September 2018

Evolution to an Artificial Intelligence-Enabled Network

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 5G, cybersecurity, robocall mitigation, IoT, artificial intelligence-enabled networks, the all-IP transition, network functions virtualization, smart cities, emergency services, network evolution, 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), and a member of the Inter-American Telecommunication Commission (CITEL). For more information, visit www.atis.org.

Notice of Disclaimer and Limitation of Liability

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 OFMERCHANTABILITY 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.

NOTE - The user's attention is called to the possibility that compliance with this standard may require use of an invention covered by patent rights. By publication of this standard, no position is taken with respect to whether use of an invention covered by patent rights will be required, and if any such use is required no position is taken regarding the validity of this claim or any patent rights in connection therewith. Please refer to [http://www.atis.org/legal/patentinfo.asp] to determine if any statement has been filed by a patent holder indicating a willingness to grant a license either without compensation or on reasonable and non-discriminatory terms and conditions to applicants desiring to obtain a license.

Copyright Information

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

Contents

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have been active areas of research and development since the 1950s. Over time, enthusiasm has waxed and waned as AI has struggled to insert itself into mainstream commercial applications. However, in recent years, advances in processing power, the availability of large amounts of data, research advances and a healthy community of open source developers have enabled AI technologies to become an essential part of many industries. In recognition of this success, this report explores how AI and ML can be leveraged to address the pressing challenges facing the ICT industry today.

1.1 What is AI?

AI can enable more intelligent and efficient automation. However, AI is much more than automation. Automation covers a wide variety of examples. Most people would not label basic programmable automation as AI but do recognize AI as the automation of tasks or functions which otherwise require human intelligence to execute properly. Alternatively, many consider AI to be intelligence exhibited by machines or computational systems that perceive their environment and take actions to satisfy an intent. As such, AI can enable automation of many routine, well-defined tasks and activities. However, AI's ability to engage active learning while analyzing very large amounts data means it also has the ability to orchestrate innovative capabilities that were previously impossible.

AI is typically considered to be systems that perform some form of reasoning, planning or object management, using knowledge as well as perceived information that, in the past, required human intervention. In many cases, AI technology can detect subtle patterns in data that humans can't easily identify. Therefore, AI applications can provide expert assistance to people responsible for a specific task or function. Leveraging this attribute, AI applications span a wide range:

- **Assisted Intelligence –** Targeted/narrow expert systems that help people to perform tasks faster and more accurately.
- **Autonomous Intelligence –** Fully automated decision-making processes coupled with ML to perform a narrow task without human intervention while adapting to changing conditions.

Figure 1 illustrates how AI, ML, neural networks and deep learning concepts are related yet distinct.

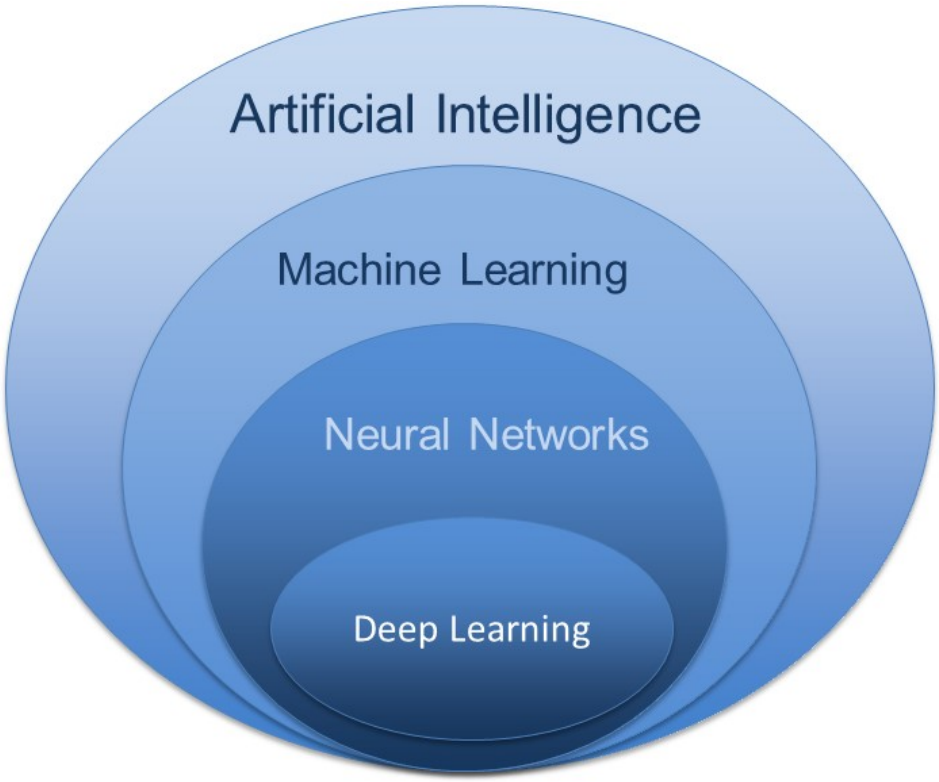

Figure 1 – Key Aspects of Artifical Intelligence

AI technology can be broadly categorized as all possible approaches for simulating intelligence, including:

- Rules-based approaches with an inference engine or semantic reasoner.
- Algorithms, dependency graphs and other expert system technologies.
- Neural networks.

In practical AI systems, these technologies are often supplemented by traditional software coding techniques to:

- Manage the AI system.
- Preprocess data used to drive the AI system.
- Implement output adapters to effectively use the decision or recommendation output.

Many forms of AI incorporate ML techniques to enable the AI systems to better adapt to a complex and potentially dynamic environment. ML involves training or data acquisition that can modify machine behavior and comes in many forms:

- Supervised learning occurs when the AI system is given training data sets where the desired output is known. The AI system then uses these data sets to learn to provide the desired output corresponding to the known input. With sufficient training, the system can then provide the correct output with inputs that differ from the training sets. Essentially, the AI system can interpolate the correct output with high probability given a properly constructed training set.
- Unsupervised learning is a type of ML where the system autonomously categorizes or describes the structure of "unlabeled" data. For example, unsupervised learning could be used to recognize patterns in the data to describe or categorize different states or conditions of a network. This information can then be used to identify anomalies. Supervised learning may be used to establish the initial state for unsupervised learning in AI systems.
- Reinforcement learning occurs when the system learns by interacting with its environment. For example, the system may receive rewards for performing correctly and penalties for performing incorrectly. These rewards are then used to enable ML, modifying future output predictions. Although in general reinforcement learning systems can be very complex, many network AI applications are well suited to this technology because networks currently provide a wealth of real time performance and quality metrics that can be used as feedback to the system.
- Online learning occurs when data becomes available in a sequential order and is used to update the model for future prediction in steps, as opposed to batch learning techniques, which operate on the entire training data set at once. Online learning is useful when the data set is very large, making it computationally infeasible to train over the entire dataset, or when the data are generated as a function of time. Both of these conditions are common to network data.

Neural networking is a specific class of AI ML systems that has been the focus of recent research advances. Use of various specialized algorithms and rules-based approaches often provide controllable deterministic results but have not been able to scale in cases where complex relationships exist which create very large numbers of potentially conflicting rules. Neural networks have shown promise in addressing complex data relationships.

A neural network is commonly comprised of columns (or layers) of nodes (representing artificial neurons). Each node receives a real-number signal from the outputs of the nodes in the previous layer. Each of these inputs has a weight that adjusts as learning proceeds, modifying the effect that input can have on that node's output. The output of each node is calculated by a non-linear function of the sum of its inputs. Artificial neurons (nodes) may be configured so that a signal is sent only when the aggregate signal crosses a certain that threshold. Different layers may perform different kinds of transformations on their inputs. Learning is accomplished by adjusting the weights and potentially the node thresholds at each layer in the neural network. Each layer or column of a neural network may represent a "layer of abstraction." Unlike many rules-based AI algorithms, AI neural networks often create outputs where humans cannot easily reason a clear justification.

Deep learning generally applies to large neural networks with thousands of hidden layers, wherein training occurs on each layer within the hidden nodes of a neural network. In recent years, deep learning neural networks have become the most promising approach to AI.

AI ML is also a key enabler for intent-based networking, where human administrators define the network's desired outcome in broad but descriptive terms. However, actual network management and operations are done using automated network orchestration and management systems that implement the desired intent of the expressed policies. Intent-based networking systems monitor, detect and react in real time to changing network conditions while automatically orchestrating new customer service deployments and configuration changes. With intent-based networking, it is often useful to think of AI as standing for automated intent rather than artificial intelligence.

1.2 Role of AI in Telecommunications

For network operators, there are two large classes of AI applications: those that run over the network for the benefit of an end user, and those that run within the network to optimize some aspect of network operation or management. Over-the-top (OTT) AI applications may place specific requirements on network performance and may be enhanced by specific network services. Network AI could be used to replace or enhance network planning, service deployment and management functions (typically operating with a long time constant). It also could be used in near-real time to dynamically optimize network performance based on rapidly changing traffic patterns.

