Perspective

Enabling the era of hyper-personalisation through agentic 6G networks

James Kirby, Adaora Okeleke, Gentiana Shiko, Janette Stewart and Caroline Gabriel

November 2025

Contents

List of figures

1

Executive summary

This white paper discusses the potential for mobile networks of the future to offer 'hyper-personalised services' (or h-p services). These h-p services might be enabled through advanced technologies such as artificial intelligence (AI) agents, and will enable mobile networks to understand user behaviours in real time, and to tailor services, products and experiences to individual preferences instantaneously.

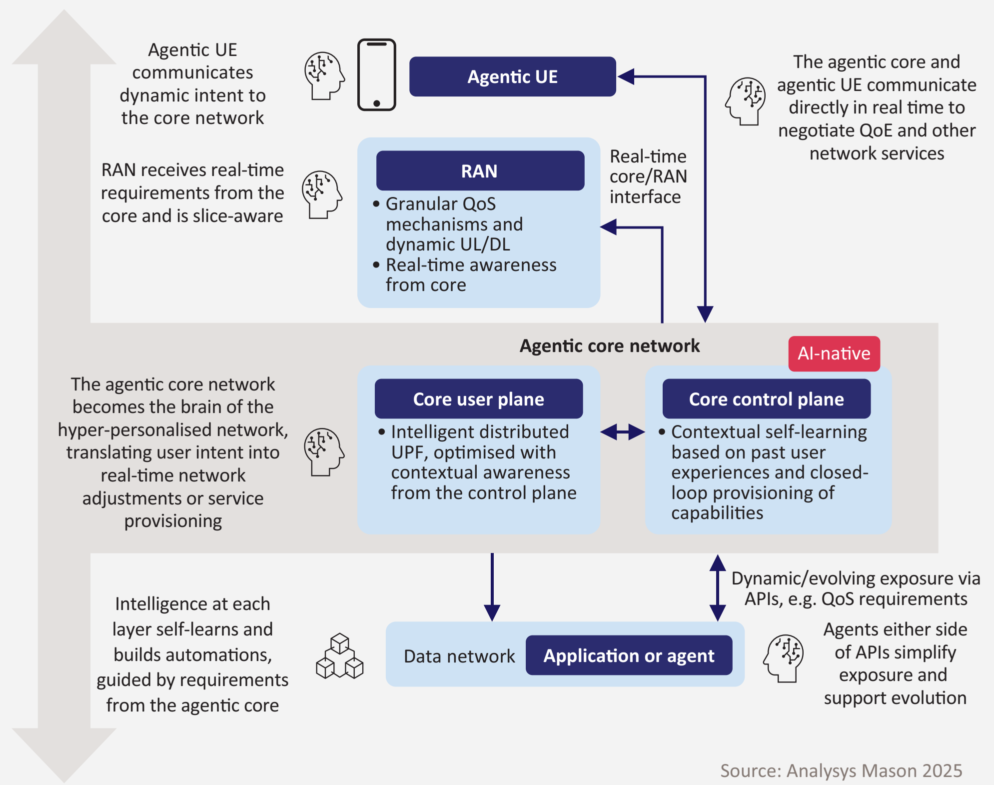
The impact of hyper-personalisation on user experience, and also on operator KPIs, could be dramatic. Imagine commuters on a busy train to work, able to enjoy an excellent and consistent quality of experience (QoE) as they work or play a mobile game, with an unbroken and unchanged experience as they move through a busy station, onto a high-speed train, and into a taxi at the journey's end. The operator that delivers this guaranteed QoE in the most challenging environments will be rewarded with customer loyalty, increased usage and, possibly, willingness to pay to subscribe to a 'QoE-anywhere' service.

This paper will illustrate how such a scenario could be enabled by mobile networks beyond 2030 (Sixth Generation or 6G), whose architectures could allow mobile networks to interact with AI agents in any device or application in any location. The capabilities that 6G could enable include the following:

- AI agents might use reasoning and self-learning to make real-time decisions on network optimisation for individual users' QoE, as well as anticipating QoE issues at user level before these arise.
- An h-p service using agentic AI would continuously test and self-improve over time, in order to enhance and adapt the user experience to the latest requirements.
- Mobile network operators (MNOs) could personalise services with high levels of precision, fine-tune all customer interactions and potentially drive improved revenue performance given increased customer retention.

In the paper, we discuss how MNOs will require open, interoperable and standardised agentic AI capability stacks if an agentic AI-enabled 6G network is to operate autonomously in response to customers' needs. A potential view of the network architecture required to deliver the full h-p experience is shown in Figure 1.1 overleaf. This type of architecture would allow AI agents on the user equipment (agentic UE) to communicate with in-built AI agents in the core network (agentic core) in order to process and learn from, and act upon, the user's intent in real time. The agentic core would play a central role in this architecture, automatically provisioning or adjusting end-to-end (E2E) network capabilities to suit user needs.

Figure 1.1: A potential 6G agentic network architecture to support hyper-personalisation



Source: Analysys Mason 2025

Integration of agentic AI in order to deliver hyper-personalisation will need to take place alongside other technology updates in the radio access network (RAN) and in the network core. In our view, evolution of the agentic core in 6G networks could unlock the full potential of network hyper-personalisation.

There are well-known technical, operational and financial challenges associated with implementing major network upgrades. Our white paper raises the following key issues:

- Development of devices supporting agentic AI is progressing at a faster rate compared to the equivalent agentic AI developments in the mobile network.
- There is a lack of connectivity capability by the mobile network which would take advantage of the agentic AI capabilities of the agentic UE (which, in turn, would support the functionalities required for delivery of h-p telecoms services to telecoms customers.)
- Working in best-effort mode, current 5G networks do not offer the real-time response needed to support the delivery of h-p experiences and services to customers in the network, especially for future real-time immersive services and AI-based burst services.
- There are uncertainty and concerns across the telecoms value chain on handling customer trust and high levels of power requirements, as well as gaps in standardisation and regulation, that need to be addressed in order to successfully implement agentic AI in the network.

However, if these challenges can be addressed, offering hyper-personalisation may benefit many MNOs, lifting network quality from a basic ‘best-effort’ service to an all-encompassing, highly tailored and adaptive connectivity offering.

The remainder of the white paper is structured as follows:

- Section 2 provides an introduction to service personalisation and uses cases that might be enabled by agentic AI technology and discusses the key challenges of achieving hyper-personalisation within current 3GPP specifications and deployments.
- Section 3 discusses how 6G networks with AI agents may enable delivery of h-p services when integrated in user devices and the wider mobile network infrastructure.
- Section 4 describes key questions and challenges to be addressed as part of standardising and commercialising an agentic AI 6G vision.

2

Hyper-personalisation, a radical change in mobile user experience, can be enabled by agentic AI

2.1 What is hyper-personalisation?

In broad terms, service and product personalisation today refers to tailored approaches to customer engagement to offer relevant products, services or content based on a user's past interactions (e.g. past purchases, online browsing history). The various forms of marketing personalisation that exist today are largely based on machine learning from customer behaviour data; examples exist across many sectors including retail, travel, hospitality, e-commerce and digital media. For customers, personalisation can improve service experience; for the service provider, the increased user satisfaction can have financial benefits (for example, increased spending, improved customer loyalty/reduced switching to alternatives).

