NEXT G ALLIANCE
An ATIS Initiative

Trust, Security, and Resilience
Sustainability
Audacious Goals
Digital World Experiences
AI-Native Wireless Solutions
Cost-Efficient Solutions
Distributed Cloud and Communications Systems
AI-Native Wireless Solutions

6G

Next G Alliance Report: **AI-Native Wireless Networks**

TABLE
OF
CONTENTS

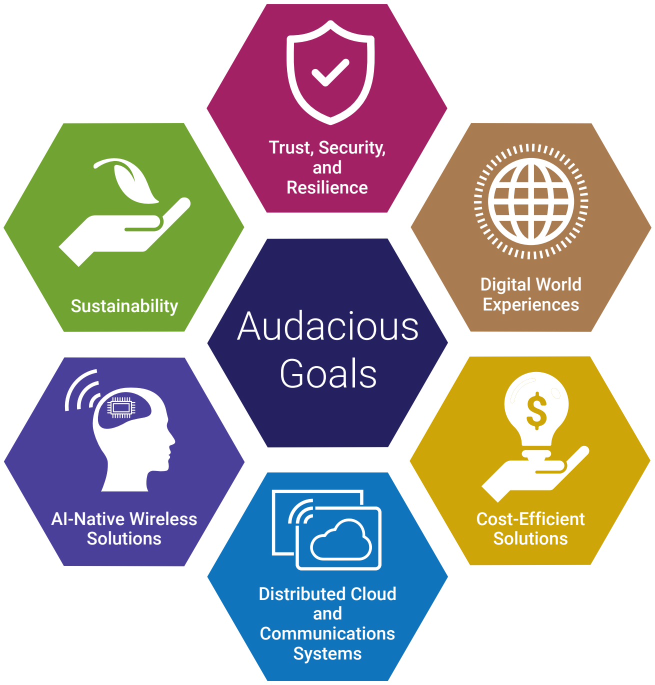
FOREWORD

As a leading technology and solutions development organization, the Alliance for Telecommunications Industry Solutions (ATIS) brings together the top global ICT companies to advance the industry's business priorities. ATIS' 150 member companies are currently working to address network reliability, 5G, robocall mitigation, smart cities, artificial intelligence (AI)-enabled networks, distributed ledger/blockchain technology, cybersecurity, IoT, emergency services, quality of service, billing support, operations and much more. These priorities follow a fast-track development lifecycle from design and innovation through standards, specifications, requirements, business use cases, software toolkits, open-source solutions, and interoperability testing.

ATIS is accredited by the American National Standards Institute (ANSI). ATIS is the North American Organizational Partner for the 3rd Generation Partnership Project (3GPP), a founding Partner of the oneM2M global initiative, a member of the International Telecommunication Union (ITU), as well as a member of the Inter-American Telecommunication Commission (CITEL).

For more information, visit www.atis.org. Follow ATIS on Twitter and on LinkedIn.

The ATIS Next G Alliance is an initiative to advance North American wireless technology leadership over the next decade through private sector-led efforts. With a strong emphasis on technology commercialization, the work will encompass the full lifecycle of research and development, manufacturing, standardization, and market readiness.


Trust, Security, and Resilience
Sustainability
Digital World Experiences
Audacious Goals
AI-Native Wireless Solutions
Cost-Efficient Solutions
Distributed Cloud and Communications Systems

EXECUTIVE SUMMARY

Native support of artificial intelligence (AI) is widely expected by the industry to be one of the major features of the next-generation wireless network and is deeply woven into the mission of the Next G Alliance (NGA) to establish North American leadership in 6G and beyond. It is one of the six audacious goals of the National 6G Roadmap working group and as such addresses key North American imperatives. It is also one of NGA's research priorities, meant to identify driving forces, technical challenges, and research directions.

Although the initial application of AI and machine learning (ML) to wireless networks began with 5G, its application in 6G will be more real time, more comprehensive, and seamlessly integrated into the wireless system design. This paper surveys the research and technology directions required to make the vision of an AI-native wireless network a reality. The most challenging task in front of us is the integration of AI/ML into the layered architecture of current wireless networks. Research directions include issues like AI/ML model lifecycle management, performance evaluation and interpretation, training data availability, the mandate for extra system capability, security, and privacy. In addition to the challenges, the application of AI/ML will also bring new opportunities, such as joint optimization of network and device operations through distributed learning and federated learning.

The path toward an achievable AI-native wireless network laid out in this paper identifies two of the major challenges: the difficulty of obtaining real data from wireless networks for model training, validation, and testing, and the impact of AI's application on existing communication standards and communication systems.

1 INTRODUCTION

The Next G Alliance is a bold, new initiative to advance North American wireless technology leadership over the next decade through private-sector-led efforts in association with government stakeholders. With a strong emphasis on technology market-readiness, the Next G Alliance's work will encompass the full lifecycle of research and development, manufacturing, standardization, and market readiness. This effort is intended to be a reference to drive North American leadership across industry, academia, and government stakeholders to meet Next G Alliance objectives.

Next G Alliance has identified six audacious goals that describe top priorities for North America's contribution and leadership in these future global standards, deployments, products, operations, and services. An AI-native wireless network is one of those six audacious goals to increase the robustness, performance, and efficiencies of wireless and cloud technologies against more diverse traffic types, ultra-dense deployment topologies, and more challenging spectrum situations.

The term *AI-native* is used to indicate that AI is incorporated into major functionality from the very beginning of the design and development cycle of a system (application, device, network). An AI-native 6G system leverages AI techniques (e.g., ML, deep learning, neural networks.) for the design, deployment, management, and operation of various network and device functions.

Next G Alliance's goal is to promote critical applications of AI in the next generation of wireless to advance the North American leadership in the field of wireless communications and fulfill market needs. AI-native networks need to be fully trusted by people, businesses, and governments to be resilient, secure, privacy preserving, safe, reliable, and available under all circumstances. Therefore, it closely relates to the trust and security audacious goal of the Next G Alliance.