A more detailed analysis of network AI application classes is given in the next section. Table 1 summarizes how AI systems are often fundamentally different from traditional software systems and as such, may require fundamentally new processes at each stage of the application lifecycle.

Traditional Software Systems	**AI Systems**
Excellent for applying well-defined requirements on structured data: • Produces deterministic results with efficiency and high capacity/performance. • Resulting actions can be traceable to code (and often to requirements).	Excellent for applying cognitive processing on unstructured information where problems may be ill defined and solutions probabilistic: • Errors will occur. Need mitigation strategies and clear assignment of responsibility. • Why AI produces a result may not be understood.
Purpose programmed to provide a specified function with well-defined features.	Use a specific platform/architecture with one or more AI technologies/libraries creating models tuned with supervised or unsupervised ML.
Well-known integration and testing tools and methods.	New processes to manage training and how the results might be utilized.

Although AI systems are excellent for applying cognitive[1] processing to complex systems, errors will occur. For network operators, very high levels of reliability and service availability are required because the financial consequences of network outages are often significant.

[1] Cognition is the process of acquiring knowledge through thoughts, experiences and senses. When applied to AI, cognitive processing includes the techniques used to simulate these human functions in electronic compute environments such as rules-based approaches with an inference engine or semantic reasoner, algorithms, dependency graphs or neural networks.

It is well known that AI algorithms may produce incorrect results in unexpected ways. Even when the AI decision process involves human engagement, decisions produced by AI systems may not be intuitively clear to people tasked to manage those systems.

A number of mitigation strategies exist to balance the need for high network performance and efficiency expected with AI-supported network systems while taking into consideration the possibility of occasional errors:

- Solutions can include redundant systems that may independently apply different AI solutions, or different combinations of AI and conventional algorithm-based automation and act only if outputs are within a certain level of agreement.
- AI systems can be segmented so that intermediate results can be verified. For example, the system may be decomposed into separate modules so that intermediate data-points are available for verification and traceability.
- Supporting AI systems can be created that are specifically designed to better explain the output of the core AI system that is providing a decision or recommendation.
- Output systems can be created to allow policy-based thresholds and limits, managed by people, to ensure the AI system cannot deviate too far from the norm. For example, the system can be designed to require human intervention/oversight if AI decisions cross certain thresholds (e.g., when changes to network parameters exceed a certain percentage or if the financial impact of a decision is very large).
- Operators can apply best practices from within the industry, as well as other industries (e.g., self-driving cars, AI stock trading automation and AI medical diagnostic applications).

Even with such risk mitigations in place, the introduction of AI technology to service provider networks is likely to start with more confined AI expert/recommendation systems and AI applications where people have some level of control over the output. As confidence builds, there likely will be many AI applications working autonomously to optimize and manage various network functions.

2. AI Network-Related Use Cases

AI use cases for wide-area network applications can be segmented into two major categories, as illustrated in figure 2: management and operations applications, and revenue-impacting applications.

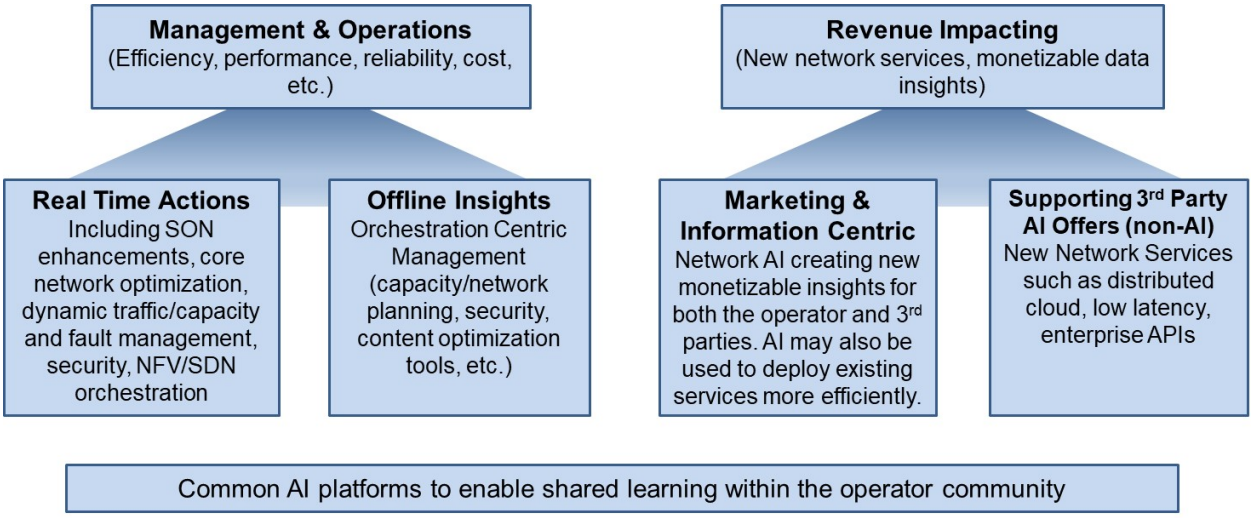

Figure 2 – Types of Network Oriented AI Applications

Management and operations use cases enable AI technologies to increase the network's efficiency, performance, availability, reliability and cost. These use cases can operate offline to create recommendations for improved capacity planning, security, service deployment, content optimization or performance of real-time actions that may optimize radio network performance, provide dynamic traffic management or real-time security detection and mitigation. They include automated service management capabilities providing traffic flow classification, fault prediction, WAN path optimization, capacity management, security, intelligent bandwidth-on-demand and service modification and restoration through the automated scaling of virtual network functions (VNFs), as well as transport level configuration of links and paths.

Revenue-impacting AI applications may enable new network services or help monetize existing network services and applications though monetizable data insights. For example, marketing and information-centric use cases can leverage network data and application data, along with opt-in customer preferences and usage, to provide an improved user experience. AI technologies can also be used in expert systems to

improve customer contact, support, sales activities and overall customer satisfaction. For example, AI techniques have the ability to perform service personalization [by?] tracking device behavior such as mobility patterns or the types of services the customer uses and predicting and applying network characteristics or presenting new services the user would enjoy. Additionally, new network services such as distributed cloud, low-latency access or new enterprise APIs can be leveraged by third-party providers interested in delivering their own AI solutions to network users. Finally, AI can be used to automate and orchestrate the deployment of customer service instances, adding new customers or customer configuration changes to the network initiated by a service order.

AI use cases generally span a wide range of complexity and control. Targeted AI applications can be used to optimize a specific aspect of a task and may operate as an aid or expert system providing insight to human managers making the final decision. Alternatively, complex AI use cases exist that may operate with vast amounts of data in hierarchical aggregates of AI modules that make real-time decisions independent of the human operator.

The use cases documented below tend to focus on the more sophisticated AI applications as these will more likely have impact on network standards and operations. These use cases are not intended to be exhaustive but rather a sampling of applications that may benefit from the unique value created by AI.

Many of the AI use cases detailed in this section utilize AI to better optimize aspects of network operations, configuration, security or content delivery in the face of a highly dynamic traffic environment. While satisfying the demands placed on next-generation networks in highly dynamic conditions is a tough challenge, the fact that these networks will be highly configurable and programmable will allow the challenge to be faced head-on. At any instant, there will be a set of demands, or stimuli, placed on the network as a result of the resources requested by the diverse applications used by various subscribers in the multitude of locations. In principle, there will be ways of configuring the network in a variety of dimensions to most optimally service those requirements while respecting any constraints and simultaneously optimizing the network's non-service characteristics. Distilling the problem down to its simplest terms, there are several components to the challenge:

- A set of stimuli in the form of the demands placed on the network by the subscribers and devices attempting to use it.

- A set of constraints that must be respected, including capabilities and capacities of network elements, and impairments to network infrastructure.
- A set of desired performance characteristics for the applications being used by the network's subscribers.
- A set of network parameters that must be configured in response to the stimuli, while respecting the constraints, to completely satisfy the desired performance characteristics.

AI models can be constructed to identify and diagnose adverse conditions and classify them in categories such as congestion, interference, loss of transport link, loss of network element or loss of coverage. AI models could also be used to make predictions about these states in the future, so the network could be prepared to mitigate the problem before it occurs. These are models that offer support to the engineers operating the network.

With more ambition, AI models can be conceived that directly link the network conditions in terms of the demands, constraints and impairments. Such models could combine these characteristics with the parameter configurations and predict the resulting performance and thus be employed to understand how well any set of configuration parameters would satisfy the desired performance characteristics. Such models would be valuable for predicting how suitable any proposed parameter configurations would be for satisfying the desired performance characteristics. Indeed, a model (or multiple models) could predict the specific set of parameter configurations that best satisfy the desired performance characteristics. If such models could be built and integrated with the network, it could form the heart of the decision-making in the network.