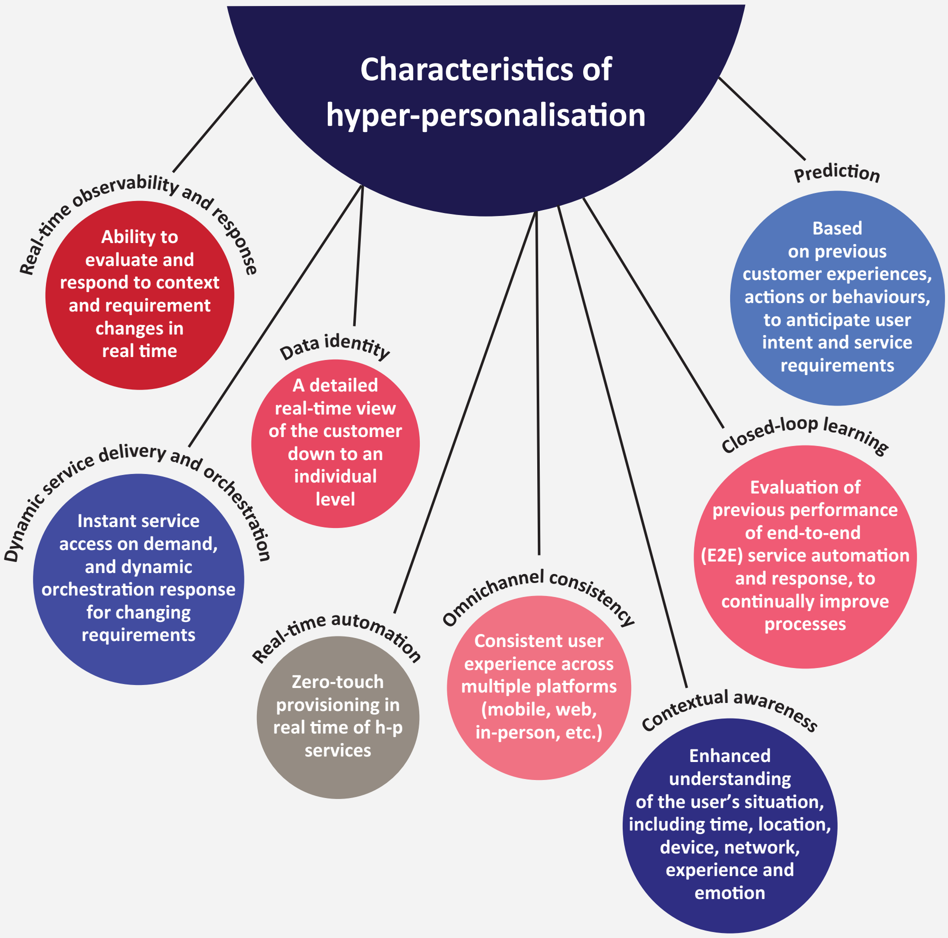
However, success has been mixed, partly due to network technology limitations which put constraints on tailoring of telecoms services by location and/or over short time intervals. In light of the increasing popularity of artificial intelligence (AI), and in particular, agentic AI,[1] there is potential for hyper-personalisation to enable mobile network operators (MNOs) to amplify and tailor connectivity capabilities to the requirements of the customer, in real time and on a per-device level. Industry interest is fuelled by consumers seeking products, services and information that are available instantly and seamlessly, and are tailored to what they want.

Hyper-personalisation refers to an enhanced level of personalised experiences, leveraging real-time data and AI. It enables the tailoring of services and prediction of user behaviour/need through learning from past interactions. In order to deliver h-p telecoms-related services, mobile networks need to evolve and include advanced AI technologies (agentic AI) throughout the network infrastructure and in the user devices. In this way, AI agents which are being built into user devices (agentic UE) can be developed to detect and predict the quality of experience (QoE) of the mobile connection, and autonomously interact with AI agents in the core network (agentic core) to dynamically and intelligently generate and adjust resources to match the user's requirements exactly.

Hyper-personalised services possess characteristics which deliver several benefits to service providers when engaging with customers. These characteristics are shown in Figure 2.1 below. However, the successful delivery of h-p telecoms-related services also brings challenges: it requires an enhanced level of transparency and trust, alongside customer control and protection and clear pricing structures, as discussed later in this paper.

[1] Agentic AI refers to AI systems or agents that can make decisions autonomously and take action based on goals, operating over long periods of time and in dynamic environments. We discuss agentic AI in more detail in Section 4 of this paper.

Figure 2.1: Characteristics of hyper-personalisation for telecoms services



Source: Analysys Mason 2025

2.2 The value of hyper-personalisation in mobile networks

MNOs and the telecoms industry ecosystem, more generally, are likely to see a number of benefits from implementing hyper-personalisation in both the services and products they offer to customers, and internally, within the myriad of routine operations and interactions between different stakeholders. For example, focusing on MNOs, implanting h-p telecoms services may result in the following benefits:

- direct benefits – e.g. increased customer loyalty and/or greater spending with the same provider, potentially leading to greater average revenue per user (ARPU) and/or improved market share for the provider; improved proposition for content, apps and advertising partners, which helps to build a strong base of services for users and potential revenue-share opportunities
- indirect benefits – e.g. improved customer feedback and customer retention, improved offering of differentiated services and onboarding of new customer segments, more granular data and insights on customer preferences to inform new commercial strategies
- internal operational value – e.g. increased operational efficiency by minimising over-provisioning of network capacity and resources, as well as operating a self-learning and self-correcting (generative) network.

There are already developments in various parts of the telecoms ecosystem that will lay the foundations for 6G agentic AI services. Several device manufacturers are already incorporating agentic AI into their products to support the concept of agentic UE with the objective to deliver deeply personalised experiences to users. Other vendors within the 'on-device AI' value chain are evolving their offerings to support the agentic UE. Examples include the following:

- Samsung is integrating agentic AI features into its devices (including smartphones, wearables and TVs) with a view to offering personalised experiences to users; these agentic AI features are performing cross-app actions and automated tasks on selected devices, on behalf of users.

- Qualcomm launched its new mobile platform, the Snapdragon 8 Elite Gen 5, which includes its Hexagon NPU (an enhanced neural processing unit (NPU)) that offers a 37% increase in AI processing power required to process more tokens at a faster rate; this capability is critical to the operations of agentic UE.

- Media Tek has been working on chipsets that support AI agents.

However, these developments focus on the device experience, and will not fulfil the full set of requirements for h-p telecoms services. In order to complete the loop needed to deliver h-p services, interactions are required between these agentic UE devices and similar agentic AI capabilities operating autonomously within the network.

2.3 Collaboration between humans, agentic UE and the agentic core will enable a myriad of h-p services

In a h-p telecoms future, networks will have to deliver personalised services to customers, but also to AI agents in devices and within the network infrastructure. In this section, we present some use cases that may become viable in an agent-driven network future. These use cases are examples of hyper-personalisation with diverse context and user intent, enabled through cross-collaboration between humans and AI agents in user devices, in the network and in humanoids,[2] and which cannot be delivered by 5G networks at present.

Figure 2.2: Examples of hyper-personalisation use cases in future networks

Use case	Operational characteristics for delivering h-p service	An example in practice
Dynamic application usage	The network adapts quality of service (QoS) or network capabilities and resources according to dynamic application usage in real time. As users navigate between different applications or in-application tasks, the network understands and compensates for their connectivity needs.	Dynamically switching between applications such as video calls, cloud-gaming, social media, live streaming or entertainment news that have varied and dynamic requirements depending on application context. In-application task changes may include different games with different requirements, or moving from text to video consumption on social media or news applications.

[2] Humanoids are robots modelled after human form factor and designed to perform tasks to augment productivity or enable actions in contexts where humans cannot.

Use case	Operational characteristics for delivering h-p service	An example in practice
Contextual user awareness	The network understands user context and the user location, device type, past usage, subscription details and current activities to accurately infer intent and network requirements to support the optimal E2E experience.	The network understands if the user is commuting, at a large event, on a long-distance journey or at work, and is able to adjust offerings or experience accordingly.
Tailored real-time service suggestions	Through contextual awareness and analysis of past user data and experiences, the network can accurately suggest relevant service improvements and additional network services in real time.	The network understands the user is commuting and offers a commuter package that provides an attractive service offering according to their specific application requirements.
Smart transit connectivity	The network understands a user's mode of transport, can predict their likely route through application and contextual awareness from the network, and can provision network resources more effectively to compensate for their movement.	The user is on the train or the highway and paying for a premium connectivity package. The network can dynamically provision a user-plane resource along the stretch of road or track during the user's travel.
User complaint handling	Users can submit simple complaints in real time and the network reacts quickly, using network, contextual and application awareness to infer more accurate user intent and adjust resources where required throughout the network.	A user complains on an agent application that a "video call is bad". The network resolves any issues or suggests priority connectivity services to improve the experience for the duration of the call.
Multi-modal agent communica-tion	The network is able to personalise experiences for agent-to-agent (A2A) communication, creating secure multi-modal communication groups that dynamically meet agent requirements for connectivity.	An onsite AI agent detects a dangerous incident involving multiple people. The on-site agent instantly communicates with the nearest hospital triage AI agent, which alerts ambulance services. These agents communicate with AI agents on board the ambulance, to describe the situation to paramedics, supporting with patient information and live imagery from the scene to prepare the crew and resources. Multi-modal (video, image, text and voice) communication among these AI agents is guaranteed and optimised for in a dynamic way. The network could instantiate cloud resources for agent interactions.
Embodied AI agents	The network meets the requirements of embodied AI agents, providing guaranteed network service and exposing network capabilities such as sensing. The network can communicate with the agent to understanding intent, solve issues and instantiate cloud application instances where required.	An industry robot experiences an object blocking its path, but onboard sensors and AI are insufficient. It communicates its issue to the network. The network instantiates cloud capabilities for AI video analytics and provides network-sensing capabilities to support identification of the issue and potential resolutions.

