may also require support from UE (two-sided mode), as well as from the core network or network management to complement AI processing. This may involve sharing data to/from these entities or directly sharing model or inference results. Therefore, it is crucial to assess whether the current interfaces and protocols allow for optimal communication for AI-transactions. If not, it will be important to evolve these interfaces or even propose new ones to make this communication efficient without impacting overall network performance negatively.

4.4 KEY CHALLENGES AND CONSIDERATIONS

While the benefits of AI integration are significant, its adoption also introduces several key challenges that require careful consideration.

- **AI Performance in RAN**

  Some evaluations reveal that many of today's AI models in RAN trained on idealised datasets and narrow operating conditions may exhibit generalised limitations across different deployment conditions, leading to context-dependent benefits, sometimes relatively modest gains, and potential increases in energy consumption. These findings highlight the need for comprehensive gain validation in real network environments, robust cross-domain data collection, and unified AI lifecycle management and interoperability frameworks.

- **Responsible AI**

  AI technologies introduce potential risks to individuals and ecosystems. It is essential to ensure the quality, transparency, and trustworthiness of AI systems. Networks must comply with regional AI regulations and adhere to operators' data and AI ethics.

- **Cost and Sustainability**

  Deployment strategies must align with cost and sustainability goals, and validation of real-world performance gains is essential. More specifically, when evaluating the benefits of introducing AI to enhance an existing network service, the net CO_2 impact of AI in terms of cost and savings should be evaluated in addition to its net financial impact.

- **Non-AI Support**

  Continued support for non-AI alternatives is required where these alternatives are necessary to ensure reliability, flexibility and openness.

- **Support for Explicit AI Service Demand**

  Network support for AI-based services would benefit from knowing their actual needs, e.g. in terms of network QoS and computing demand. It would facilitate the network resource allocation, thereby allowing meeting the required QoS at minimum cost and environmental impact.

- **Considerations for Interconnection with Legacy Systems**

  Some existing legacy systems cannot interpret AI-based requests (e.g., control or management requests in the context of autonomous networks) due to hardware constraints, which makes interworking with new systems more challenging. Therefore, converged management interfaces should be designed to control both legacy and new systems, while also considering the gradual replacement of legacy hardware where required.

- **Pace of Technological Change**

  AI capabilities are advancing very quickly indeed. There is a risk of standardising 6G functionalities and protocols for AI that will be out of date at the time 6G networks are deployed.

4.5 IMPLICATIONS FOR 6G NETWORK ARCHITECTURE EVOLUTION

6G network architecture should ensure the proper integration of AI across all domains and network layers. 6G is not expected to be a clean slate/ or disruptive revolution; however, the architecture should be flexible enough to incorporate new services. Based on the analysis in previous sections, the following architectural requirements should be considered:

- **Start from SBA**

  The 5G Service-Based Architecture (SBA) will be considered as the starting point for 6G architecture, serving as the foundational framework for the 6G core.

- **Intent-based Interaction and Agentic Communications**

  AI and AI agents are expected to be pervasive across the 6G network, enabling intelligent communication and coordination between key components, especially between external systems such as UE and third-party applications. These interactions will facilitate intent-driven management, intelligent network and service control, autonomous operations, dynamic resource orchestration, and simplified UE-network interactions.

- **AI Agent Framework in Core**

  AI communication protocols could be adopted independently while leveraging the current SBA as shown in the figure below. This may also provide support to AI agents in selective functions to address new use cases, when justified.