2.1 Network Anomaly Detection

Story Highlights

A large cellular service provider's network can generate several million performance measurements every minute. Some of these events can have common signatures during network outages. Data-powered ML can be applied to correlate these signatures with network anomalies.

This large volume of data can be processed by a network anomaly detection ML system using an AI data processing platform. ML algorithms can be applied to this data to correlate the signatures with network anomalies. ML algorithms can be used to correlate the data and determine patterns that can be presented visually using various ML applications. To support network anomaly detection, ML models can be created, on-boarded, trained, executed and shared using, for example, the Acumos AI platform and marketplace (see section 3.2).

Business Drivers

Finding a network anomaly in millions of events is like looking for a needle in a haystack. Network operations personnel spend countless hours searching for network anomalies and pinpointing the root cause. ML is ideally suited to analyze such a large amount of data to identify a small number of network anomalies. Benefits include reduced time and cost for resolving network anomalies.

Deployment Model

This use case needs the following environments:

- Development environment to create ML models using a rich set of ML toolkits.
- Sandbox environment to execute ML models using real-world production data.
- Production environment to run ML models and realize business benefits.
- An AI platform and marketplace such as Acumos to share and exchange ML models across teams.

This use case also assumes that network data are available and are of sufficiently large size over a large enough time period to detect network anomalies.

Actors

Key actors associated with this use case include:

- The network provider, whose organization and systems enable data collection and control.
- Business and individual customers making mobile phone calls on this service provider's LTE network.

High-Level Architectural Context

The management of voice over LTE (VoLTE) calls is a good example of this use case. VoLTE calls travel thru the radio access (RAN), mobile packet core and services core networks. These calls can fail if an outage hits any one of these three networks, as illustrated in figure 3. The service core network can report an outage signature via call detail records (CDRs). The proxy – call session control function (P-CSCF) is the first point of network entry for making VoLTE calls; hence the P-CSCF usually reports failed CDRs for the service core network. CDRs have this outage signature data, which can then be used with other network data to better assess service anomalies.

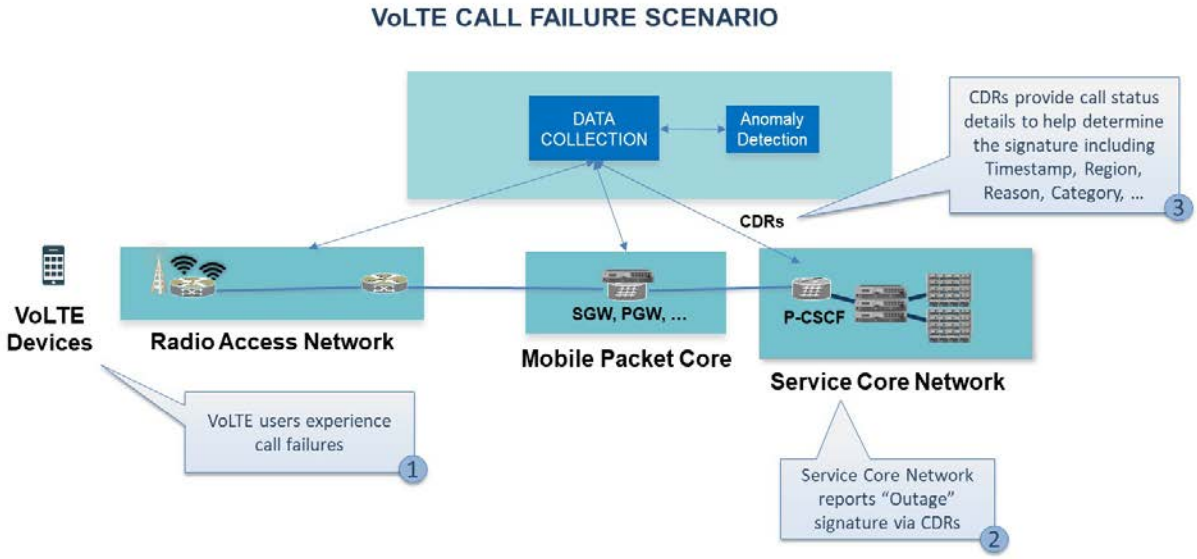

Figure 3 – VoLTE Call Failure Scenario

Related and Derivative Use Cases

Unsupervised ML technology is well suited for applications that detect pattern anomalies in network data. Proactive anomaly detection can be used to prevent or minimize the occurrence of potential future failures by performing preemptive maintenance based on the detected anomalies. In optical networks, for example, unsupervised ML has been applied to network data to identify trends that suggest aging and future failure of an optical port:

- First data are collected during normal operation of the optical network to train the ML model.
- Subsequently new incoming data are analyzed over some time period to determine their probability of occurrence under normal conditions and the overall trend.
- A risk factor is derived to indicate when preemptive action may be needed to avoid a future failure in the optical port.

In addition, anomaly detection can be utilized in security related use cases for applications ranging from the detection of unwanted text messages to caller ID spoofing and robocalling mitigation. In the case of unwanted text messages, network AI systems can be used to analyze typical text message data patterns and identify text messages that differ from normal texting behavior. Caller ID spoofing and robocalling are being addressed through the SHAKEN framework as documented in ATIS-1000074, *Signature-based Handling of Asserted information using toKENs (SHAKEN).* SHAKEN is an industry framework for managing the deployment of secure telephone identity technologies with the purpose of providing end-to-end cryptographic authentication and verification of the telephone identity and other information in an internet protocol (IP)-based service provider voice network. This specification includes a call validation treatment (CVT) application server function for applying anti-spoofing mitigation techniques once a signature is positively or negatively verified. The CVT can also provide information in its response that indicates how the verification results should be displayed to the called user. Unsupervised ML anomaly detection technology can be applied within the CVT function to better identify and mitigate telephone identity spoofing.

2.2 Network Security

Network security and associated cybersecurity threats represent a significant and ongoing challenge for the industry as a whole. In the US, interest in this area has been heightened by the May 11, 2017, Executive Order on *Strengthening the Cybersecurity of Federal Networks and Critical Infrastructure*. That order called for "resilience against botnets and other automated, distributed threats," directing the Secretary of Commerce, together with the Secretary of Homeland Security, to "lead an open and transparent process to identify and promote action by appropriate stakeholders" with the goal of "dramatically reducing threats perpetrated by automated and distributed attacks (e.g., botnets)." The resulting report highlights areas where network/infrastructure security can

be enhanced including recommendations to provide the filtering of traffic as it enters and exits a network and to deploy other anti-DDoS services.

The complex landscape of the ever-changing threat environment associated with network security provides the opportunity for AI-based solutions to out-perform existing network security capabilities. The following use case provides one example of such a deployment.

Story Highlights

Today's networks create a significant amount of data used to describe data traffic attributes and performance. Some of this data are currently collected for network analytics. Additional network data that may be indicative of security-related attacks could also be collected. This data could be supplemented by information provided by active network probes, data center customers and/or enterprise customers.

This large volume of data could then be processed by a network security AI system, which might then identify potential security issues and isolate the traffic associated with these issues to specific ingress (or egress) links. Once identified, it would be possible to use network functions virtualization (NFV) and software-defined networking (SDN) constructs to instantiate fine-grained anomaly detection and mitigation functions and re-route suspect traffic through these functions to increase the potential of successfully addressing security threats attacking or passing through the network. The AI system is not required to be 100 percent accurate as its purpose is to direct the application of a limited number of complex and highly specialized, real-time intensive security functions to potentially damaging traffic.

Business Drivers

Although security services products are commercially available in the marketplace, providers and consumers of network services and applications generally expect that network security is built into the products and services they use. In effect, network security is substantially a cost center and as such, providing sufficient levels of security at the lowest cost is a common goal. Implementing security recommendations often requires large capital and operational expenses in the deployment of specialized (and ever-changing) security capabilities across the network (e.g., the ability to filter traffic as it enters and exits a network and to deploy other anti-DDoS services at these points).

A system that first identifies traffic anomalies using existing network data and then instantiates specific detection and prevention tools targeted to the identified traffic could potentially increase network security at lower cost points. However, to be successful, the initial identification of traffic anomalies must be accurate enough to thwart a high-percentage of security attacks while limiting the application of expensive detection and mitigation functions to a reasonable cost.

Deployment Model

This use case assumes the existence of a robust and ubiquitous NFV/SDN infrastructure. It is further assumed that NFV data centers are widely deployed with sufficient SDN controls to allow traffic at all ingress/egress links to be redirected to an NFV data center where appropriate "security scrubbers" can be service-chained into the link.