6G will build up its capabilities on distributed clouds, and a tight AI-native integration between communication and computing is the natural evolution. It is expected that 6G wireless standards need to be developed in an AI-native way, enabling a large set of applications that may rely on enhanced real-time capabilities. Overall, the application of AI/ML to 6G will bring a shift in how networks are designed and implemented. The application of AI to 6G will come at multiple stages and different flavors [1]. AI/ML can be used to improve individual functions or modules, as well as in an end-to-end fashion, jointly optimizing multiple functions or modules. Initial applications of AI/ML are expected by 6G's launch, with more advanced applications emerging by 2030 and beyond.

Application of AI/ML is being developed in 5G and ongoing 5G Advanced standardization, but with a focus on enhancing the existing air-interface and network modules and functions specified in more traditional ways. AI/ML in 6G will be an inherent enabler rather than an overlay, as shown in Figure 1.

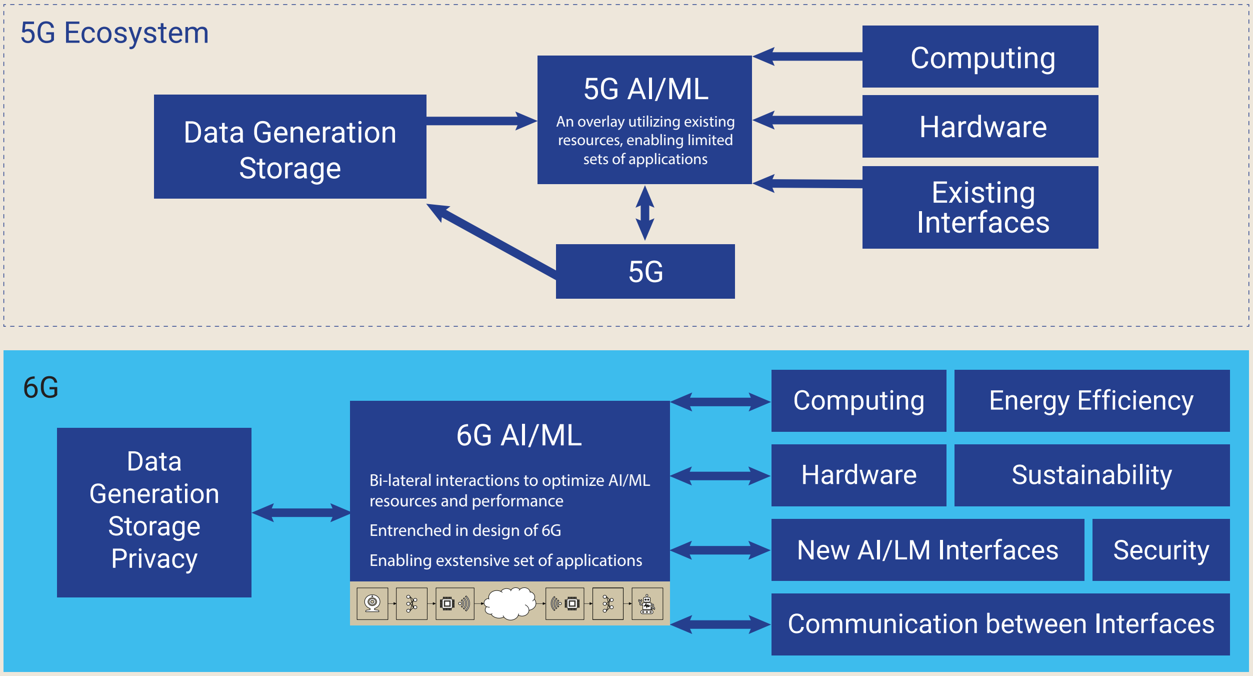

5G Ecosystem
Data Generation Storage
5G AI/ML
An overlay utilizing existing resources, enabling limited sets of applications
5G
Computing
Hardware
Existing Interfaces
6G
Data Generation Storage Privacy
6G AI/ML
Bi-lateral interactions to optimize AI/ML resources and performance
Entrenched in design of 6G
Enabling exstensive set of applications
Computing
Energy Efficiency
Hardware
Sustainability
New AI/LM Interfaces
Security
Communication between Interfaces

Figure 1: 5G AI/ML and 6G AI/ML [1]

Therefore, it is necessary that AI/ML be entrenched in the design of radio layers with interfacing to AI and data-collection frameworks. These interactions need to be enabled with a strong emphasis on security and privacy. Such an AI-native approach in 6G could be a foundation and may allow the wireless technology to evolve at a faster pace independent of standards cycles.

It is our understanding that, when comparing the applications of AI/ML in 5G/5G Advanced, the applications of AI/ML in 6G should have the following features:

> Although AI/ML applications for 5G are mostly operated in near-real-time fashions, many AI/ML applications for 6G will operate in the real-time manner.

> The applications of AI/ML in 5G are limited, but its application in 6G will be comprehensive.

> The design of 5G didn’t take AI/ML into account, whereas AI/ML will be blended into the design of 6G.

> Although data collection was not present in 5G, in 6G it will be at various layers and within the network.

> In 6G, the transceivers may be fully redesigned to be AI-native. It is also possible that AI/ML in 6G will be able to create/learn new signaling and procedures based on the environment it is working in.

2 DRIVING FORCES AND NORTH AMERICAN IMPERATIVES

Over multiple generations, the networks designed and deployed have become increasingly complex involving coexistence of radio access networks (RANs), multiple frequency bands, and with increasing variety of optimizations for different use cases and scenarios. These factors drive the need for automation using machine intelligence. Access to data from many components of wireless networks — user devices, network nodes, and sensors — along with improved centralized and distributed computing capabilities, are other drivers of AI-native wireless networks. Recent advances in AI/ML are enabling near-real-time applications of AI/ML to mobile networks. Using traditional methods, suboptimal solutions may be preferred because of the optimal solution's complexity. The air interface has variable latency, reliability, and sensitivity to channel variations. In many cases, it is difficult to even analytically model the problem or scenario involved, and an optimal solution may be unknown. Networks have become significantly more complex, and performance has been improved incrementally using classical techniques for signal processing and receiver design. AI will be incorporated into major functionality from the very beginning of the system's design and development cycle. AI-native wireless networks are expected to have a better performance-complexity tradeoff by tackling the traditional mobile network problems in a completely new way using data-driven approaches instead of traditional algorithms. That may lead to new 6G use cases which have not been considered before.