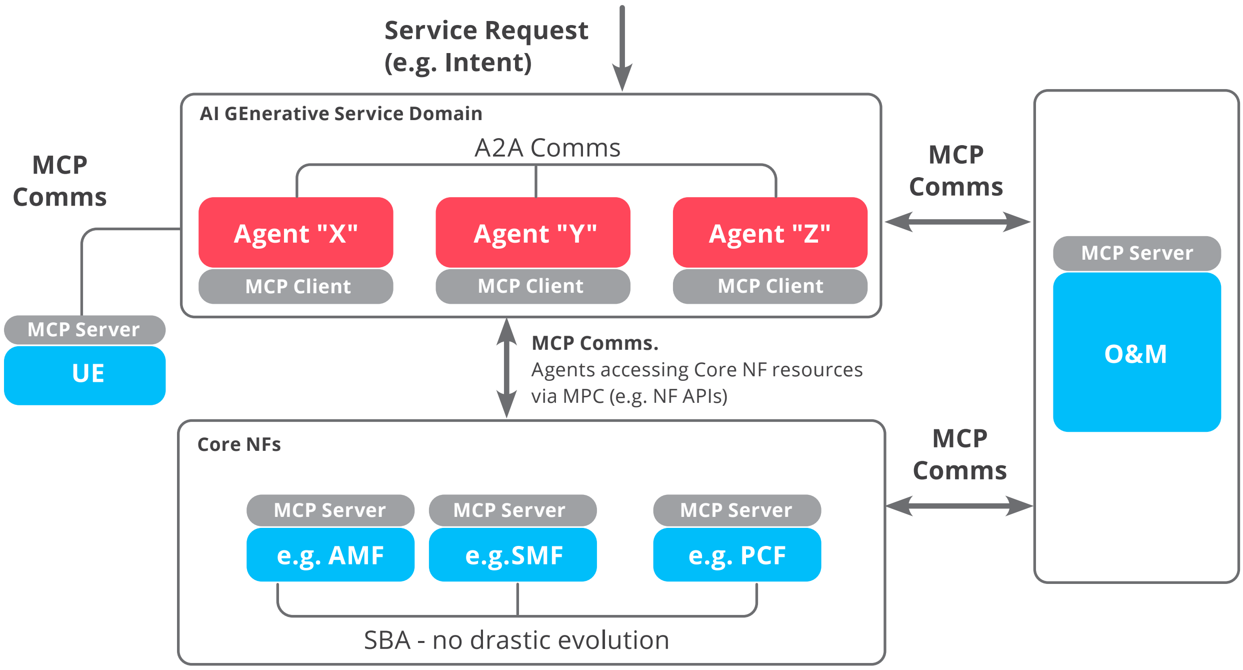

Fig. 1 An Example of AI Agent Framework in Core

- The Model Context Protocol (MCP) is a potential mechanism for integrating AI agents with network function resources (e.g. NF APIs). To enable AI agent communication between the network and the UE—thereby simplifying UE configuration—the MCP protocol can also be utilised.
- For use cases and procedures where current APIs are suitable, SBI architecture should be prioritised.
- For new NFs or use cases not supported today, when clearly justified, the potential use of AI agent-based (agentic) communications could be assessed in core network architecture, ensuring multi-vendor interoperability.

- **AI in RAN**

  The evolution from 5G to 6G shifts AI deployment from isolated single-node processing to a collaborative multi-node intelligent cluster, enabling dynamic sharing of intelligence and computing resources across network elements. AI integration in the RAN should consider two layers: an upper layer, responsible for controlling and optimising RAN functions, and a lower layer, responsible for managing radio resources and the air interface. The requirements for AI/ML implementation differ significantly between these two layers in terms of data management, real-time execution, and computing capabilities. RAN node may require sharing AI models, inference or RAN data with multiple CN NFs and/or upper layers for cross-domain AI-implementation. In that case, RAN-CN interfaces will need to evolve to support AI-services more efficiently. Hence, it is proposed that the corresponding control plane interfaces will evolve to consider new transport protocols, such as Quick UDP Internet Connections (QUIC), or any other enhancements which are more efficient for future evolution.