This use case could be implemented with fixed (non-virtualized) security assets using other mechanisms to dynamically direct traffic through a service chain that includes these assets. From an AI perspective, it is not so important how the traffic is segregated and treated. Rather, the use case proposes that AI systems be used to identify the traffic and associated links, with the assumption that the AI system can dynamically learn new threat vectors and more accurately detect both existing and new attacks.

The use case also assumes that network data are available and enough to detect security anomalies and identify an ingress or egress link associated with the traffic. As such, the data collection system may:

- Collect data from all edge routers and gateway elements associated with the ingress and egress of traffic. These elements should have the ability to detect and report key packet header attributes that can be used to identify potential traffic anomalies.
- Instantiate probes in various places in the network to collect very specific security-related metrics and attributes.
- Collect data from other control and operations systems to enable the AI system to understand network topology and potentially the context of the data.
- Collect data from traffic terminated internal to the network (e.g., DNS traffic) and other identifying control information.
- Support interfaces with enterprises and data center customers of the network to enable these entities to provide insight into traffic attributes. These interfaces

may use existing standards or may require new standards and/or business agreements.

Actors

Key actors associated with this use case include:

- The network provider, whose organization and systems enable data collection and control.
- Enterprise and data center customers who may be originating or terminating suspect traffic and who may be able to alert the network provider about this risk.

High-Level Architectural Context

An example illustration of the use case in a service provider network environment is shown in figure 4.

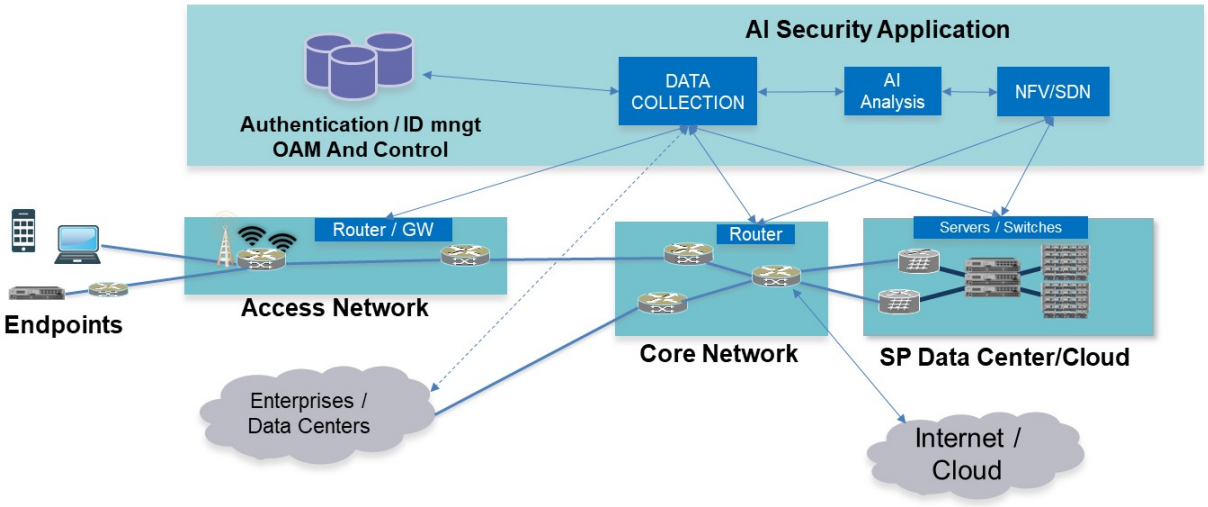

Figure 4 – A Network AI Security Application

In considering the deployment of this use case, several questions arise. For example:

- Do network elements have the necessary capabilities to detect and report attributes associated with malicious traffic?

- Do network operations and control systems provide enough information to correlate the detection of traffic anomalies to enable the traffic to be intercepted?
- Can all this data be sufficiently reduced to enable an AI system of reasonable capacity to operate on the data and provide results in sufficient time to limit damage associated with an attack?
- How can this AI system be trained to achieve sufficiently high accuracy?
- Do network APIs exist to enable the AI system to instantiate specialized detection and mitigation functions and route the suspect traffic through these functions?

Related and Derivative Use Cases

Related use cases may also engage active probes or other data sources to better detect and isolate security anomalies.

2.3 RAN Optimization

Moving into the era of 5G, network demands will be driven by new IoT applications, public safety organizations requiring highly reliable critical communications and the extra data volume from new high-bandwidth services and applications for consumers. These demands include increasingly dynamic utilization of the network across various dimensions including location, application, subscriber and time. Satisfying the demands placed on the network has always been a challenge for wireless network operators, and this challenge will magnify in the future.

RANs provide communication services in a complex and dynamic environment. The degree of control offered in the 5G radio network will be unprecedented. As in previous generations, there will be parameters that determine the characteristics of the physical radio interface. These include transmit power and the orientation of the antennas, some of which can be controlled programmatically. It also includes the configuration of the parameters controlling user equipment (UE) behavior, such as when and how to make and send measurements and how to behave in idle mode. Parameters controlling the behavior of the network, such as how the network responds to UE movement and how different network layers (carriers and radio technologies) are utilized together, can also be configured. The programmability of the network extends to where different network functions are located.

Therefore, optimization is vital for the network to reach its full potential. From early generations of mobile networks, optimization relied to a great extent on manual tuning. Support for automatic data recording and network tuning was introduced in 2G networks to reduce the manual effort. These early solutions were typically located in the central operation and maintenance systems. With the evolution of 4G and a more distributed architecture, the automatic functions have also been distributed into the RAN nodes. Examples of such automation functions are automatic neighbor relations and mobility robustness optimization, which have significantly increased the network efficiency. 3GPP has addressed many of these optimizations in its work on self-organizing networks (SONs).

The need for automation has been increasing continuously, and so has the development of autonomous functions. AI and ML offer the potential to achieve higher levels of automation and efficiency in terms of performance, coverage and capacity. The challenges lie in utilizing these computationally heavy procedures in near-real-time control loops. It is necessary to identify the data of most relevance to design effective solutions, which is why domain knowledge is still important.

Increased compute capabilities in the access network enable distributed AI/ML in smart network nodes to avoid extensive signaling to a centralized system and facilitate low-latency actions. Distributed AI also can simplify multi-vendor interoperability by enabling simple interfaces and can automatically adapt connection management to individual user service needs. Additionally, user consent and anonymization mechanisms facilitate the use of user data for AI-enabled network management to enhance overall user experience.

Story Highlights

The 3GPP RAN avails a great deal of performance and state information that can be used as a basis for RAN optimization. The various performance metrics available enable an operator to create a cost function specifically structured to optimize network performance consistent with operator goals (often a balance between coverage, capacity and overall fairness across the subscriber base).

Examples of potential RAN state information and performance metrics may include:

- RAN cell physical aspects such as topology, spectrum used (number of carriers in each band), power levels and cell neighbor relationships including small cell overlays within macro cells.
- Number of active UEs in a cell, along with some measure or indication of UE specific traffic characteristics (e.g., voice, data, SMS) and performance.
- UE signal, interference and noise ratios (or distribution of such) in the cell.
- Some measure of UE mobility.
- Measurements of recent UE and/or cell performance including call/session drop rates, handover performance, overall packet capacity and performance.
- Network transport and routing infrastructure metrics indicating packet loss, latency and throughput, along with information indicating what applications are running on what devices in which subscriber groups and network slices and at what locations.

Additionally, the network operator has control of several parameters that can be used to modify RAN network performance dynamically. Examples of potential controls include:

- Neighbor lists (list of acceptable cell neighbors appropriate for handover).
- Carrier power levels.
- Sub-tone power distribution including inter-cell interference management profiles.
- Ability to balance load across frequency carriers on various bands.
- Functional decompositions of network elements and associated deployment options on orchestrated virtualized platforms will dramatically increase the programmability and provide additional capabilities to apply AI-based optimizations.

Using network performance and state information, along with historical network data, an ML-based AI system can be used to optimize RAN performance, as defined by the operator, potentially using non-supervised real-time learning techniques that work to minimize a network-data-derived cost function. These AI functions may be centralized, distributed to the edge or both.

Business Drivers

It is well known that a large percentage of overall network cost exists within the “last mile” of access. That is, network cost is dominated within the access network at the

boundary between the network and the user device. By optimizing performance of the access network, particularly in highly dynamic and uncontrolled environments typical of the 3GPP RAN, operators can provide better service to subscribers at lower cost points.

Addressing a network's dynamics by configuring multitudes of parameters that operate in concert to achieve optimal performance requires intimate knowledge of the network. The flexibility in the configuration of the functional elements between the core and the radio means that in addition to configuration changes having wide geographic impact, the impact will also span the traditional silos of core, transport and RAN. Grasping this complexity will place enormous demands on the humans operating the network to understand their networks in extraordinary breadth and depth. Finding enough people with those skills will be enormously costly.