In 5G systems, it is already well known that there are thousands of parameters that have to configured. As a result, optimizing the system manually will not be an option in the future. Moreover, the optimal network configuration also depends on the surrounding environment, calling for fine-grained adjustments. With AI/ML-based techniques, the network may better adapt to users and applications seamlessly and automatically in real time, which means that the network achieves a larger portion of its theoretical capacity. With a fully AI-native 6G network, it could even be possible to improve the physical layer based on the prevailing channel conditions, allocated frequencies, etc.

North America already excels in disruptive AI research and can therefore widen the gap with other competing nations. AI-native networks are a significant opportunity for North American operators and industry players develop new ways to optimize networks and device operations and to enhance customer experiences. A successful, broad application of AI/ML to 6G can have an economic impact on the telecommunications industry through automation of network operations, a shift in how networks are designed and implemented, eliminating the need for new hardware platforms, and improving network performance across areas like energy efficiency, computational efficiency, and spectral efficiency.

Autonomous, energy-efficient, sustainable telecommunications networks will impact North American industry players, societies, and workforces. Enabling network automation through AI/ML-aided management and orchestration will reduce long-term costs. As for the workforce, the shifts in required skillsets can have a major impact on education and other training programs. It is important that society as a whole prepares for the impending changes for the next decade, with governments playing a significant role. North America needs to be at the forefront of such economic activity and pursue substantial research efforts across the various dimensions of the application of AI/ML to communication systems.

3 RESEARCH AND TECHNOLOGY DIRECTIONS

3.1 Options for Integrating AI/ML into Future Networks

AI and ML offer many potential benefits and enhancements for the operation of 5G and 6G wireless networks and may also enable new capabilities and functions within the networks. Hence there are currently many technical options and approaches being considered, researched, and proposed for the application of AI/ML into wireless networks.

A first step that is already being studied in 3GPP for standardization, as well as being implemented in various proprietary solutions, is to add AI/ML to existing functions to enable improvements in those functions. This is normally implemented by using AI/ML to optimize the selection of parameters or variables to provide a more optimum set of operating parameters in the related function. The optimization may be selected for performance, speed, energy saving, or any other key parameter of the function. Note that the actual function is unchanged. Instead, AI/ML powers the selection of parameters within the function. For example, a radio resource scheduler can use AI/ML inference to select appropriate resources for an individual user.

A second step is to replace an existing function within the network with an AI/ML-powered function, where the operation of the function has been designed for the use of AI/ML. This approach enables the operation of the function to be more specifically optimized for AI/ML and to benefit more directly from the AI/ML capabilities. This approach can be applied to virtual network functions (VNFs) in the 5G Core (5GC), where the function's external inputs/outputs remain the same, but the actual operation of the function is powered by AI/ML. The approach can also be applied to individual function blocks within the RAN and user equipment (UE) such that a specific function in the RAN or UE transmitter/receiver chain is now powered by AI/ML. An example here may be for beam management, which is currently being studied in 3GPP Release 18 with AI/ML approaches, utilizing either spatial or temporal domain patterns of the beams. In 6G, we expect functions like beam management will be further enhanced with the AI/ML approach.

Beyond the steps of function-level application for AI/ML, it is also considered that joint learning/inference can be deployed across multiple network functions. This is expected to enable further improvements in the related functions because the AI/ML can take a wider set of inputs and provide inference with a wider context. An example here is load balancing between cells in a network, where the scheduling and admission/mobility control across multiple cells can use AI/ML inference to determine optimum mobility and scheduling actions across the multiple cells. Although this load-balancing capability is already able to use ML in 5G networks to improve performance, we could see the 6G implementation having enhancement of interface APIs to enable joint inference across multiple network components. This would enable more advanced schemes to be implemented.

As a further step in the application of AI/ML, significant new concepts may be introduced via AI/ML, such as a redesign of key functionality including radio resource management. Furthermore, AI can be extended over service chains of multiple network functions to provide a system-level approach to improving performance. An example of this might be to use AI inference to manage mobility and cell configuration based on the real-time geographical movement of users in the cell, applications and services currently being used in the cell, together with the load in the cell. The load and configuration of a group of cells located at a sports stadium, traffic intersection, or airport terminal can have dynamically changing load requirements. AI/ML can create better ways of managing and configuring those cells.

RF sensing is another candidate for AI/ML utilization in 6G. AI/ML is already demonstrated for object recognition in video, lidar, and radar applications, so there can be an application for using AI/ML inference to power both the RF sensing waveform selection and the object detection/classification/recognition functions.

3.2 Applying AI/ML Learning Algorithms and Models

It is arguable that the entire gamut of AI/ML concepts and algorithms — such as deep learning, reinforcement learning, regression, decision tree, K-means clustering, and federated learning — are applicable to wireless network evolution, as well. The deployment of these algorithms within multiple layers of the network will be governed by resource intensity, latency requirements, power requirements at end user devices, and several other considerations. The choice between real-time and non-real-time tuning of these algorithms should also be made based on the considerations of cost and performance.

3.3 Computing Architecture for AI/ML

Current AI and ML methods are considered to be power intensive due to a high load of compute operations that are required, typically measured as Tera Operations Per Second (TOPS). In addition, the compute complexity of AI and ML can be considered to increase significantly when the number of input parameters is extended. These factors lead to the requirement for careful consideration of compute architecture with AI/ML-powered functions to ensure the required level of TOPS processing power is available.

The training functions for AI are normally performed in a central location, prior to the models being deployed into use in the inference engines. This normally leads to the offline learning and model training being located on a central compute platform where a high level of compute resources is available. With federated learning, some of this learning task may be passed down to the end device, with the model updates from each user then being shared back to a central location for a federated update to all users. In 6G we may consider a revised compute architecture using standardized interfaces and APIs to define the transfer of data for this purpose. This type of data transfer between different network entities (such as between UE and core network operator) may also require suitable business arrangements between the different entities to agree on permission for the data transfer.