- **Standardisation for Multi-Vendor Deployments**

  3GPP should conduct a comprehensive study to determine whether communication protocols for AI agents should be formally standardised or addressed through the future de facto industry practices —similar to how microservices, container-based architectures, and for example Kubernetes orchestration have been adopted in supplier products during 5G core network development while NFs and NFs' functionalities / operations were standardised by 3GPP. The objective is to avoid constraining future innovation and the emergence of new use cases by mobile operators. Regardless of the ultimate approach, ensuring interoperability across vendors and minimising integration complexity remain critical considerations that must be addressed.

05 CONCLUSION & STANDARDISATION FOCUS AREAS

5.1 CONCLUSION

This document has examined the implications of the AI surge for 6G system design from an operator's perspective, focusing on three key dimensions: AI traffic, network for AI, AI for network and implications for 6G architecture evolution.

Despite the exponential growth of AI applications, their current impact on network traffic remains modest, and the scale and likelihood of future traffic will remain uncertain. This uncertainty underscores the need for flexibility as a core principle of 6G standardisation, ensuring the network can adapt seamlessly to diverse and unpredictable demands.

AI-driven applications will not only reshape traffic patterns—introducing more uplink-intensive and machine-oriented communications—but also demand new network capabilities beyond traditional connectivity.

These include dynamic resource orchestration, intent-driven management, trust and authentication frameworks, and flexible compute integration across edge and cloud domains. Meanwhile, AI will become a key component of 6G networks, enabling autonomous operation, intelligent orchestration, and proactive decision-making across all layers of the system architecture.

To achieve these objectives, 6G architecture should ensure the proper integration of AI across network domains, with the 5G SBA serving as the starting point for the core network.

Enhancements such as AI agent frameworks, AI agent communication mechanisms (e.g., current options such as MCP), intent-based interfaces, and multi-vendor standardisation will be instrumental in enabling seamless AI-driven communication and collaboration between network functions, UEs, and third-party applications.

5.2 RECOMMENDED STANDARDISATION FOCUS AREAS

The transition towards embracing more AI technologies in networks presents both opportunities and challenges for MNOs. On one hand, it promises greater operational efficiency, service differentiation, and new business models; on the other, it requires addressing interoperability, trust, and security concerns across increasingly open and intelligent ecosystems.

To advance this evolution, standards development organisations, such as 3GPP are encouraged to consider the following areas:

- Standardised architecture, protocols and interfaces enabling efficient end-to-end support of AI functionalities, integrated across all domains (RAN, core, transport) and all network layers, including devices.

- Standards that allow explicit demand of the actual needs of AI services in terms of e.g. network QoS and computing.

- Standards that allow adaptability to support changing traffic patterns to accommodate the uncertainty on the impact of evolving AI use cases.

- Evolution of the existing (5G SBA) network architecture should be justified by value-driven AI use cases and service scenarios, ensuring alignment with societal and business needs.

- It is expected that Agent-to-Agent and Agent-to-Network communication are enabled during the 6G era.

- Standardisation of framework for agent discovery, identity, policy and trust, enabling secure and interoperable agent to agent and agent to network interactions across domains and vendors.

- Functional and performance requirements for AI capabilities across the 6G system.
- Establishment of interoperability and trust frameworks to enable secure, multi-vendor, and multi-agent deployments and operations.
- Emphasising the reuse / adoption or enhancement of “AI interfaces” from telco and non-telco where appropriate and mainstream. (e.g. A2A or MCP).
- Requirements and architectural support for using 6G sensing capabilities as a foundational input for a distributed AI data platform, on top of which AI agents can operate consuming data from sensing capabilities.

By pursuing these directions collaboratively, the industry can ensure that 6G evolves into a flexible, sustainable, and intelligent network—one that supports continuous innovation, operational simplicity, and meaningful value creation for society, end-user and industry alike.