Another business driver is revenue growth. The industry aspires to deliver network slicing, and this is being defined and standardized in 3GPP, GSMA and other standards bodies. This technology will allow different services to be hosted on the same physical infrastructure but with logical separation, enabling each logical service to have its own heterogeneous quality-of-service (QoS) requirements. For example, an IoT layer may be required to achieve deep indoor coverage and long battery life across a very high density of devices, but with small data volumes. These are very different characteristics from a public safety network that must be highly available and have low latency with very dynamic utilization. These both are different from consumers using smartphones who typically need high volumes of data and the ability to support a variety of commercial applications.

Network slicing allows these services to be delivered simultaneously on the same physical infrastructure and with logical separation, where each slice meets its service level agreements (SLAs) and there is prioritization in cases where QoS requirements conflict. Network slicing facilitates a market in connectivity with SLA guarantees, which can become a new revenue source for operators, underpinning the business case. This depends on AI because it is another dimension in the complexity of the next-generation network that must be managed, making manual configuration even less practical.

Deployment Model

AI and ML technology can be deployed in the RAN at a variety of layers. As figure 5 illustrates, RAN automation can take many forms and operate at many levels in the network at different time scales.

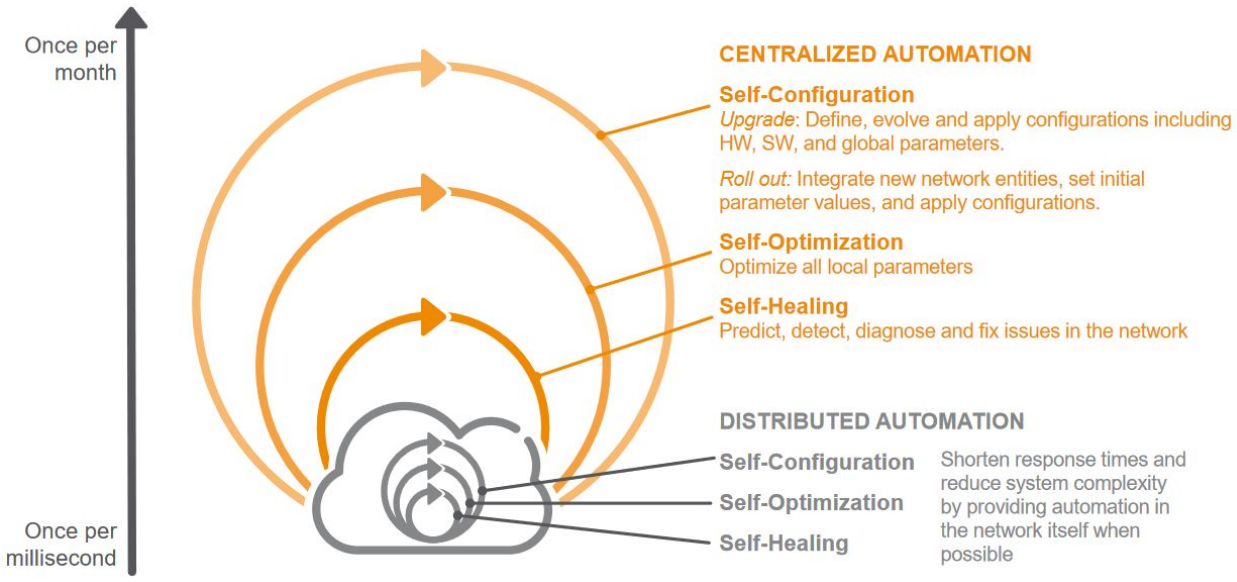

Figure 5 – RAN Automation Control Loops

Automation can include activities such as self-configuration, self-optimization and self-healing. Automation executed in a centralized mode covering a broad physical area typically operates at longer time scales. Automation distributed out to the network edge can operate very quickly, sometimes at millisecond speeds. Due to the different time scales, it is possible to run both automation types simultaneously in the network.

An AI system operating as a component of network automation will depend on models of the network. These models might be partially trained in isolation from the actual network in which they are deployed. However, it is anticipated that there also will be a significant part of the training that is empirical and based on the actual network concerned so that the network's nuances and specific characteristics can be captured. Moreover, these models must be capable of representing situations out of the network's normal operational envelope. This will allow them to effectively handle new types of anomalies on demand, previously unseen impairments and other unusual events.

Actors

AI/ML-based RAN automation may include the following actors:

The network operator decides the commercial objectives for achieving performance in a dynamic and non-uniform environment. It sets policy that translates into the goals for the automation systems that control the network and how it operates in response to demand, impairments and other stimuli. Specific network operator subsystems impacted may include:

- Network operations systems, which may provide both historical and real-time operational state and performance metrics.
- Network operator policies, which may guide or limit automation activities.
- Network APIs, which enable automated actions.

Another type of actor is the entities procuring services from the network, especially services with SLAs such as network slices. This includes the subscriber-owned UE, which may provide additional information related to the state and performance of the RAN from a user's perspective.

High-Level Architectural Context

Figure 6 illustrates 3GPP SON configurations for centralized, distributed and hybrid architectures. In all cases, automation may be operating in a RAN context that may include macro cells and small cells operating in a variety of frequency bands using licensed (3G, 4G and 5G), unlicensed or shared air interface standards.

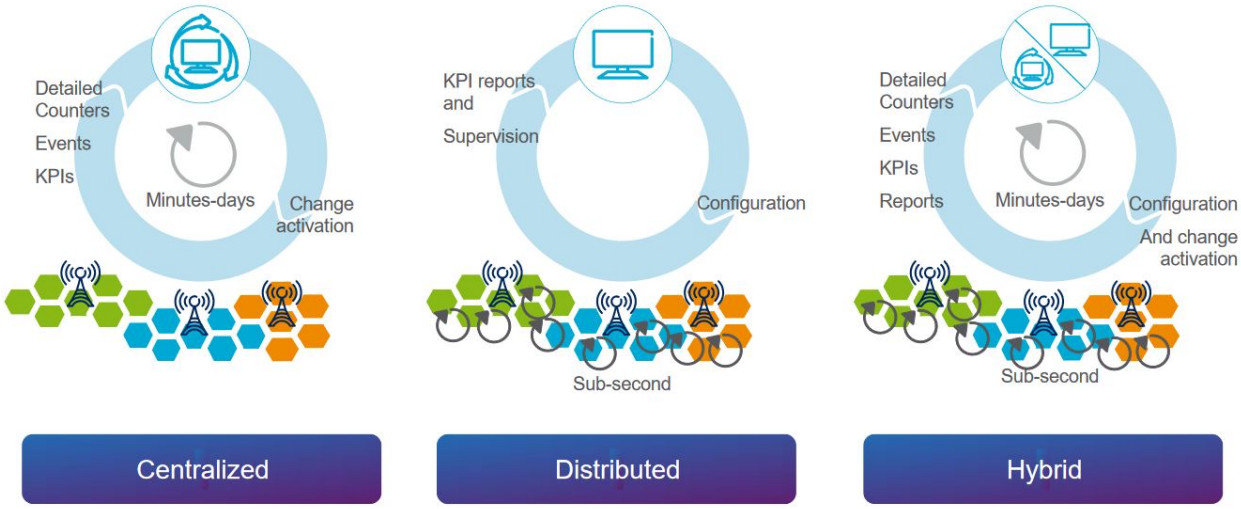

Figure 6 – SON Configuration Options

AI/ML technologies can be used in all approaches, both centralized and distributed, to provide enhanced automation results.

Parts of the functionality of the RAN have been decomposed by 3GPP into a centralized unit and a distributed unit. In some cases, these will be VNFs with a choice of where these are instantiated. Various factors influence the choice of where to physically locate these functions, typically driven by constraints on latency and jitter along with transmission costs. For example, these factors include the need for extreme capacity delivered by massive MIMO and the need to overcome inter-cell interference through coordination of transmissions. How these nodes are associated and the routing between them is also a configurable decision that can not only allow the latency and jitter required by the services to be achieved but can also enhance the resilience of the network to impairments. Some aspects of these networks will be controlled by vendor-specific parameters, such as how the schedulers work, when and how different network layers are used and when handovers are performed. However, some degree of configuration will be exposed via APIs, orchestrators or similar entities.

2.4 Dynamic Traffic and Capacity Management

Story Highlights

Abundant capacity and performance data are available in-service provider networks. Typically, a network performance monitoring system might collect metrics such as latency, packet loss and throughput on a per-link or per-path basis for every node and link in the network every 10-15 minutes. For large networks, this represents a daunting amount of data to be analyzed. Additionally, optimal network configuration and routing may be time variant to account for time-of-day and day-of-week traffic patterns.