Source: Analysys Mason 2025

[3] For example, cloud gaming is the application, but the specific game (with unique requirements) is the task; TikTok is the application, but content upload or live streaming is the specific task

Different applications of hyper-personalisation will have different combinations of network requirements, included but not limited to specific levels of latency, throughput, jitter, uplink, downlink and security. Increasingly, as users switch between different applications or within applications,[3] h-p features can enable networks to predict and adjust capabilities in real time to provide the QoE needed for that application, wherever the user is located and over the duration that the user requires it.

Offering of personalised telecoms services at present is limited

There are challenges in bridging the gap between today's standards of mobile services and an h-p future. Network-specific requirements will include, amongst others, exposing some user data, training for application awareness, supporting high automation, managing unstructured data, and delivering closed-loop automation. Devices may provide an additional challenge in that the device capabilities will limit how many h-p services the user can purchase or use, before requiring a device upgrade or an additional device.

Recently, MNOs have started to leverage some QoS and personalisation capabilities and there is growing interest in supporting quality on demand (QoD) through network application programming interfaces (APIs). Figure 2.3 provides examples of commercial offerings by MNOs; however, at present, the level of personalisation available commercially is limited. Many of the personalised services offered through 5G networks are only semi-personalised, in practice. The current approaches support differentiated QoS for a small number of pre-defined user groups and, often, leverage a limited number of network capabilities, with relatively static and pre-defined policies.

Figure 2.3: Examples of quality on demand/personalised services using 5G

Operator	Service	Description
China Unicom	Service tiers for streamers, gamers and others	Leveraging a Network Data Analytics Function (NWDAF) and an inter-personalised experience function in the core network
China Mobile	Service tiers for UE in transit	Service guarantee on high-speed rails
AIS Thailand	Service tiers	Option to purchase QoS enhancements for specific use cases for a given time period
AT&T Turbo	Per-user boost service	Not tailored, per-user service
Deutsche Telekom	Improve QoS for gamers of specific platform	Using network slicing in partnership with gaming platform
T-Mobile T-Priority	Priority/improved service for emergency services	Priority access and pre-emption through 5G network slicing for public safety

Source: Analysys Mason 2025

2.4 Current 5G architecture and capabilities make it challenging to deliver h-p services

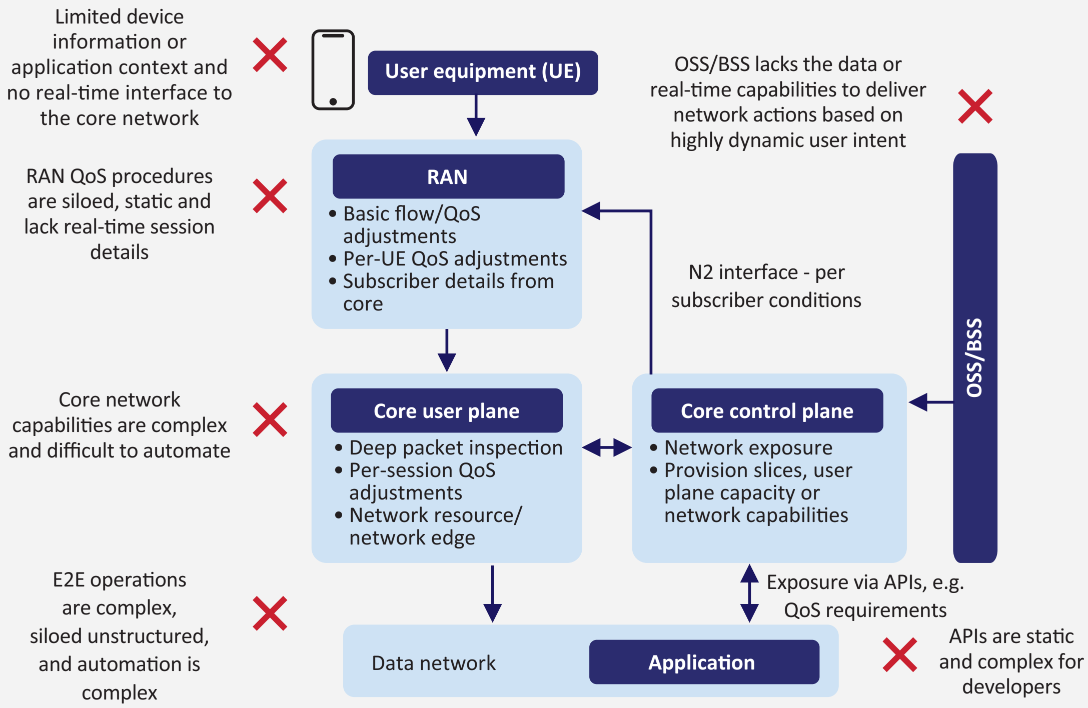
Personalisation in 5G networks has been supported by the introduction or enhancement of some fundamental capabilities. These include network slicing, enhanced location/positioning data, enhanced radio intelligence and massive MIMO, digital twins, open APIs to expose network functions to developers, and intent-based automation. Furthermore, 5G has provided a foundational blueprint for supporting personalised services for users or applications within the network; the approach to delivering personalised services in 5G networks and its major challenges are represented in Figure 2.4.

There are four major personalisation flow concepts in the 5G context:

- only the core network is used to provision network slices
- radio access network (RAN) QoS improvements are used to deliver differentiated service tiers (static QoS profiles from the 5G core)
- user requests flow through operations support systems (OSS)
- a combination of the approaches above, whereby the control plane communicates directly with the user plane functions (UPFs) and the RAN in order to optimise performance on a per-session basis across RAN and core

These four concepts of personalisation flow in 5G networks have remained nascent in commercial terms, due to technical and operational challenges. At present, some 5G networks use either the RAN or the core method combined with OSS/BSS information. However, implementation of the RAN and core methods as a fully integrated solution is rare in practice, as the real-time exchange of session-level parameters between the core and RAN remains an immature concept and orchestrating QoS parameters interdependently across these domains is complex. Instead, the RAN and the core are, most often, managed by separate rules driven by an orchestration layer. This makes it challenging to personalise the user experience from end to end and limits the potential for dynamic response. Furthermore, capabilities such as accurate positioning and network edge are limited, and require robust automation before being able to support personalisation.

Figure 2.4: Personalisation flow and challenges for advanced 5G networks



Source: Analysys Mason 2025

As shown in Figure 2.4 above, MNOs face a number of key challenges in delivering high levels of personalisation in 5G:

- **Poor availability of data from UE devices and applications** – session-level QoS needs rely on packet inspection, which gives an incomplete view of user or app-specific requirements. Deep packet inspection offers more detail but is resource-heavy and slow to enforce. Privacy laws and encryption further restrict data collection from the UE – these are constraints that will persist or grow in future 6G networks (as discussed in Section 4).

- **Challenges in the 5G network API model** – using the latest 3GPP standards, data exposure via network APIs can allow applications to request QoS and other services. However, this concept faces hurdles such as complexity for developers, fragmented and static APIs, weak integration with automation and analytics, poor capability discovery, limited scalability and limited abstraction layers.

- **Insufficient network or operational capabilities** – MNOs have lacked the network intelligence and automation capabilities to understand the available data on a large scale, and to implement relevant network or service changes in an efficient (and reliable) way.

- **Inconsistent data and interworking across domains** – RAN, core, transport and OSS/BSS produce inconsistent and siloed data. QoS is provisioned with different mechanisms across domains, and the current mobile network architecture struggles to implement E2E QoS parameters based on real-time data in the network.

- **Technical and operational challenges with network edge** – network edge services have demonstrated challenges with roaming and mobility, have been challenging to operationalise, and must be highly automated to become a viable service offering.

- **Lack of real-time feedback from UEs** – customer or user data is not detailed nor received by the network in real time; hence, the network is not able to establish intent easily from user requests.

- **Limited support for other personalised features** – examples include the provision of network sensing/location capabilities when required, or the deployment of computing resources and services for applications.

In the remainder of this white paper, we explore how evolution to 6G networks and use of agentic AI may help to address some of the challenges observed in 5G networks. In our view, the core network will be the centre of hyper-personalisation, as we discuss in the next section.