Inference at a central compute platform is attractive because it enables the inference engine to receive input with many parameters and enables the use of high-performance compute platforms to perform complex AI inference tasks. However, this architecture puts a high load on the communications network that connects the end device to the inference engine because all the raw data from the device must be transferred across the network. Moving the inference function out to the end device will remove the need to transfer data across the network but requires costly high compute power (and related size, weight, and power requirements) in the end device. So, a combination of edge compute (on device or in a local compute facility close to the network connection point of the end device), distributed compute (resources placed in the region to provide larger scale compute without transferring data to a distant location), and central compute (to provide hyperscale compute power for highest performance) is designed into the architecture of the AI/ML function.

An example of this architecture can be seen in surveillance camera recognition, where an on-device or edge compute AI function can be used to recognize and classify certain objects in an image (e.g., vehicle license plate, a human face). This object class can then be extracted from the image and only this sub-portion of the available data is transmitted to the central compute, where full classification will take place (e.g., optical character recognition of the license plate, or facial feature extraction and profile matching for a human face).

In this example, a relatively simple AI task (with low compute power requirements) is performed at the edge and reduces the volume of data needing to be sent across the network to central compute resources. Then the more complex and resource-intensive operations still take place in a central compute location, where suitable resources are available. This split of inference and compute load (across the network from device/edge, to distributed location, to central location) is becoming more flexible as more flexible cloud compute resources become available. The availability of virtualized compute resources, within a standardized platform architecture, offers the ability of distributed cloud systems to meet the requirements for optimizing the load of AI/ML compute resources.

In addition to the volume of data load put onto a network when AI/ML is performed in a central location, the centralized location of the compute resource can also create significant latency in the response time of an AI/ML function. Where an AI/ML function is related to a latency-critical task, then the transmission time latency of the communications network should also be considered. This may lead to the requirement for low-latency tasks to be located at the edge of the network (e.g., a 1 ms latency connection must be within 300 km, providing a round-trip response time of 2-3 ms). So the design of the compute architecture for AI/ML must also consider transmission latency when allocating inference workloads between different distributed compute elements.

As noted above, AI/ML learning and inference of complex functions using a complex data set is relatively intensive for compute resources and compute operations. In addition, the transfer of large volumes of data sets from an edge device to central compute locations consumes significant communications network resources. Both of these require significant energy resources to power them, and so the widespread deployment of AI/ML can be seen as contributing to increasing energy consumption of 5G and 6G networks and services. This requires that significant work and progress in reducing the power consumption to ensure that the carbon footprint is reduced and that environmental sustainability of 6G networks is achieved. Research directions are including the design of more efficient and low-power AI/ML inference methods, more efficient model training and learning methods, and the design/optimization of compute resources to spread compute load in the most power- and cost-efficient manner.

Finally, it can be noted that AI/ML offers the capability to improve the management of power consumption within a network. An example is using AI/ML inference to activate/deactivate a radio cell and reduce the energy consumption of the entire RAN. If the expected traffic volume is below a fixed threshold, AI/ML can turn off the cell and offload terminal communication to another cell. From the data collected by the RAN network, AI/ML is used to predict energy efficiency and load status, and cell activation/deactivation is performed to save energy. As the telecom industry moves to focus more activities into reducing power consumption and minimizing carbon footprint, 6G specifications can be expected to provide a framework for a compute architecture where such AI/ML functions can be implemented and operated with better power efficiency.

3.4 Performance

AI/ML performance is key for its successful adoption and use in applications or functionalities intended to overcome the shortcomings of current 5G AI/ML enhancements and to evolve toward a truly AI-native RAN. Remarkable performance milestones have been achieved in the broader field of AI/ML, from smart end user devices or applications to unmanned and/or autonomous vehicles and robots. AI/ML methods are steadily replacing conventional algorithms with AI/ML models and structures, such as in speech/image recognition and video processing, to name only a few examples. Notably, one open-source application of so-called "generative AI" is chatGPT (by OpenAI), with applications ranging from writing and debugging code in select programming languages to writing and composing music, lyrics, or scripts to language translation — albeit not without controversy or disruption.

Despite such remarkable progress in the broader field of AI/ML, its application to AI-native wireless communications is still — *ad literam* — nascent. Certain enablers will be required for the latter to mature, such as hardware acceleration tailored to ML operations, minimum performance guarantees, and consistency of performance for the end-to-end AI-native system.

Complexity of implementation and cost efficiency will also be factors in the selection of the type of AI/ML employed by AI-native wireless communications. While complexity is inherently related to performance, the latter may choose to focus on either success rate, cost efficiency, or accuracy. This reflects the dichotomy between AI and ML inasmuch as the former (very much in infancy at the time of this white paper) aims to achieve intelligence ("success rate") via decision-making. Meanwhile, the latter seeks algorithms that allow systems to learn from data (and past experience; thus "accuracy") via, for example, supervised, unsupervised, or reinforcement learning (*cf*. above).

While ML is a subset of AI, a further sub-class of ML is, in turn, deep learning (DL), which currently involves Deep Neural Networks (DNNs) in one of the following configurations: pretrained but unsupervised, recurrent (chain) or recursive (tree), and convolutional NNs. DNNs are expected to achieve superior accuracy (over ML per se) by exploiting and leveraging large amounts of data (datasets) along with higher model complexity. In one aspect, AI-native problems with a smaller dimensionality will likely lend themselves to plain ML algorithms having *lower* complexity. Meanwhile, a large number of dimensions and parameters (vs. hyperparameters) and/or a higher nonlinearity will require DL and implementations that can handle increased complexity — either algorithmically or by using dedicated acceleration hardware.

DL and DNNs have seen a remarkable revival and advancement since the early 2000s. Their superior performance (in terms of accuracy), while coming at a complexity cost, relies on more than one hidden layer. This enables them to extract high-level features present in the input space by using statistical learning along with large datasets to obtain an efficient representation of the input. Here, efficiency refers, for example, to dimensionality and noise reduction by efficiently projecting the input onto subspaces while preserving information and optimizing the (information) statistic for subsequent processing. In principle — from a performance-improvement perspective based on architectural refinements — multiple and fully connected layers along with multiple, nonlinear activation functions allow AI/ML structures to model high-dimensional spaces along with nonlinearities. This, in turn, can help bridge the gap between *linear* yet *sub-optimal* algorithms to *optimal nonlinear* statistical signal processing.