Given the availability of SDN-controlled traffic routing, advanced transport network configuration capabilities and other advanced routing techniques, networks can implement dynamic network configuration changes without physical or manual intervention. By applying AI techniques to network performance data, an AI system can dynamically and automatically (or manually) enable significant changes in network configuration to optimize traffic flow and thus minimize network cost while increasing network performance and efficiency.

Business Drivers

Network topologies and configurations have traditionally been costly and time consuming to create, optimize and deploy. While modelling tools are available, they are only effective if a network conforms to the tool's assumptions. Additionally, traffic patterns can change dynamically from hour to hour and day to day so that creating an "optimal" topology/configuration becomes a quickly moving target. By addressing these issues with automation, the network can be made to perform better at lower cost points, all of which benefits the network service provider and its customers.

Deployment Model

This use case can be deployed in access and metro networks, as well as the core backbone in service provider networks. Deployments assume the existence of the necessary APIs and control functions to control traffic routing, establish new paths/links and reconfigure transport bandwidth. The use case also assumes that network data are available and enough to manage traffic in the network. As such, the data collection system may:

- Collect data from all edge routers and gateway elements associated with the ingress and egress of traffic. These elements should have the ability to detect and report key traffic performance and capacity metrics such as packet loss, throughput and latency.
- Leverage active probes that may be dynamically instantiated in network data centers as needed to monitor critical assets.
- Collect data from other control and operations systems to enable the AI system to understand network topology and potentially the context of the data.
- Collect data from traffic terminated internal to the network (e.g., DNS traffic) and other identifying control information.

Architectural Context

Figure 7 illustrates the example use case of a service provider network environment.

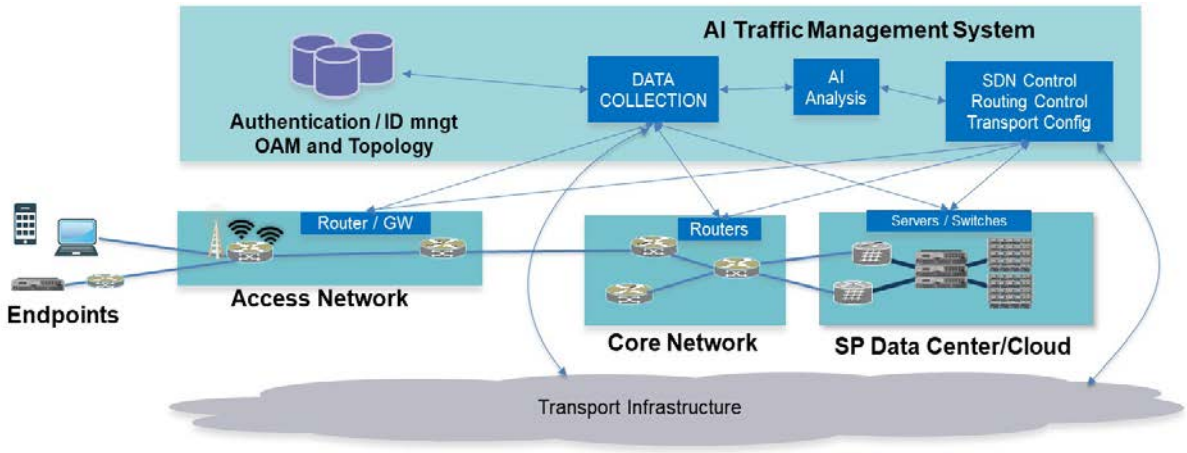

Figure 7 –AI Based Network Traffic Management

In addition to the need for good data collection that can sufficiently assess network performance on a granular basis, the AI system must include a robust backend system that can implement, in an automated fashion, the network configuration recommendations proposed by the AI system. Initially, AI recommendations can be forwarded to a control center where people can make the final decision about configuration changes and run the appropriate scripts to implement the changes. A real concern for this use case is the possibility that network configuration recommendations could negatively impact network performance or even take the network down. Additionally, it's possible that an AI recommendation might be good for a majority of

users but negatively impact critical systems such as emergency services or SLA-guaranteed enterprise services. As such, until robust control systems exist to properly manage configuration changes in a fully automated fashion, this use case is likely to have significant human intervention to implement the recommendations.

Related and Derivative Use Cases

AI-based network optimization use cases may also utilize the general architecture an operational structure of this use case. Network optimization has been addressed by ATIS in past reports including:

- *Network Optimization Focus Group (NetOp-FG) Assessment and Recommendations*, September 2011, ATIS-I-0000023.
- *Emerging Opportunities for Leveraging Network Intelligence*, October 2014, ATIS-I-0000046.
- *ATIS Big Data Analytics Focus Group: BDA Data Value Chain Reference Model & Use Cases*, October 2013, ATIS-I-0000043.

Generally, these reports explored a wide range of network optimization use cases, identifying the required service capabilities, various implementation options, regulatory considerations and areas recommended for further standards development. The specific network optimization use cases addressed in these reports include:

- Congestion-aware fairness.
- Subscriber-application-aware network optimization.
- Network-aware scheduling of content.
- User rate plans.
- Reasonable network protection and management.
- Load- and policy-aware multi-RAN selection.
- Optimizing sse of wireless non-bearer resources.
- Network-wide application detection and usage support.
- Prioritization of traffic for regulatory and enterprise services.
- Personalized broadband.
- Public safety spectrum sharing.
- Enhanced fault resolution.
- Outage alerting, avoidance and reporting.
- Network-wide intrusion detection.

- NFV automated network growth/degrowth.
- Inter-datacenter congestion mitigation.
- RAN-aware time-shifted content delivery.
- Dynamically inspect traffic and predict network performance.

In most cases, these use cases utilize a generalized architectural framework wherein network data are collected and aggregated via a variety of data sources and data analytics systems. The data are then utilized by an optimization application. Finally, the recommendations from the application are used to implement configuration changes in the network. AI technologies can be applied within this architecture in support of the optimization application. In the past, these optimization applications were implemented in software and programmed to execute a specific optimization algorithm. Going forward, AI systems may replace the bulk of the software to provide a higher level of "intelligence" and ML to better optimize the aspect of interest

Network Resiliency and Self-Healing

An important derivative use case related to dynamic traffic and capacity management is the ability to dynamically respond to network failures and anomalies to provide enhanced network resiliency and self-healing. The basic ingredients for this derivative use case include the ability to apply AI analysis technology to the mass of network data along with:

- Network topology information.
- Network orchestration functions to manage SDN-controlled traffic routing, advanced transport network configuration capabilities and other advanced routing techniques.

Networks with these capabilities possess the foundation to implement dynamic network configuration changes without physical or manual intervention. This would enable the AI system to not only dynamically manage network traffic capacity, but also utilize outage information in deploying network changes to enable self-healing and resiliency capabilities.

2.5 AI and Orchestrated Management

Orchestrated management refers to operations that require coordination of activities among several network or OSS resources. Examples of orchestration include managing

customer orders, implementation of data centers, complex expansion and reconfiguration processes. These overall processes have many sub-processes (components), which may succeed, fail and/or yield different outcomes. Traditionally, the overall process is implemented using a complex workflow that addresses the triggering of various activities, exception handling and recovery.

Story Highlights

The complexity of such processes typically increases on a higher order (e.g., n^2) as the number of components grow. After several generations, managing such complexity becomes a hindrance in offering new services, modifications of existing services and taking advantage of changes in technology and market demand. AI-based approaches can alleviate these issues by breaking the complex flows into a number of smaller yet more intelligent entities that, with assistance from ML and rule-based programming, can significantly simplify the evolution of these flows. The two key elements of AI and ML are:

- Entities must be independent (each with their own rules). Entities must allow ML to influence their "decisions."

Business Drivers

One of the major barriers to new service introduction is complexity of existing orchestration processes. New services can require enhancement of operations processes, and the complexity of such operations can add cost and delay. Lack of experience with new services can result in initial processes that are sub-optimal and must be enhanced, causing further delay and cost.

Deployment Model

In this deployment model, the complex processes are broken down into smaller components (sub-processes). A proper breakdown is key. These partitions must be independent, with clear triggers (conditions that activate them) and outcomes (which overall state variables they modify).

The triggers are based on overall "state variables" visible to (but not necessarily used by) other modules. The independence requirement is also key in that these modules must be able to run from start to finish without waiting for another process to complete (albeit it can trigger other processes in the beginning, duration and the end but not be

dependent on them for its own completion). Each process may also be invoked several times based on the outcome of other steps and go through several iterations until it completes.

The AI aspect arises from the fact that these independent processes each behave according to their own rules, which may be written by different developers and/or teams. The interaction of these rules can yield solutions that no one has perhaps foreseen. Also, as new process components are added, these will bring their own rules and can interact with other components in novel ways, as well.

This can be further enhanced by introduction of ML. As the various solutions are produced by the interaction of these independent modules, resulting key performance indicators (KPIs) can be fed back to an ML infrastructure to provide additional state variables which can be used in triggers and process component operations. This set of additional information can influence each component to make decisions more conducive to previous successes, trigger more suitable components or both.