06 LIST OF ABBREVIATIONS

3GPP	3rd Generation Partnership Project
A2A	Agent-to-Agent Protocol
AI	Artificial Intelligence
AF	Application Function
AI	Artificial Intelligence
API	Application Programming Interface
AR	Augmented Reality
CN	Core Network
DL	Downlink
HARQ	Hybrid Automatic Repeat Request
L2	Layer 2
LLM	Large Language Model
MAC	Medium Access Control
MCP	Model Context Protocol
ML	Machine Learning
MNO	Mobile Network Operator
NF	Network Function
NGMN	Next Generation Mobile Networks Alliance e.V.
NPU	Neural Processing Units
QoS	Quality of Service
QUIC	Quick UDP Internet Connections
OTT	Over The Top
RAN	Radio Access Network
SBA	Service-Based Architecture
SBI	Service-Based Interface
UE	User Equipment
UL	Uplink
VR	Virtual Reality
XR	Extended Reality

07 REFERENCES

[1] NGMN, 6G Position Statement, Sep 2023, https://www.ngmn.org/wp-content/uploads/NGMN_6G_Position_Statement.pdf

[2] NGMN, 6G Drivers and Vision, April 2021, https://www.ngmn.org/wp-content/uploads/NGMN-6G-Drivers-and-Vision-V1.0_final.pdf

[3] NGMN, 6G Use Cases and Analysis, February 2022, https://www.ngmn.org/wp-content/uploads/220222-NGMN-6G-Use-Cases-and-Analysis-1.pdf

[4] NGMN, 6G Requirements and Design Considerations, Feb 2023, https://www.ngmn.org/wp-content/uploads/NGMN_6G_Requirements_and_Design_Considerations.pdf

[5] NGMN, 6G Key Messages – An Operator View, Jun 2025, https://www.ngmn.org/wp-content/uploads/2506_NGMN_6G-Key-Messages_An-Operator-View_V1.0.pdf

[6] NGMN, Network Architecture Evolution towards 6G, Feb 2025, https://www.ngmn.org/wp-content/uploads/250218_Network_Architecture_Evolution_towards_6G_V1.0.pdf

[7] Recommendation ITU-R M. 2160, Framework and overall objectives of the future development of IMT for 2030 and beyond, https://www.itu.int/dms_pubrec/itu-r/rec/m/R-REC-M.2160-0-202311-I!!PDF-E.pdf

[8] AppLogic Networks 2025 Global Internet Phenomena Report, https://www.sandvine.com/hubfs/AppLogic_Networks/Collateral/Global%20Internet%20Phenomena%20Reports/GIPR%202025.pdf

[9] Tridens Mobile Data Statistics 2025: Global Usage Trends and Consumption, https://tridenstechnology.com/mobile-data-statistics/

[10] Ericsson Mobility Report: GenAI's impact on network data traffic, Jun 2025, https://telecomlead.com/5g/ericsson-mobility-report-genais-emerging-impact-on-network-data-traffic-121471#:~:text=While%20most%20mobile%20network%20traffic%20follows%20a%20-90%3A10,the%20interactive%20and%20content-generation-heavy%20nature%20of%20these%20applications.

[11] Google Gemini Nano for pixels

08 FIGURES

Figure 1

ACKNOWLEDGEMENTS

The following people and companies participated in the development of this publication.

Project Leads:	Quan Zhao (China Mobile), David Lister (Vodafone).
Editors:	Quan Zhao (China Mobile), Tao Sun (China Mobile), Jose Antonio Martin Martinez (Telefónica), Viraj Abhayawardhana (Liberty Global), Guenter Klas (Vodafone).
Contributors:	Eric Hardouin (Orange), Elena Serna Santiago (Telefónica)