The architecture of DNNs can be — at a very high level — conceptualized by three basic structure types: Multilayer Perceptron (MLP), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). The MLP and CNNs are *feedforward* (FF) types of NNs. They rely on stochastic backpropagation algorithms — during model training without affecting the FF nature, with variations such as gradient descent with momentum — in order to learn the weights and biases that interconnect the hidden layers. Moreover, CNNs are not necessarily "fully connected" and use spatial filtering in order to extract features from the input of the DNNs (images). Information travels only forward — during the inference step of AI/ML — and such DNNs are generally suitable for nonlinear classification and prediction problems. On the other hand, sequential data, and the implicit memory (or "correlation") embedded in a sequence, require RNNs, which have loops and states and are configured to memorize parts of the input and to make local predictions.

In one fundamental aspect, RNNs suffer from the problem of so-called vanishing gradient (during backpropagation, which can be solved by variations like Long-Short-Term Memory (LSTM), bi-directional LSTM, and Gated Recurrent Unit (GRU) RNNs. A possible further refinement of RNN (with promising potential) involves the mechanism of "attention," whereby every unit of a GRU is allowed to look at a larger pool of information than that of the immediate past state. A related flavor of DL is the autoencoder, suitable for compression (as in source coding, *cf*. below). Not only does this have promising potential, but the transformer architecture is used in GPT-3 and powers ChatGPT.

To date, performance in many applications of AI/ML predominantly relies on the user's judgment or perception. For both success rate and accuracy, the introduction of AI-native in wireless communications will require more indirect and abstract measures of performance, with less human subjectivity — even while preserving some of the cost functions for training the model(s).

AI/ML performance will be ultimately influenced by the speed and quality of model(s) training, the availability and quality of the datasets, and by model lifecycle management. Training is the process whereby an AI/ML system learns to perform its task(s) by optimizing the values of its parameters (weights and biases in a DNN). During model inference, the learned model is used to perform the designed task. Hardware acceleration will be key — as will allocation of the computational burden across RAN, cloud, and user devices. AI-native wireless communications still face a need to research, refine, and specify data collection procedures for model training. Considerations for user data privacy, along with specifications for data collection enhancements and associated signaling, will be essential in 6G. In addition, a balance between model generalizability and specificity will influence performance. General models might have lower accuracy but are simpler to deploy. Specific models may perform better, but managing many models might become difficult.

AI/ML model training latency is a fundamental performance aspect with several components. Model transfer between network nodes can further add latency. *Communication* latency depends on downlink (DL) and uplink (UL) data rates for model distribution and respectively for trained model updating. *Computational* latency depends on the computation/memory resources available on training devices.

Overall latency is determined by the larger among the two. To reduce the training latency, more efficient compute engines tailored for ML operations are often used, such as graphics processing units (GPUs) and network processing units (NPUs). Although model inference is less computationally demanding, its complexity, including the pre- and post-processing of data, needs to be considered (e.g., FLOPs). Model complexity, such as the number of parameters (or size), will affect storage requirements and the overall latency. At the time of the release of this white paper, the state-of-the-art in GPUs for AI is NVIDIA's H100, which will replace the current A100 GPU. The H100 is up to 9X faster than A100, at least with regard to training, and is expected to be released in 1Q23. Although the price point may need to ameliorate, it is expected to significantly accelerate AI-native algorithms for wireless communications in the cloud, or at nodes like gNBs, perhaps even leveraging Open Radio Access Network (ORAN) architectures.

Nevertheless, AI-native solutions at the end-user terminals will be more sensitive to complexity aspects because the availability of hardware acceleration will likely see a delay due to affordability and/or power consumption. In another aspect, the tight coupling between training and inference will render model monitoring and updating more stringent. Consistency of performance and minimum performance guarantee will require appropriate lifecycle management (LCM) procedures, including initial model training, model deployment, model transfer, model monitoring, and model updating/selecting.

AI/ML techniques can reduce the over-the-air overhead, such as the reference signals (RS) for beam management or positioning, or the overhead associated with CSI feedback. This reduction will be countered by new types of overhead due to data collection, information exchange, model delivery and transfer, or other AI/ML-related signaling. The overall dynamic needs to be studied and understood.

According to NGA's vision for 6G (*cf*. elsewhere in this document, including the discussion about datasets and security aspects), AI-native wireless communications is expected to continue and build on the performance of current 5G AI/ML enhancements by addressing their shortcomings and expanding their scope. As for continuity and precedent, the entry point for AI-native 6G networks will reflect the status of 5G AI/ML enhancements, as briefly reviewed below.

5G networks have become increasingly complex and capable of generating huge amounts of data. This enables data-driven AI/ML techniques throughout AI-native wireless networks. Further leveraging AI/ML techniques beyond 5G requires industry alignment through global research and standardization efforts. A variety of AI-related activities have emerged in many standardization bodies, including 3GPP and the O-RAN Alliance, among others. In Release 17, 3GPP initiated a study on AI-enabled NG-RAN covering high-level principles, functional framework, potential use cases, and associated solutions for AI-enabled RAN intelligence. 3GPP Release 18 aims to specify data collection enhancements and signaling support for a selected set of AI-based use cases, including network energy savings, load balancing, and mobility optimization.

An AI-native 5G NR air interface is being studied in 3GPP Release 18 with the goal of enhancing performance, including the reduction of complexity/overhead. The study is expected to set a common foundation for a scalable AI/ML framework for the air interface, identify areas where AI/ML can improve air interface functions, investigate the description, characterization and management of AI/ML models/lifecycle, assess AI/ML techniques to understand their gains and complexity, and assess standardization impacts. To gain traction towards these objectives, 3GPP Release 18 is focused on an initial set of selective use cases, specifically CSI feedback, beam management, and positioning.