6G will enable mobile hyper-personalisation using agentic AI to reach its full potential

Supporting hyper-personalisation could bring value to the telecoms industry and its customers. However, to do so effectively, 6G must build on the capabilities of 5G through addition of real-time generative hyper-personalisation. This means solving the technical challenges remaining with key personalisation capabilities such as slicing, QoS tiers and network edge.[4] It also means achieving advanced levels of data integration and automation across network domains, facilitating dynamic, robust and adaptive E2E QoS and capability management based on detailed user and session information.

It is yet to be established how 6G will conclusively achieve all these capabilities; official 3GPP standards discussions have only just begun for the next mobile generation. However, it is clear that agentic AI will play a pivotal role in achieving the level of automation and contextual awareness required for hyper-personalisation. Agentic AI could be integrated at multiple layers of the mobile network to help support the network architecture required for hyper-personalisation, including integration of AI agents throughout the network and device layers.

3.1 Implementing a 6G agentic AI network architecture delivers many benefits

An agentic network and device architecture in 6G could deliver a range of benefits in terms of user experience and potential monetisation, addressing the challenges currently faced in 5G networks and delivering a high level of personalised service. Figure 3.1 demonstrates how an agentic architecture could help to solve the challenges currently facing hyper-personalisation in mobile networks.

Figure 3.1: Addressing 5G personalisation challenges with the 6G agentic framework

5G challenge	Value of the 6G agentic framework
Poor availability of data from UE devices and applications	• The agentic connection between the UE and the core network improves awareness on both sides of the air interface. This is not limited to application intent and contextual information, but also includes device and inter-application usage intent. • The agentic UE can send intent and contextual information to the network, based on detailed information held on the device. The network would otherwise struggle to benefit from this data due to security and privacy concerns or regulations.
Challenges in the 5G network API model	• The agentic UEcould take the role of communicating intent and application requirements to the network in a range of scenarios. This represents a more detailed and dynamic communication of user intent to the network • For other use cases, agents on both sides of network APIs could automate network exposure for operators and developers, simplifying the process of network capability exposure. These agents could also seek to evolve data requirements via supervision and validation.

[4] Network edge could provide the processing infrastructure for user plane distribution and the running of applications or agentic AI features within the network.

5G challenge	Value of the 6G agentic framework
Insufficient network, data or operational capabilities	• AI agents within the core network would be able to act based on a wide range of data insights, helping operators to build closed-loop provisioning for a range of network capabilities – such as E2E slicing, dynamic policy rule, bearer mapping, location/positioning, and 6G sensing. Data sources could include agentic UE intent, application intent and E2E network QoS.
Inconsistent data and interworking across domains	• By making the agentic core network the centre of the agentic framework, this becomes a single source of truth for other domains. It would act as a logical starting point to build automations into other domains. However, standards and operators themselves will still need to prioritise interworking and improve data consistency between domains. • Agents could become the new orchestrators of the network and could support the normalisation of network data and communication on an E2E basis.
Technical and operational challenges with network edge	• Network edge faces fundamental challenges relating to roaming and location determination that must be addressed. • However, the operational challenges of network edge (such as real-time provisioning of cloud or UPF resources to where they are needed) could be supported through closed-loop automation in the agentic core network. Improved contextual awareness within the core network could also make these capabilities more dynamic and attractive.
Lack of real-time feedback from UEs	• In the same way the agentic UE would communicate intent to the network, the network would communicate network information to the device. This would allow both sides to be context and network aware, dynamically negotiating closed-loop adjustments to the user experience. It would also allow the network and device to understand QoS issues in greater detail.
Limited support for other personalised features	• In addition to network edge and slicing, the network could play an enhanced role in 6G. Capabilities such as sensing or co-hosting enterprise workloads on the RAN or core platform could offer additional opportunities for operators. However, the success of such services will require the network to be context aware and deliver these capabilities on demand in real time. The agentic framework could support both network awareness and the operational capabilities to support these service functionalities.

Source: Analysys Mason 2025

Beyond addressing the challenges of hyper-personalisation in 5G networks, a 6G agentic framework may enable networks to respond to highly frequent changes in application and user context (e.g. meeting different network demands as users switch between applications or physical environments, in near real time), or to address network QoS issues or requirements before a user complains or enquires. The latter may be especially valuable to MNOs in order to retain customers, lower churn and offer new services dynamically where there is a requirement. In Section 2.3, we discussed some examples of possible future user environments in mobile network with agentic AI enabling responsive, predictive and real-time optimisation of network resources (see Figure 2.2 above).

3.2 Technology enablers for realising the 6G agentic AI network architecture

Potential multi-layer agentic network architecture for delivering 6G hyper-personalisation

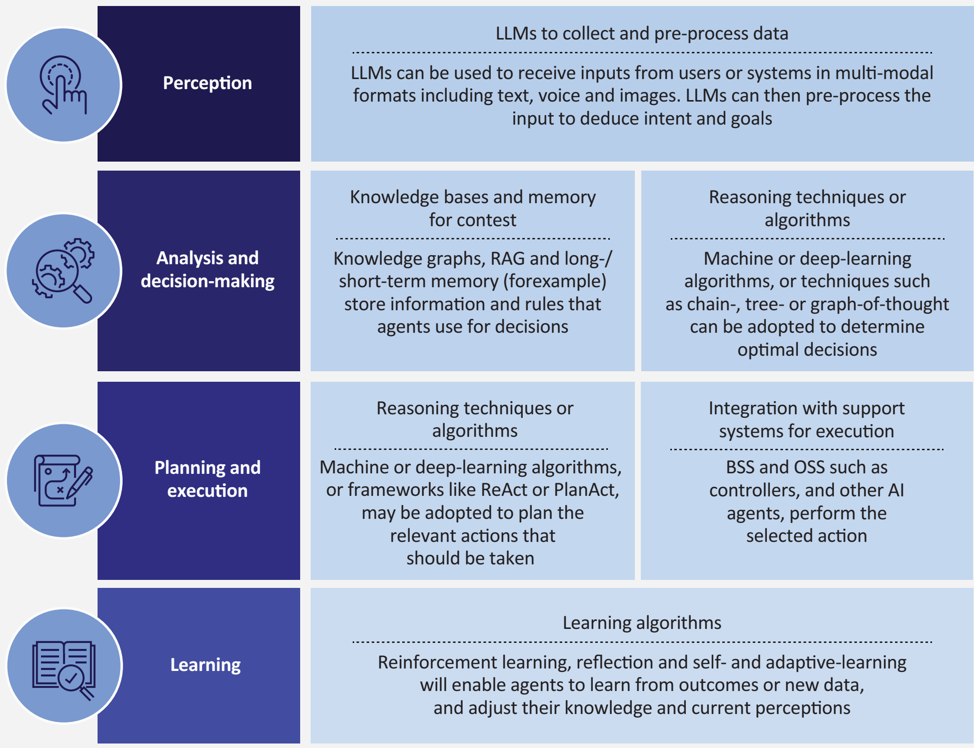
Agentic AI

Agentic AI is an advanced form of artificial intelligence that is focused on developing systems that can make decisions and take actions autonomously. Agentic AI capabilities transcend the creative functions of generative AI (GenAI) or the specialised functions of other AI capabilities (such as prescriptive or predictive AI), because they directly integrate into organisation's workflows and do not require human oversight.

Figure 3.2 summarises some of the characteristics of AI agents in relation to mobile networks.

Agentic AI systems can potentially perceive user intent, make decisions on actions to take to fulfil user intent, and/or act autonomously to respond dynamically to customer needs. When performing these functions, these systems adopt workflows that decompose complex goals, iteratively optimise actions and actively adapt to dynamic environments. Hence, agentic AI systems hold promise in supporting the operational complexity of hyper-personalisation in mobile networks.