Architectural Context

Typical implementations include an environment where state variables can be communicated among all process components and constantly evaluated to ensure proper triggering of process components. It is critical that this infrastructure is distributed and efficient in its operations as constant evaluation of activation conditions (triggers) requires a near-real-time processing and efficient handling of race conditions (e.g., update of the same variable by multiple processes).

The ML infrastructure is also an enhancement where available. This element requires data collection, processing, model development and a connection into the overall state of the operation where it can inject/modify state variables to affect the execution of the process components. It further must have a learning environment where the outcomes are properly modeled so they can be assessed and fed back into the system.

2.6 AI-Based Subscriber Insights

AI-based platforms can be used to collect, store and analyze data from across an operator's entire customer base to achieve real-time behavioral insights. Network information is aggregated, anonymized and/or combined with consent-based solutions, enabling operators to leverage a wealth of information about their customers' behaviors,

preferences and movements to mutual advantage. This information could help cities better manage their infrastructure, help businesses reach customers more likely to have interest in the products and services being offered or help health officials track diseases. More specific examples include:

- The creation of demographic information about how users arrived at a specific event/location and/or the kinds of applications they use once they arrive. These insights provide more efficient and effective event planning and management functions.
- The creation of location- and time-dependent data traffic profiles that city planners may use to better manage vehicular traffic flows, as well as by local businesses to facilitate user awareness of local products and services.
- Enhanced advertising effectiveness by better matching the types of ads delivered to user preferences and observed behaviors.
- The creation of advanced analytics analysis for systems of connected machines that can be used for anomaly detection, diagnostics, forecasting and other optimizations.

In all of these cases, we assume that either the user privacy policy enables the sharing of anonymous and aggregated subscriber data with outside parties and/or explicit consent to use data has been acquired.

Story Highlights

In many cases, a significant network data collection and analytics system may already be in place to support a variety of network traffic management, capacity planning and optimization use cases. This data infrastructure, used in combination with AI-based application functions, can create a variety of offline insights. This information may be valuable to cities and governmental agencies, health care organizations and businesses to better manage infrastructure, operations or support more effective product and service marketing.

Business Drivers

Offline AI insights derived from network data are valuable in a wide variety of contexts. As such, they enable the network operator to better monetize the wealth of network data available within its network.

Deployment Model

This use case also assumes that network data are available to create the necessary traffic density maps and associations needed for the specific AI application. Data may be collected from network elements such as edge routers and gateway elements, control and operations systems, as well as application specific functions. This data enables the AI system to understand network topology and potentially the context of the data. In some cases, specific subscriber information may be used in either aggregated or anonymized form or based on explicit subscriber consent.

Architectural Context

Figure 8 illustrates an example use case in a service provider network environment.

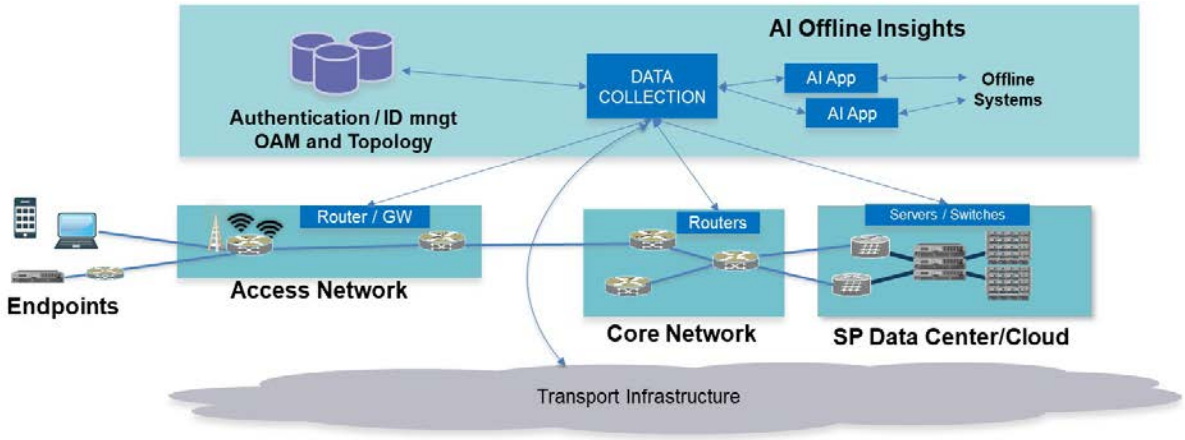

Figure 8 – Network Architecture for AI Offline Insights

In many cases, the data collection and analysis function may be shared across many different applications, both in real time and offline.

2.7 AI-Assisted Customer Support and Sales

Expert systems designed to assist in areas such as customer support, sales and network troubleshooting have been available for many years with varying degrees of success. However, as AI technology advances, the effectiveness and applicability of these AI applications also increases.

In performing support functions, people tend to use:

- Experience/history, which is information and knowledge based on past events.
- Explicit rules and heuristics, which are simple, efficient, learned or hard-coded rules or processes often based on experience.
- Focused system knowledge to solve problems.

Successful AI-based expert systems will use ML techniques to reproduce human problem solving as applied to these areas. These systems can be used to increase the accuracy, speed and effectiveness of support and sales activities by providing people executing the support/sales activities with expert information and recommendations.

Story Highlights

ML AI-enabled expert systems utilizing network data, subscriber data and domain specific rules and knowledge can assist personnel performing sales, support and troubleshooting tasks to increase speed and effectiveness in addressing the specific task. These systems also can potentially anticipate when customers are experiencing a service issue, giving the service provider an opportunity to address the problem prior to customer contact.

Business Drivers

These systems increase productivity by enabling faster and more effective problem resolution at lower costs and with increased end-user satisfaction.

Deployment Model

These AI-enabled expert systems can be deployed in isolated task specific systems, as well as in broader application/network areas on a national or regional basis.

Actors

Key actors associated with this use include:

- Support and sales personnel utilizing the AI-enabled expert system.
- End user/customer.

2.8 AI-Based Content Processing and Management

Content curation is an excellent example of how AI-based content processing and management systems can be used as part of a network service. Content curation is the act of sorting through large amounts of content, both web-based and from content libraries, to provide a content index to a user in a meaningful and organized way. The process can include recognizing user-relevant patterns associated with content and then sorting the content into specific themes for publication. Using advanced AI-based content analysis systems, customers can be provided with a personalized content and video experience.

Story Highlights

Using an AI-based content processing and management system, a service provider can search a wide range of web-based content and subscription-based content libraries, to sort and prioritize content personalized for a specific user. Content recommendations can then be presented to the user for subsequent viewing.

Business Drivers

Content curation systems can provide a value-added feature to existing access or content subscriptions to provide differentiation and increased customer satisfaction. These systems can also drive traffic growth through greater user engagement.

Deployment Models

This use case would generally be deployed as a network application with access to broad web-based resources and other content libraries. Access to a user profile and viewing history (with user consent) would be critical for accurate personalized recommendations.

Actors

Key actors associated with this use include:

- Content users that have opted in to receive personalized recommendations.
- Service provider(s) providing content recommendations.
- Content providers.

3. AI Architectures and Technologies

3.1 Network Architecture Aspects of AI

Network-based AI systems tend to share a common high-level fundamental architecture that includes three major components:

- Data collection and feature extraction. This process is about getting the right data in the right timeframe with the ability to derive the right data features to enable efficient and accurate cognitive processing in the AI core.
- An AI core using one or more cognitive analysis technologies such as a deep learning neural network. This step may involve making choices regarding the use of an AI framework and associated library and tool set to best match the given task. Given a framework and library, a model will need to be created and trained to respond to the given environment.
- Output adaptors and/or formaters that take the raw AI output and actualize the goal of the specific AI function or module. In some cases, this is a matter of applying appropriate data visualization technologies so that human operators and administrators can better understand or use the result. In other cases, the output may be used to effect automated real-time changes in the network. For this situation, appropriate APIs may be required to apply the necessary controls. Additionally, orchestrators and policies may be used to implement the end result on a given topology constrained by these policies.

Figure 9 illustrates these steps at a high level and is followed by an in-depth discussion of each step and how these components may fit in existing networks.