At the network level, general principles for AI-native RAN, an AI/ML functional framework, and potential use cases were reviewed in a 3GPP Release 17 study item (SI), with identified potential solutions (*cf*. below). A subsequent Release18 work item (WI), in progress at the time of this white paper, aims to specify data collection enhancements and signaling support within existing NG-RAN interfaces and architecture (both split and non-split architectures). AI/ML will attempt to leverage the collection of RAN data in order to:

> Optimize network energy saving (NES) by predicting energy efficiency and traffic load in subsequent states and enabling proactive, adaptive actions of traffic offloading, coverage modification, and cell (de)activation.

> Build datasets with measurements and feedback from network nodes and UEs in order to predict load and improve network performance and user experience.

> Improve handover performance, predict UE location and performance, and steer the traffic to achieve quality network performance.

3GPP Release 18 will mark the first attempt at standardizing AI/ML in a wireless air interface by examining potential performance enhancements in three initial use cases:

> **Channel state information (CSI) feedback:** An auto-encoder aims to capture spatial, angular, frequency- and time-domain correlations in the channel matrix, for which CNNs or a type of RNNs are strong candidates. Initial evaluations indicate significant CSI feedback overhead reduction (even up to 60%, depending on traffic load) versus Release 16/17 Type II codebooks.

> **Beam management (BM):** Data-driven AI/ML methods can exploit the historical information in the training dataset to construct a mapping function from sparse beam sweeping measurements to a best beam pairing. Spatial domain beam prediction with AI/ML (and 64-DFT codebook) can outperform the legacy approach in most cases with regard to beam selection accuracy. For example, AI/ML-based top-5 beam prediction can reach 94.95% prediction accuracy while further reducing overhead by 67.17% versus 55.3% with a legacy approach (for the same overhead reduction).

- **Positioning accuracy:** AI/ML-based methods promise solutions for positioning in the more demanding NLoS scenarios by learning, in essence, an acceptable mapping function from measurements to UE positions. Both Direct ML/AI and ML/AI-assisted methods show significant improvement in positioning accuracy even in heavy NLoS environments. RF fingerprinting (RFFP) shows promise when used on the same site where training occurred. Further improvements and generalizations down to tens of centimeters are being pursued.

3.5 Datasets

Training and retraining are essential aspects of the ML workflow. The models generated by the AI engine are only as good as the training datasets, which must satisfy some essential conditions. The training dataset must have a statistical similarity to the operational data of the system in deployment. It must provide sufficient coverage for all possible deployment conditions and outcomes envisioned for the specific deployment model, including outlier scenarios. The industry as a whole must tackle the scarcity of such datasets in the public domain, which hampers innovation. One possibility is to create a commonly available pool of datasets pertaining to each of the use cases being considered for AI/ML-based optimization.

However, there are oncerns about security, privacy, and bias in this area that must be tackled. This may be a domain where a North America-specific solution may be explored. The use of synthetic data based on realistic emulations in digital twins is also a legitimate alternative, as long as questions about their similarity to real-world data are sufficiently answered. On the one hand, we discuss the lack of data, but on the other, in a counter-intuitive fashion, where there is data (such as within the operators' domain), there is an abundance of it. The thousands of possible measurements and performance metrics amount to a glut of data that must be adequately harnessed. Proper labeling and classification of data, and the removal extraneous or redundant streams, are all prerequisites to creating viable training datasets.

Datasets from mobile network operators are highly proprietary and are likely to remain so. As a result, access to the data requires a business relationship between the model developers and mobile network operators.

3.6 Security

AI/ML has been adopted in current 5G network, such as O-RAN's Radio Intelligent Controllers (rApps and xApps), and the core network's NWDAF for core NFs optimization. In 5G beyond and 6G, native AI and cross-domain AI are the research focus across network management and composition, signal processing and physical layer, service-based communication, and data mining.

Data poisoning is a security attack of tampering with ML training data to produce undesirable outcomes from the AI/ML serving model/algorithm. Typically, a data poisoning attack either targets availability by injecting or manipulating data into the data store to render a total ineffective and/or inaccurate ML algorithm, or targets integrity by leaving an unnoticeable backdoor into the data set controlled by the attacker to deal a fatal blow at certain time to a seemingly working ML model. There are four steps to protect against such data poisoning attacks:

- Implement tight security controls to protect the data store(s) access wherever it resides (central or distributed).

- Data extraction/validation/preparation for generating training data set from the data store(s) to support capabilities to detect and filter out outliner/bad data.

- Wireless device or network element endpoint protection ensures that the individual raw data or logs are secured from tempering or manipulation.

- Data (raw or training) in transition should be secured with encryption.

Finally, continuous testing/validation of the security countermeasures listed here will keep AI/ML secured from the data poisoning attack. Adversarial threat simulation and verification of network response has to be a continuous process in order to ensure service assurance. The security threat surface is constantly evolving, so the network response also must evolve.

3.7 New Opportunities

Joint optimization of network and device operations may potentially be expanded in scope. In AI/ML-based CSI feedback, compression and decompression are performed jointly in an autoencoder DNN, yielding enhanced performance. There exist many scenarios where joint optimization of network and device operations can improve the overall performance, and neural networks are natural candidates. Similarly, cloud-based training requires that enormous amounts of training data be moved from devices to the cloud, facing prohibitive communication overhead (and data privacy issues).

Alternatively, AI/ML model training can be performed jointly among multiple network nodes. Distributed Learning and Federated Learning are two such examples. In Distributed Learning, each computing node trains its own local model using local data, thus preserving privacy. Network nodes will communicate with one another to exchange the local model updates and build a global DNN model. In Federated Learning mode, a cloud server trains a global model by aggregating local models partially trained by individual end devices. Within each training iteration, a UE performs the training based on the model downloaded from the AI server, but it uses local training data. Then the UE reports the interim training results (e.g., gradients) to the cloud server via 5G UL channels. The server aggregates the gradients from the UEs and updates the global model. The updated global model is distributed, via 5G DL channels, to UEs, which can perform the training for the next iteration.