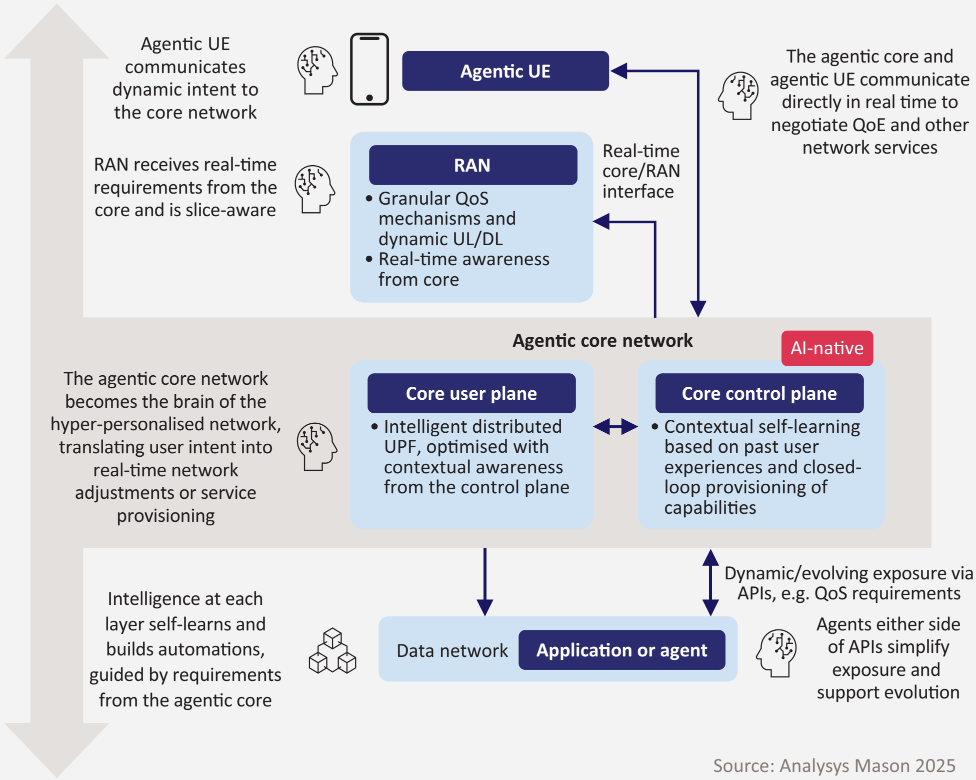
Figure 3.2: Characteristics of AI agents and related technology enablers



Source: Analysys Mason 2025

In 6G, agentic AI could be applied at multiple layers of the network as well as within the UE to support hyper-personalisation. While the depth of AI agent integration for 6G is not yet established, Figure 3.3 demonstrates what an agentic network architecture could look like in 6G. In this architecture, the agentic core network could play a valuable central role in in achieving hyper-personalisation.

Figure 3.3: Potential multi-layer agentic architecture for 6G hyper-personalisation



Source: Analysys Mason 2025

Agentic UE

Agentic user equipment (UE) is a concept currently being explored for 6G. It relates to the use of a built-in AI agent (or a system of multiple AI agents) embedded within a device in order to perform specific tasks autonomously. This device could be used by humans (e.g. mobile handsets) or they could be devices used in industry or within organisations. Examples of such devices include cameras, drones, embodied AI agents, humanoids, industrial machines and autonomous vehicles.

In an agentic AI network, in order to dynamically and automatically improve the experience in real time, the AI agents need to detect the intent, context and QoE needs of each end user, within each application's context and experience. The agents also need to know how to interact with the application data and/or the network infrastructure that links to the UE. A version of UE with agentic AI services could support this requirement, if the devices are able to communicate intent effectively to the network. In order to perform their agentic AI functions, the agentic UE would need to support AI models that use highly personalised data to reveal user context.

The agentic UE could perform the following functions:

- **Perceive** – detect changes in users' context by collecting and analysing data from several sources in real time, including active applications and the environment, so as to reveal users' context and experience.
- **Reason** – use an internal AI model so as to determine the goal (or intent) to achieve given current context and to identify actions to take to achieve the goal.
- **Plan** – break down the goal (or intent) into smaller and actionable tasks and allocates these tasks.

- **Act** – once the agentic UE decides that the network needs to take an action, it generates an intent[5] and relevant contextual information to communicate the users' requirements in real time to the mobile network.

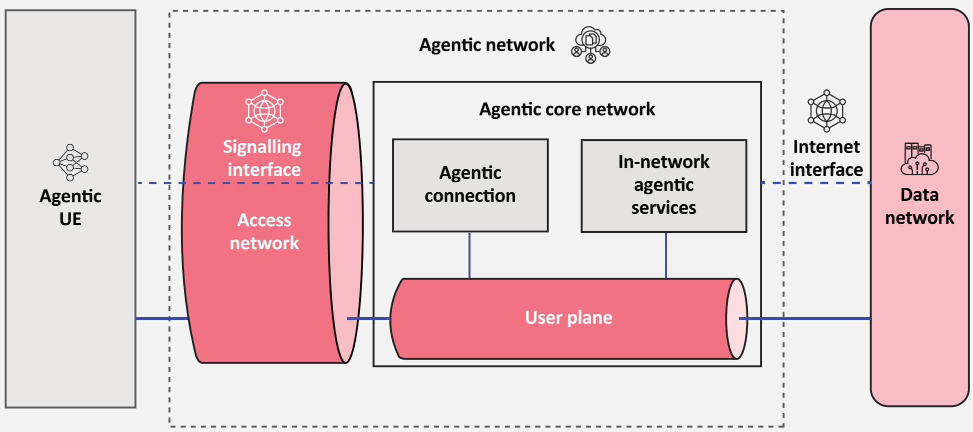
- **Evaluate** – assess the outcome of results fed back by the network in order to verify whether the user requirements were met (or not) and in order to learn from actions and adapt behaviour for future scenarios.

In order to most effectively support hyper-personalisation, the agentic UE and the network would need to allow for a two-way closed-loop communication between intent and outcome (i.e. where an intent is sent and the outcome is assessed to ensure that the outcome aligns with the intent). This communication could occur using interfaces that enable agentic communications between the agentic UE and the agentic core network. AI agents on both sides of these interfaces can process unstructured or semi-structured content. The agentic core can then adjust resources or capabilities from the user plane and RAN according to data from the UE side and the network. This process becomes a two-way negotiation, which adapts to changing network conditions and user intent. The agentic UE is unlikely to interact directly with the agentic network through the RAN, due to security and privacy concerns. Hence, our view is that the agentic core network lies at the centre of 6G capabilities.

Agentic core network

An agentic core network holds promise as the central brain for a 6G agentic network architecture. This is because core networks naturally contain the most relevant subscriber, application and network data. The core network is also the central command centre for personalisation capabilities such as network slicing, policy rule, bearer mapping and accurate positioning/location data. As shown in Figure 3.4, in order to understand and process UE intent in real time, the agentic core could expose interfaces to both the agentic UE and the data network. Existing web- and cloud-based services in the data network can engage with the agentic core through the internet. The agentic UE could then interact with the agentic core via the shortest signalling path to support real-time response.

Figure 3.4: Overview of an agentic AI network architecture (illustrative)



Source: Analysys Mason 2025

[5] The intent could be in an unstructured or semi-structured format; it could not be in a structured format due to the varied application characteristics and scenarios.

The agentic core could host network agents with different expertise. These network agents would focus on executing specific tasks linked to the customer intent provided by the agentic UE, including network resource optimisation for real-time provisioning of network resources on demand. Core network agents could also support the provisioning of capabilities such as network sensing or edge processing (in-network agentic services) in response to a specific application or UE intent.

The network agents could leverage a large-scale AI network model in order to make decisions in their subject domains, forming multi-agent systems to work in collaboration via agent-based interfaces. These agents could drive closed-loop processes that deal with device and network data,[6] simplifying and improving the accuracy of delivering network capabilities to meet user needs. These actions would also be predictive, using past personalisation data and user actions to anticipate user requirements and recommend services to the user through the agentic UE. By learning from the behaviour and experience patterns at the level of the individual UE (across applications, services and traffic flows), the network could further optimise scheduling of resources and workflows in order to enhance the hyper-personalised experience for the UE.

Agents within the 6G core would benefit from a single control plane function, referred to in Figure 3.4 as the 'agentic connection', which would be critical to understand user intent and convey instructions to the rest of the network. This component would be embedded with capabilities such as GenAI to understand user intent, and would communicate with other functions within the core to provision required capabilities. It could also benefit from having direct connection with other domains or domain agents to support real-time management of capabilities end to end across the network. Such an agentic connection could also deliver rule-based automation for users with basic connectivity.

The agentic core could also communicate to OSS/BSS layers for billing and wider network context, however for real-time adaptation, intent should be processed in the core network. OSS/BSS has neither the user data nor real-time capabilities required to achieve dynamic hyper-personalisation.

Finally, for interactions that occur between the agentic UE and the agentic core, the in-network agentic services would include an internal agentic orchestrator that has the information needed to co-ordinate these workflows between both environments.