1finity, Andrew Tanamas
Amdocs, Andrew D'Souza (until Dec 2025)
Anritsu, Jonathan Borrill
Anritsu, Marsh D'Souza
Apple, Mona Mustapha
Apple, Ralf Rossbach
Apple, Kome Oteri
BT, Kevin Holley
BT, Richard MacKenzie
BT, Aaron Walker
China Mobile, Tong Zhang
China Mobile, Xiaoting Huang
Chunghwa Telecom, Che-Wei Yeh
CICT Mobile, Hui Xu
CICT Mobile, Shaoli Kang
CISCO, Fran O'Brien
Deutsche Telekom, Konstantinos Chalkiotis
Deutsche Telekom, Peter Stevens
Ericsson, Janne Peisa
Ericsson, Massimo Condoluci
Ericsson, Gunnar Mildh
Ericsson, Torbjörn Cagenius
Fraunhofer Institut, Zoran Utkovski
Fraunhofer Institut, Hauke Buhr
Huawei, Peiying Zhu
Huawei, Xueli An
Huawei, Yan Chen
Huawei, Fei Li
Huawei, Xu Li
Huawei, Hui Lin
Huawei, Shaoyun Wu
Huawei, Fang Li
Huawei, Said Tatesh
Institute of Information Industry, Kay Chung
Institute of Information Industry, Yountai Lee
Institute of Information Industry, Frank Su
InterDigital, Ulises Olvera
InterDigital, Sebastian Robitzsch
ITRI, Mitch Tseng
ITRI, Charlie Chang
LG Electronics, Ki-Dong Lee
MTN, Farhan Khan
Nokia, Omar Elloumi
Nokia, Devaki Chandramouli
Nokia, Atte Länsisalmi
NTU, Hung-Yu Wei
Orange, Jean Schwoerer
Orange, Arnaud Braud
Orange, Sandrine Lataste
Orange, Dinh-Thuy Phan Huy
Orange, Steve Tsang Kwong U
SK Telecom, Hyeonsoo Lee
Smart PH, Francisco Mangaliman
Telia, Ulf Nilsson
Telia, Anders Rudolphi
TELUS, Aleksandar Prekajski
TNO, Ljupco Jorguseski
TNO, Relja Djapic
Telefónica, Alexander Chassaigne
Telefónica, Jesús Martín
Telefónica, Manuel Nuñez
University of Dresden, Mohammed Elabsi
Vodafone, Chris Pudney
Vodafone, Rishikesh Chakraborty
Vodafone, Jörg Sternagel
ZTE, Feng Xie
ZTE, Yin Gao
ZTE, Yuzhou Hu
ZTE, Jiajun Chen

NEXT GENERATION MOBILE NETWORKS ALLIANCE

NGMN - Next Generation Mobile Networks Alliance - is a global, operator-driven organisation established by leading international mobile network operators (MNOs). As a global alliance of operators, vendors, and academia, NGMN provides industry guidance to enable innovative, sustainable and affordable next-generation mobile network infrastructure.

Key focus areas include Mastering the Route to Disaggregation, Green Future Networks, and 6G, while supporting the full implementation of 5G. NGMN drives global alignment of technology standards, fosters collaboration with industry organisations and ensures efficient, project-driven processes to address the evolving demands of the telecommunications ecosystem.

VISION

The vision of NGMN is to provide impactful industry guidance to achieve innovative, sustainable and affordable mobile telecommunication services to meet the requirements of operators and address the demands and expectations of end users. Key focus areas include Mastering the Route to Disaggregation, Green Future Networks and 6G, while supporting the full implementation of 5G.

MISSION

The mission of NGMN is:

- To evaluate and drive technology evolution towards the three **Strategic Focus Topics:**
    - **Mastering to the Route to Disaggregation:**

      Leading in the development of open, disaggregated, virtualised and cloud native solutions
    - **Green Future Networks:**

      Developing sustainable and environmentally conscious solutions
    - **6G:**

      Providing guidance and key requirements for design considerations and network architecture evolution
- To define precise functional and non-functional requirements for the next generation of mobile networks
- To provide guidance to equipment developers, standardisation bodies, and collaborative partners, leading to the implementation of a cost-effective network evolution
- To serve as a platform for information exchange within the industry, addressing urgent concerns, sharing experiences, and learning from technological challenges
- To identify and eliminate obstacles hindering the successful implementation of appealing mobile services.