In fast-changing environments, AI/ML models running on devices or on the network side must be continuously adapted to new environment to keep the desired performance. One solution is to perform online model updating, which requires retraining the model with new training data. However, online model updating might be computationally expensive, especially for UE devices. A solution to this challenge is to adaptively select the model for inference from a set of trained models as shown in figure below, known as continual and dynamic adaptation of network and device operations.

Perhaps the most obvious candidate for AI-native RAN is the Self-Organizing Network (SON) functionality (introduced in 3GPP Release 8). A SON self-adjusts and fine-tunes a range of parameters. By 2030, 5G SON will transition from a rule base to predictive AI/ML implementation of objective optimization. The difference between 5G and 6G is the tighter integration of AI/ML with 6G. SON is somewhat limited by the data collection interval, whereas in 6G, AI/ML improvements will bring it closer to real-time—up to constraints caused by moving data or updating models across interfaces.

Technology Direction	Summary
Integration of AI/ML into networks	> Improve existing function > Replace existing function > Joint learning/inference across multiple functions > Incremental improvement over 5G networks > Need to enable AI/ML models that are optimized for a wide range of applications > Need to move from near-real-time AI/ML to real-time AI/ML
Computing architecture for AI/ML	> Transition from large central compute platforms to distributed/edge/local compute > Flexible ML function split across the above > Workload management to optimize inference latency > In-built energy management
Datasets	> Essential conditions to be met by training datasets > Need to tackle industry-wide scarcity of datasets > Curation/labeling/classification of data and extraction of relevant streams from large datasets
Security	> Security vulnerabilities make an AI/ML-based system susceptible to data poisoning > Security should be enforced at all key data repositories and interfaces > Continuous, adversarial threat analysis is needed to tackle evolving threat surface

Table 1: Summary of Research and Technology Directions

4 PATH TO REALIZATION

In recent years, AI has been widely used in many industries (e.g., image and speech recognition, automatic driving), and has also made great achievements. Yet its application in the field of wireless communication is still in its infancy. In order for AI to achieve comprehensive and in-depth application in the field of wireless communication, the following two major obstacles must be overcome.

The first is the difficulty of obtaining real data from wireless networks for model training, validation, testing, and performance monitoring, as detailed in Section 3.5. A fundamental rule about AI/ML is that the performance of an AI model depends greatly on the quality of the training data it is fed with. However, in the field of wireless communications, it is very difficult to obtain operating data from real networks. The biggest issue, among others, is the privacy of customers. Mobile data is more or less related to the user confidentiality and privacy, as well as business intelligence of operators. Therefore, operators are generally extremely cautious when sharing training data, which is also very understandable. Another source for real data about wireless networks would be universities and research institutes in the US. Some have their own wireless network platforms, which generate mobile network data in real operating environments. Therefore, to address this lack-of-data issue, universities and research institutes could play a big role by sharing data obtained from their private or experimental mobile networks. The Next G Alliance also plans to release a mobile network dataset description guideline so that people can have a clearer description of the dataset when sharing.

The second obstacle would be the impact of the application of AI on existing communication standards and communication systems, such as the uncertainty of performance. Simple AI applications in wireless networks do not need to affect communication standards. But if we want to fundamentally apply AI to wireless communication networks, it is bound to affect the formulation of existing communication standards. Such applications, especially in the early stages of AI applications, will encounter many problems. The following is just a partial list of questions related to the issue.

> Does such an application require changes to existing communication standards? Is it necessary to increase signaling exchange and data transmission at the control and data planes?

> Will the application of AI itself place more burden on wireless communication networks? For example, model training may require a large amount of sometimes real-time network data. Transmitting this data itself has the potential to place a greater burden on an already congested channel. In addition, it is clear that the application of AI itself will consume both system resources and energy.

> How can AI systems be seamlessly integrated into existing communication systems and existing communication standards? Do we need to define any interfaces? Is it necessary to roll back to the traditional approaches when the AI model cannot meet the performance requirements?

In summary, if we want AI to be widely used in wireless communication systems, the most impactful way would be to introduce it into wireless communication standards (e.g., 3GPP and ITU-R standards). This requires shifting more of our research and efforts to the application of AI when developing the next generation of wireless communication standards. To do this, we will need to rely on the appropriate allocation of resources and the development of interdisciplinary talents mentioned earlier. Close cooperation across the industry is also crucial, especially the sharing of training data.

5 CONCLUSION

As one of the six audacious goals identified by the Next G Alliance, AI-native wireless networks are expected to increase the robustness, performance, and efficiencies of wireless and cloud technologies against more diverse traffic types, ultra-dense deployment topologies, increasing energy consumption, and more challenging spectrum situations. The Next G Alliance's goal is to promote critical applications of AI in the next generation of wireless to advance the North American leadership in the field of wireless communications and fulfill market needs.

Over multiple generations, the networks designed and deployed have become increasingly complex and are in need of automation for performance optimization and network configuration/management. However, the application of AI/ML to the wireless industry faces many unique challenges unseen in other industries, both technical and non-technical.

On the non-technical side, the largest challenge would be the difficulty of obtaining training data from wireless networks and, at the same time, protect user privacy and the carriers' core business secrete. On the technical side, major challenges include, but are not limited to:

- Non-deterministic model performance (a fullback mechanism is needed, doubling the network complexity).
- The layered communication architecture of the existing networks (requiring different levels of integration inside the existing network; in general, AI/ML favors a layerless architecture).
- Limited resources with mostly battery-powered UEs (e.g., limited capability for model training and inference).

These challenges would not prevent North America from excelling in disruptive AI research in all fronts and widening the gap with other competing nations. Given that the US is the leading country in the research and application of AI, and houses world-renown research universities and companies in the industry of wireless communications, there is no doubt that it will lead again in this AI/ML era. The NGA is thrilled to be the leader and promoter for this revolutionary use of AI/ML for the wireless industry.