The AI stack for implementing agentic AI network architecture in mobile networks

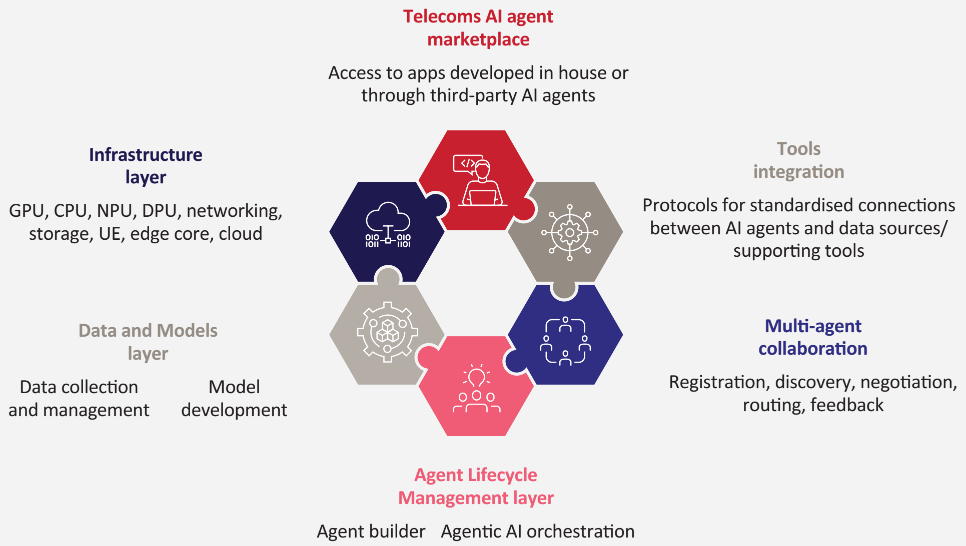
We believe MNOs will require an open, interoperable and standardised agentic AI stack of capabilities that enable the network to operate autonomously in response to customers' needs. A potential view of this stack, as shown in Figure 3.5, includes multiple layers of capabilities starting with the infrastructure layer, and progressing through to the data and models layer, all the way up to the agentic AI marketplace layer.

Infrastructure layer

This foundational layer includes cloud compute, networking, and storage tools for developing, deploying and managing agentic AI networks. Compute options such as central processing unit (CPU), graphics processing unit (GPU), and tensor processing unit (TPU) [7] should be selected based on the scale and type of AI workload. Cloud-based execution environments like Docker and Kubernetes, and runtime environments will provide sandboxes in which agents can run. To meet low-latency, high-performance scenarios, especially at the edge, network and compute environments may need to converge, and operators could invest in supporting edge infrastructure directly or via partners.

[6] The user intent could be in an unstructured or semi-structured format, but it is unlikely to be uniform due to the varied scenarios and application characteristics.

Figure 3.5: Summary of the AI stack for implementing agentic AI network architecture in mobile networks



Source: Analysys Mason 2025

Data and models layer

Data collection and management: Data is essential for training, monitoring and inference. Data collection and analysis will occur within both agentic UE and networks, using environment-specific datasets.

- **Telecom-specific AI models:** Includes generative AI models (e.g. large language models or LLMs) and non-generative AI models (predictive/prescriptive). These models must be telecoms specific and fine-tuned to reduce compute overhead. Model sharing among operators could accelerate development.

- **Ontologies:** AI agents could use domain-specific ontologies to structure knowledge, linking agentic UE intent, service features, customer preferences and network actions.

Lifecycle management of AI models and agents

Lifecycle management spans training, inference, monitoring and deployment. Dynamic orchestration ensures agents and models adapt to environmental and user needs within guardrails.

- **Agent orchestration:** An orchestration engine within the core network assigns tasks, synchronises sessions and manages agent priorities. A lead agent interprets user intent from the UE and delegates subtasks to other agents.

Multi-agent collaboration

Agents with specialised roles could collaborate via shared memory or intermediate outputs, using protocols to exchange state, assign tasks and co-ordinate actions. This framework supports task decomposition, resource distribution, conflict resolution, and planning. The process of defining multi-agent collaboration mechanisms is crucial for complex, distributed networks with privacy constraints.

Tools integration

AI agents would require access to tools operating within the network in order to initiate and execute tasks autonomously and deliver on user intent. This may entail directly leveraging network functions in the core control plane, accessing elements of OSS/BSS systems or even

[7] A tensor processing unit is an AI accelerator application-specific integrated circuit (ASIC) which is used in machine learning.

communicating with tools in the external data network. How these agents will interact with tools has not yet been standardised. However, in the IT industry, the Model Context Protocol (MCP) is being considered and adopted by several OSS/BSS players.

Telecoms AI agent network marketplace

The telecoms AI agent network marketplace will provide a central hub where MNOs and developers can buy, sell and deploy network AI agents; such an approach would mimic current app stores but, instead, focus on future agentic AI capabilities. This marketplace will be key to streamlining the implementation of the agentic AI network, since pre-built and ready-to-use agents will reduce implementation cost, time and required expertise.

3.3 Design principles for implementing an agentic AI network architecture in 6G

The implementation of the agentic AI network architecture will demand the standardisation of key requirements such as providing support for all device types and applications, network resource optimisation, and ability to respond in real time to deduced user intent. Requirements will include:

- **Support any terminal and any application –** for a true h-p experience, 6G network architecture should support any device and any application running on the device and/or embodied AI agent, ranging from smartphones, AR/XR devices, machines, drones and other UEs.
- **Operate on-demand and dynamically –** hyper-personalisation could be produced by UEs on demand, or proactively triggered by the agentic AI 6G network; for the same UE, it may dynamically change over time and in based on context.
- **Real-time network adjustments and workflow orchestration –** network resource adjustments and service workflow orchestration for hyper-personalisation will be generated in real time.
- **Run as a closed loop –** a closed-loop, collaborative interaction between agentic UEs and networks ensures AI-generated h-p implementation is verifiable; to be truly dynamic, either users, applications or UE's AI agents might be involved in verifying whether hyper-personalisation meets expectations.
- **Optimise network resource value –** since network resources, including in the RAN, are scarce, agentic-AI 6G networks will need to dynamically adjust resources to meet priorities of both the operators and the users, creating challenges to be addressed at the point of 6G commercialisation. For operators, the revenue potential from investing in network capabilities which enable dynamic assignment of resources for h-p services will be important. By comparison, users will be interested primarily in network connection satisfaction and the value of the applications being consumed.
- **Fulfil basic connectivity needs –** related to the above point on challenges in commercialising 6G, the reliance that societies and businesses place on mobile connectivity means that 6G, as with previous generations, needs to balance between providing advanced capabilities when users demand h-p experience, compared to meeting more basic requirements for any type of network connection, ensuring satisfaction for all.
- **Unified data framework –** unified framework covering data reporting from various sources, data sharing and exposure, and data lifecycle management; the unified data framework will be essential in establishing the foundation for hyper-personalisation through agentic AI.
- **Go beyond network connections –** consideration will be needed on integrated sensing, computing and AI agent communications services that may reside within the 6G network and can support agentic AI with real-time situational awareness.

Lack of considerations of these factors could slow the pace at which the industry meets growing demand for h-p services and experiences from customers. However, MNOs will face several challenges when trying to fulfil these requirements. For example, a lack of non-standardised interfaces within the agentic network could limit the realisation of the closed-loop operations required to assure h-p experiences.

Key questions and concerns must be addressed for 6G to support agentic AI optimally

The evolution of AI technologies from providing insights and generating content to enabling autonomous operations comes with benefits, but also with challenges that need to be addressed. Questions remain on critical aspects such as:

- How would AI agents on either side of the air interface (the agentic UE and the agentic network) communicate with each other?
- How do we assure and validate decisions made by agents?
- How do we secure the network from unexpected or malicious attacks?

There are several essential but challenging developments that are required to address these questions, such as ensuring uninterrupted communication between devices and the network, mitigating the risk of AI model hallucinations and security vulnerabilities, and managing their unpredictability. Regulatory and ethical considerations are also very important; the specifics of data collection need to be transparent to users, and user consent must be revocable.

Concerns around bias stemming from the data used to train AI systems are also widely recognised, and, hence, adoption of strong ethical frameworks is required.

4.1 Standardisation gaps in the agentic network architecture need to be resolved

There are significant gaps in standardisation of the agentic network architecture. We discuss the shortcomings in agentic network interfaces and orchestration in turn below.

Standardisation gaps in agentic network interfaces

Existing interfaces and APIs that are used to exchange information between telecoms networks and devices are not well suited for agentic network architecture. Figure 4.1 provides a summary of the requirements that these agentic interfaces need to satisfy. In most cases, APIs are incapable of fulfilling these requirements at present, including enabling capability discovery, context-aware messaging features and the low-latency forwarding capabilities needed to co-ordinate agents' tasks.

Figure 4.1: Requirements for network interfaces when supporting agentic interactions

Requirement	Explanation	How this can be achieved
Trust, governance and failure management	This ensures that the interfaces enable communications only between trusted agents and that the data exchanged is protected. It should also make provision	Interfaces should support authentication, authorisation, policy publication and auditing/logging of actions. They should also use failure management systems that
Semantic meaning	This provides a framework of the data, tools and systems supported by the interface to enable agent interactions	Interfaces should include ontologies or schemas that define the data and actions of the tools or systems they support; metadata offering this information should be descriptive
Discoverability	This enables agents to understand the capabilities of other agents and tools on their own	Interfaces should describe their tools and services to potentially enable semantic search and capability querying
Composability	Interfaces that can link multiple tools or services can support AI agents to perform complex tasks	Interfaces should clearly model inputs, outputs and dependencies and provide declarative descriptions of preconditions and effects
Contextual awareness	This enables AI agents to discover context linked to another agent's input and the tools and systems that an interface supports	Interfaces should expose contextual parameters and provide the ability for agents to query or infer the state of connections
Memory	This ensures that the context of conversation or interactions between AI agents is maintained	Network interfaces should be stateful interfaces that can hold and share session-state information
Support for error handling and conflict resolution	This makes provision for agents to recover from failures	Interfaces should use failure management systems that detect, report and respond to failures in agent communication

Source: Analysys Mason 2025

While protocols such as MCP[8] and Agent2Agent (A2A)[9] are already being explored, they mainly address IT-specific scenarios, while missing out on some communication-specific requirements, including:

- **high-reliability requirements –** supporting error detection, congestion control, retransmission mechanisms and redundant paths

- **security vulnerabilities –** addressing cross-domain threats (e.g. prompt injection, unauthorised access) that may occur between the agentic UE and the agentic network, or between components of the agentic network.

There is no consensus yet within 3GPP regarding the introduction of protocols such as A2A and MCP. MNOs may also need to consider customising these protocols, for example, enabling the protocols to include mobile identity within the payload of interfaces that enable communications within the agentic network. Standards organisations such as the Internet Engineering Task Force (IETF) are considering creating a dedicated working group that would address some of the issues related to the A2A protocol. Such work may provide good foundations for future 3GPP work on the agentic AI network. In addition, in h-p scenarios which require real-time adjustment in the RAN, MNOs may need to assess current signalling interactions between the agentic

[8] The Model Context Protocol (MCP) provides a secure and standardised 'language' for LLMs to communicate with external data, applications and services.

[9] The Agent2Agent (A2A) protocol is a communication protocol for communication and collaboration between AI agents.

core network and the RAN in the control plane or user plane so as to ensure it can fulfil this requirement.

Standardisation gaps in agentic orchestration

Current orchestration capabilities within the mobile network are based on static rules and are unable to manage dynamic network behaviours. In order to implement the agentic network, MNOs will need to explore solutions with features such as orchestration pattern recognition and selection, task execution sequencing, and tool/agent allocation. These features will be essential in evolving current static orchestration capabilities to more adaptive and dynamic operating systems.

4.2 Lack of trust in agentic AI must be addressed urgently

Trust is a key issue that operators must address if they are to successfully implement agentic AI network architecture. Multiple studies, including a recent report by Capgemini,[10] show that enterprises have low trust in agentic AI, with 60% of the organisations not fully trusting AI agents to manage tasks and processes autonomously.

Key contributors to mistrust in agentic AI systems include model hallucination, model accuracy and model vulnerability to malicious attacks. This perspective is supported by a recent study by Analysys Mason,[11] which found that 80% of telecoms operators face challenges in achieving the desired accuracy from GenAI models. Such inaccuracies and unpredictable agent behaviours may lead to unintended actions within networks, potentially causing adverse network effects.

Components that need to be monitored to avoid inaccuracies in agent behaviour include the LLM, the AI agents and the closed-loop system between the network and the agentic UE.

There are several interventions for MNOs to consider, including the use of techniques such as chain of thought (CoT), constrained decoding or retrieval augmented generation (RAG) to ensure syntax and semantic correctness of the LLM outputs. For inaccuracies associated with the AI agent operations within the network, the concept of the runtime assurance architecture (RTA)[12] could be adopted. The RTA system design uses a trusted monitor to oversee a more complex or untrusted controller, such as an AI model. If the complex controller's behaviour is close to being violated, the monitor can pre-emptively transfer control to a simpler, pre-certified back-up controller to ensure continued safety and system reliability. Once the trusted monitor confirms the untrusted system has recovered, it signals a switch to restore normal operations. This set-up enables high-performance systems in safety-critical settings by ensuring a reliable fallback.

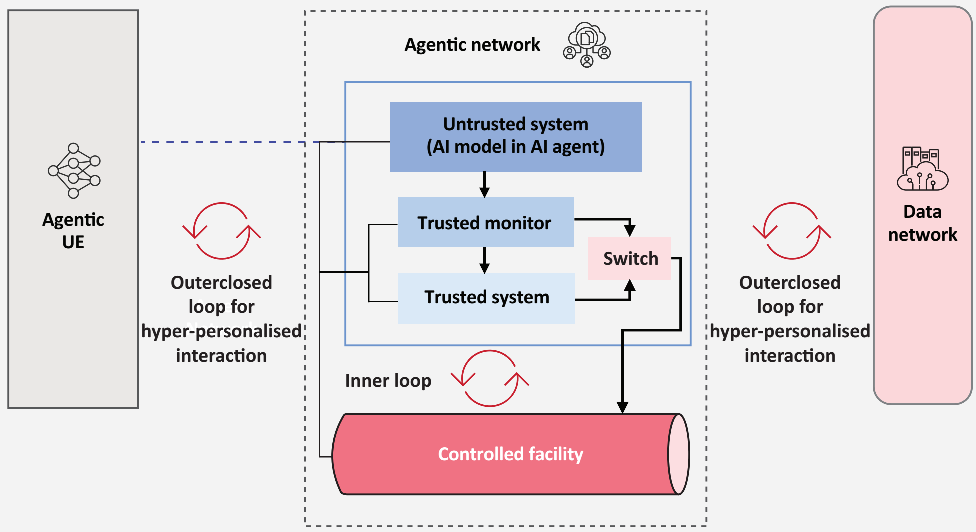
The RTA can be applied to the agentic network to achieve a trustworthy system, as illustrated in Figure 4.2. The AI model or models running within the agentic network should be treated as untrusted environments. This is because the AI model would be susceptible to hallucination, which could yield harmful outcomes for the network. Consequently, every action triggered by the result of the AI model's inference must be assessed by a trusted monitor. In this scenario the monitor, the system and the switch will all be rules-based in order to ensure a more deterministic system takes back control from the untrusted system. An inner loop (within the agentic network) and an outer loop (between the agentic network and the agentic UE or data network) will operate as closed loops (i.e. independently from the rest of the agentic network) in order to provide feedback for post-validation of actions performed by the agentic network in response to requests for hyper-personalisation. Finally, the controlled facility will include network functions such as the UPF.

[10] "Rise of Agentic AI", Capgemini, July 2025. See here https://www.capgemini.com/wp-content/uploads/2025/07/Final-Web-Version-Report-AI-Agents.pdf

[11] "GenAI in the network: CSP progress in adopting GenAI for network operations", Analysys Mason, May 2025. See here https://www.analysysmason.com/research/content/perspectives/genai-network-google-rma16-rma14/

[12] Runtime assurance architecture was first proposed by Joseph Sifakis and David Harel, in their paper titled "Trustworthy Autonomous System Development" published in April 2023.

Figure 4.2: Application of the RTA architecture to the agentic network



Source: Analysys Mason 2025

4.3 Other questions also merit attention before agentic AI can become mainstream

Uncertainties regarding regulatory requirements

AI systems have vast potential to benefit multiple sectors of society. Frameworks such as the European Union (EU) AI Act are being developed to foster safe and responsible AI development and deployment, given the significant risks around data transparency, consents and ethics. Key risks include visibility, accessibility and suitability of user rights, online safety and security concerns, as well as wider systemic risks of data models used to train agentic AI systems (for example, there is a risk of systems creating harmful biases that prevail across many applications).