6 REFERENCES

1. J. Hoydis, F. Aoudia, A. Valcarce, H. Viswanathan. *Toward a 6G AI-Native Air Interface*. IEEE. April 2021. https://arxiv.org/pdf/2012.08285.pdf

2. 3GPP TR 38.843: *Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface.* 3GPP. July 2022. https://www.3gpp.org/dynareport/38843.htm

COPYRIGHT AND DISCLAIMER

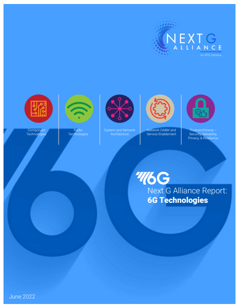
Published April 2023

Alliance for Telecommunications Industry Solutions
1200 G Street, NW, Suite 500
Washington, DC 20005

The information provided in this document is directed solely to professionals who have the appropriate degree of experience to understand and interpret its contents in accordance with generally accepted engineering or other professional standards and applicable regulations. No recommendation as to products or vendors is made or should be implied.

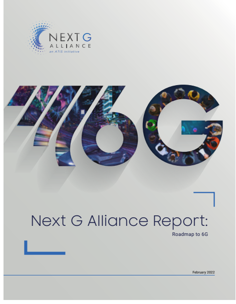
NO REPRESENTATION OR WARRANTY IS MADE THAT THE INFORMATION IS TECHNICALLY ACCURATE OR SUFFICIENT OR CONFORMS TO ANY STATUTE, GOVERNMENTAL RULE OR REGULATION, AND FURTHER, NO REPRESENTATION OR WARRANTY IS MADE OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR AGAINST INFRINGEMENT OF INTELLECTUAL PROPERTY RIGHTS. ATIS SHALL NOT BE LIABLE, BEYOND THE AMOUNT OF ANY SUM RECEIVED IN PAYMENT BY ATIS FOR THIS DOCUMENT, AND IN NO EVENT SHALL ATIS BE LIABLE FOR LOST PROFITS OR OTHER INCIDENTAL OR CONSEQUENTIAL DAMAGES. ATIS EXPRESSLY ADVISES THAT ANY AND ALL USE OF OR RELIANCE UPON THE INFORMATION PROVIDED IN THIS DOCUMENT IS AT THE RISK OF THE USER.

NEXT G ALLIANCE REPORTS



6G Technologies

6G Applications and Use Cases

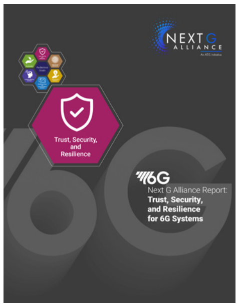


Roadmap to 6G

Green G: The Path Toward Sustainable 6G

NEXT G ALLIANCE
Distributed Cloud and Communications Systems
6G
Next G Alliance Report:
6G Distributed Cloud and Communications Systems

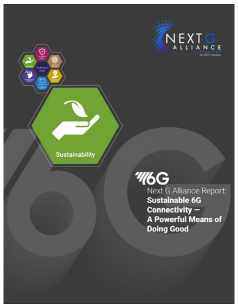
6G Distributed Cloud and Communications System


NEXT G ALLIANCE
Trust, Security, and Resilience
6G
Next G Alliance Report:
Trust, Security, and Resilience for 6G Systems

Trust, Security, and Resilience for 6G Systems

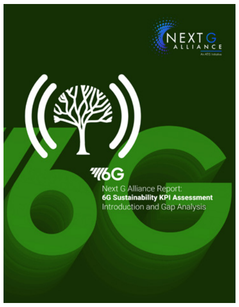
Digital World Experiences

Cost-Efficient Solutions



Sustainable 6G Connectivity — A Powerful Means of Doing Good

Next G Alliance Report: AI-Native Wireless Networks


NEXT G ALLIANCE
6G
Next G Alliance Report:
6G Sustainability KPI Assessment
Introduction and Gap Analysis

Next G Alliance Report: 6G Sustainability KPI Assessment Introduction and Gap Analysis

Next G Alliance Report: Multi-Sensory Extended Reality (XR) in 6G

Next G Alliance Report: Terminology for Frequency Ranges

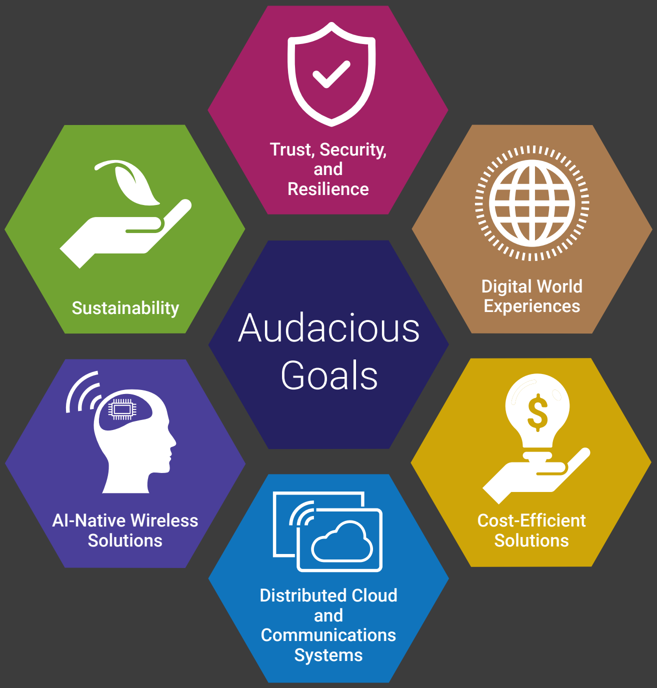
The Next G Alliance (NGA) is a bold new initiative to advance North American mobile technology leadership over the next decade through private sector-led efforts.

With a strong emphasis on technology commercialization, NGA's scope of activities encompasses the full lifecycle of research and development, manufacturing, standardization, and market readiness.

NGA's growing membership reflects support from stakeholders in academia, government, and industry.

This Perspective outlines the strategic importance of 6G and the imperatives that will shape North America's competitiveness, economy, and global leadership.

6G Market Development: A North American Perspective



Audacious Goals

Trust, Security, and Resilience

Digital World Experiences

Cost-Efficient Solutions

Distributed Cloud and Communications Systems

AI-Native Wireless Solutions

Sustainability

Building the foundation for
North American leadership in
6G and beyond.

nextgalliance.org

NEXT G
ALLIANCE
An ATIS Initiative