Frameworks such as the EU AI Act apply to those developing AI systems, as well as those deploying them. In the context of agentic AI in a 6G network, the ecosystem of parties involved is complex. As highlighted in this paper, there may be multiple application providers, device vendors, MNOs and third-party solution providers involved in enabling the agentic AI 6G network to operate autonomously in response to customers' needs. Under the definitions in the EU AI Act, it seems likely that the types of use cases that can be enabled by agentic AI and 6G networks will fall under a high-risk category. Questions therefore remain unanswered about the specific requirements and regulation that agentic AI 6G devices and networks will be subject to.

High power consumption

There may be concerns that large-scale deployments of agentic AI will consume vast amounts of energy. Software and model efficiency efforts will be key, and use of clean energy sources will be important. Recent studies and claims from ecosystem players such as Google Cloud[13] and NVIDIA[14] indicate that intensive efforts are being made to address this concern.

Despite current efforts, MNOs must find ways to improve the energy consumption levels of agentic AI systems. They must consider factors such as the hardware used in implementing the agentic AI systems, the algorithms and optimisation techniques applied to improve power efficiency, and the amount of data required for agentic AI tasks. Some of the practices that MNOs could consider include: using energy-efficient hardware, adopting smarter model architectures, using more optimised training and inferencing techniques, and dynamically shifting models to run on less energy-consuming chipsets at peak times.

[13] https://services.google.com/fh/files/misc/measuring_the_environmental_impact_of_delivering_ai_at_google_scale.pdf

[14] https://resources.nvidia.com/en-us-sustainable-computing/accelerated-ai-energy-efficiency

Operators should also look at managing the prompt requests for each connection. A recent report published by the GSMA[15] found that energy typically accounts for ~20% of an MNO's operational costs and that energy costs are the only costs expected to continue to increase, in part driven by increasing consumption demand. The GSMA has also started tracking average energy efficiencies of mobile core networks and data centres.[16] However, MNOs are using AI-powered network planning and engineering in order to improve energy efficiencies within their networks. As AI agents would be part of a network control plane, it will be very important to use such AI-powered tools to manage the number of prompt requests per connection and consider how to best serve the prompts. For example, one way to improve efficiency could be for large portions of processing to be handled by rule-based tools and network functions, and agentic AI to be used only in cases where hyper-personalisation is part of a user's contracted service with the MNO.

Limited access to telecoms-specific LLMs

The industry faces challenges in creating telecoms-specific LLMs, which will be instrumental in driving interactions between the components of an agentic AI network. In a recent report, the GSMA[17] raises the following as key factors limiting the development of these industry-specific models:

- **Fragmented data environments –** MNOs have been unable to share data to create a sufficient corpus of data to pre-train and then refine the models for telecoms-specific scenarios. Such consolidated datasets would drive an accelerated improvement in the quality of the outputs generated by the models.
- **Inadequate model customisation techniques –** customisation techniques that would be relevant to the telecoms environment are currently limited. Consequently, the cost to create these telecoms-specific models is likely to be high.
- **Limited access to evaluation frameworks –** telecoms-specific evaluation benchmarks would be helpful in testing and validating the performance of agentic AI, including its ability to deduce user intent effectively, and evaluating the reasoning, decision-making and planning capabilities of the agentic AI network.

Development of standardised data modelling

A standardised framework for data modelling will be key for stakeholders across the ecosystem (i.e. across the network, device and application data levels). Any delays in developing and finalising a standardisation model will likely slow down the realisation of the agentic network.

Data modelling is essentially a mapping of data assets generated from different environments, which enables the contextualisation of the data the AI agent needs in order to understand its environment and make informed decisions. A standardised data model is especially important since data which the agentic AI network will leverage to execute tasks are locked away in silos. MNOs should evaluate how device data and various components of the agentic network can be aligned with a unified and standardised data model, with a view to support the transition towards agentic networks and enable advances in hyper-personalisation.

However, some positive steps are being taken by the 6G working group within 3GPP, which is discussing the building of data framework and data plane to for example collect, store, process and expose, data from various sources (including devices, RAN, core network, sensors).[18] This suggests that the current data-modelling framework may be replaced by a data framework and/or data plane configuration.

[15] See Section 2 of the GSMA's mobile economy report 2025, https://www.gsma.com/solutions-and-impact/connectivity-for-good/mobile-economy/wp-content/uploads/2025/02/030325-The-Mobile-Economy-2025.pdf

[16] However, year-to-year trends on network energy efficiencies are inconclusive as yet, due to more and more operators being added to the study each year.

[17] Agentic AI for Telecom: Charting the Course for an Intelligent Future, GSMA, 2025. See here https://www.gsma.com/solutions-and-impact/technologies/artificial-intelligence/agentic-ai-for-telecom-charting-the-course-for-an-intelligent-future/

[16] See here https://www.ericsson.com/en/blog/2025/6/blog-6g-standardization-technology-step-to-publish

Glossary of terms

3GPP	Third generation partnership project
A2A	Agent-to-agent
AF	Application function
AI	Artificial intelligence
API	Application programming interface
AR	Artificial reality
BSS	Business support system
CPU	Central processing unit
CSP	Customer service provider
DN	Data network
E2E	End to end
EU	European Union
GenAI	Generative AI
GPU	Graphics processing unit
GSMA	Global System for Mobile Communications Association
h-p	Hyper-personalised
IR	Immersive reality
ITEF	Internet Engineering Task Force
LCM	Lifecycle management
LLM	Large language model
MCP	Model context protocol
MIMO	Multiple-input multiple-output
ML	Machine learning
MNO	Mobile network operator
NEF	Network exposure function
NPU	Neural processing unit
NWDAF	Network Data Analytics Function
OSS	Operations support system
QoD	Quality on demand
QoE	Quality of experience
QoS	Quality of service
RAN	Radio access network
RTA	Runtime assurance architecture
TPU	Tensor processing unit
TV	Television
UE	User equipment
UPF	User plane function
VR	Virtual reality
XR	Extended reality

About the authors

Caroline Gabriel (Partner) leads Analysys Mason's *Networks and Cloud* research practice, as well as leading many research activities related to 5G, 6G and network economics that span multiple programmes. Her focus is on critical issues and trends related to mobile and wireless infrastructure, particularly operator deployment intentions for 4G, 5G, virtualised RAN, end-to-end open networks and other technologies. She has led research and consulting projects with a wide range of clients, including mobile infrastructure vendors, large and start-up operators, regulators, trade bodies, government agencies and financial institutions.

Janette Stewart (Partner) has 30 years' experience of working on wireless technologies and spectrum-related matters, including wireless network strategy, wireless infrastructure deployment (2G through to 5G, FWA, satellite), frequency planning, spectrum strategy and valuation. She is part of Analysys Mason's consulting team in the UK, and has advised on 4G and 5G network architectures, RANs and spectrum deployment and is actively tracking developments on 5G evolution and 6G.

Adaora Okeleke (Principal Analyst) leads Analysys Mason's *AI and Data Platforms* research programme. Her research focuses on service providers' adoption and use of data management, artificial intelligence, analytics and development tools to support the digital transformation of network, customer and other business operations. Adaora tracks vendor strategies for the telecoms industry to understand how they are evolving their product portfolios to include data, AI and development capabilities. She also provides key industry insights to operators and vendors on strategies for adopting these technologies.

James Kirby (Senior Analyst) is a member of the *Networks and Cloud* research team based in Analysys Mason's London office. He manages the *Wireless Technologies* programme and leads research across key topic areas such as RAN, Open RAN/vRAN, 5G/4G/3G, core networks, 5G standalone (SA), spectrum and antenna technology.

Get in touch

enquiries@analysysmason.com

Analysys Mason is a global technology, media and telecoms management consulting firm

We help clients navigate complex transformation journeys and make and act on the decisions that steer their progress and connect our world.

We have been at the forefront of pivotal moments that have shaped the technology, media and telecoms industry for four decades — from guiding government agencies to forge the path for mass connectivity, to defining the strategies for companies exploring the opportunities of space and satellites.

Together we are shaping the next

Analysys Mason has its head office in London, and offices in Bonn, Dubai, Dublin, Frankfurt, Kolkata, Madrid, Manchester, Milan, Munich, New Delhi, New York, Oslo, Paris, Riyadh, Singapore and Stockholm.

www.analysysmason.com