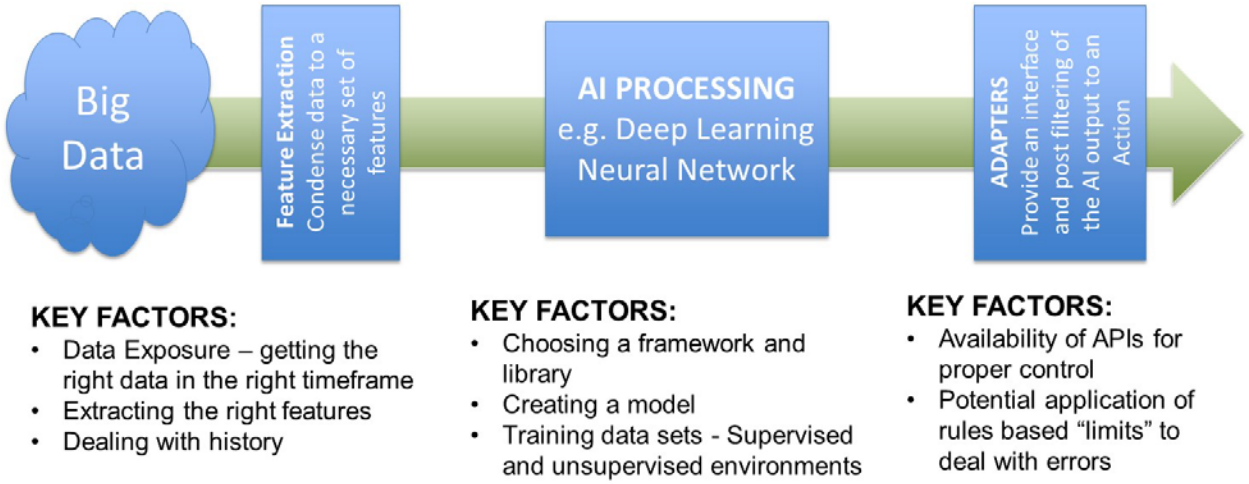

Figure 9 – High Level Structure for Network Based AI Applications

Data Collection

Most service provider networks already collect and store large volumes of data associated with network operations. This data can be categorized along four basic dimensions; traffic-based attributes, network/subscriber state, topology/location and time/history.

Traffic-Based Attributes:

Packet-traffic-based attributes include raw packet router/switch or link-specific metrics such as throughput, packet loss, latency, packet length (short or long), burstiness and queue fill. Although more information may be available through deeper packet header inspection, the widespread use of TLS/HTTPS on transport flows limits the utility of this approach. In addition, it is often impossible to implement deeper inspection at wire speeds on modern high speed links, switchs and routers.

Nevertheless, application-specific attributes may also be available for network flows. These attributes can often be aggregated by application class and may be collected directly from the application (at either the client or the server) or at an application gateway. Probes may also be used to simulate application traffic. Examples of application specific attributes may include:

- Adaptive bit rate/video-streaming-related metrics, and measures of video display quality.

- File transfer attributes.
- Interactive communication attributes including mobile network dropped calls, and voice or video quality.

Traffic may also be categorized based on the aggregate of application classes used by a subscriber or group of subscribers into subscriber classes. Often, subscriber traffic can be identified by address markings or associations with links using various network segmentation mechanisms (e.g., VPNs). In some cases, subscriber classes can be mapped to SLA parameters, for example, associated with an enterprise service.

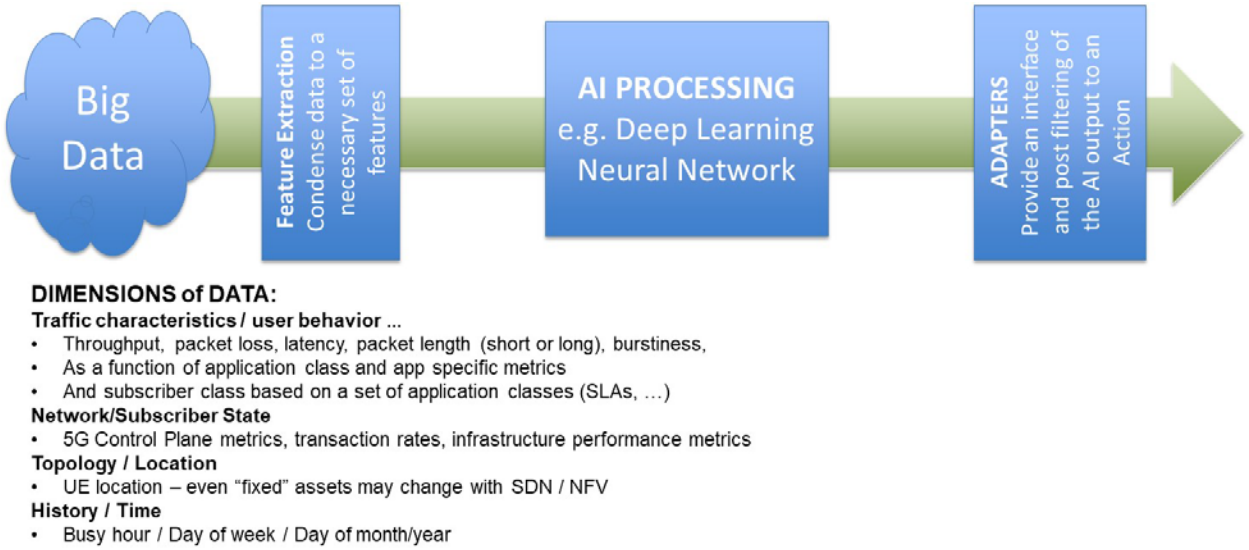

Figure 10 – Dimensions of Data for Network Based AI

Network/Subscriber State

The network is rich with data associated with the performance of the network infrastructure, as well as subscriber-specific state that the network uses to manage each subscriber session. This state information ranges from data center compute and storage metrics to 5G control plane properties associated with subscriber sessions.

Topology/Location

Network topology is a very important class of data for most network-related use cases. Although much of the network topology is fixed as [a?] function of fixed assets, network operators are constantly adding capacity, resulting in topology changes. In addition, the

introduction of SDN and other automated configuration and routing tools can effectively modify network topology in near real time. For mobile networks, the UE's location is an ever-changing variable, and the widespread use of nomadic devices (e.g., Wi-Fi tablets) adds to this dynamic transport environment.

History/Time

At each point in the network topology, the packet-related traffic attributes will vary in time, creating a history of activity that can be useful in predicting and managing traffic flows. Traffic is often time-dependent as evident with the existance of busy hour and busy day metrics.

The AI Congnitive Processing

Network AI uses cases are generally shown in isolation, with only one AI processing core (e.g., neural network instance). But in practice, many different AI processing modules, each potentially utilizing different AI technologies, may exist, working in concert.

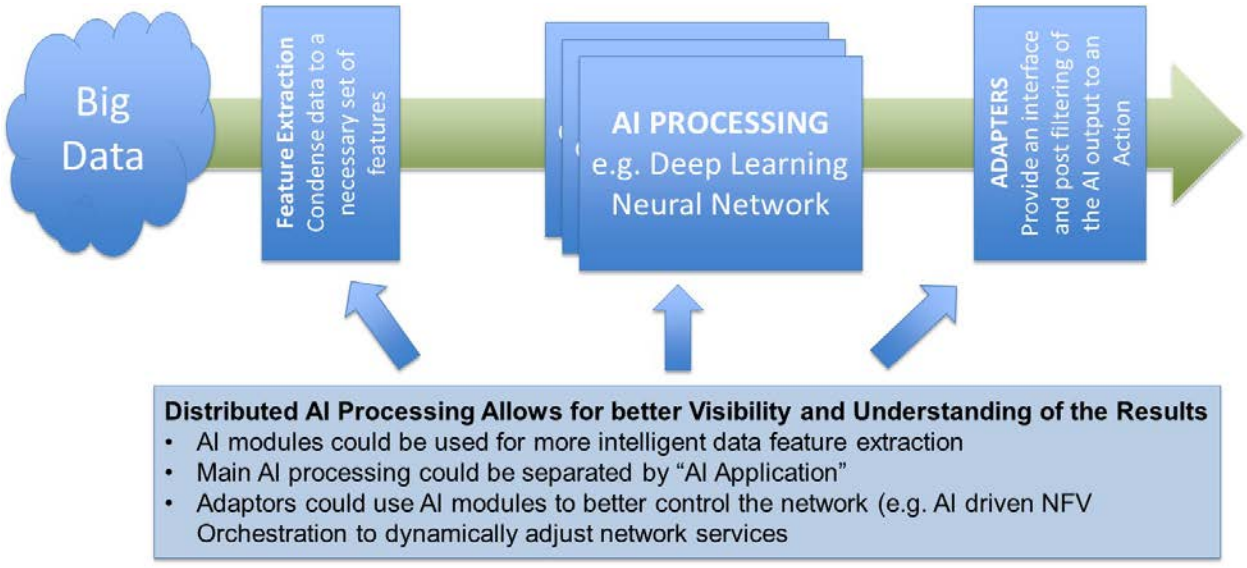

Figure 11 – Distributed AI Processing in Network Applications

For example, even within a single AI use case, AI processing could be used in each step of the process for specific functions:

- AI could be used in the feature extraction process to better compress large amounts of time-dependent data across a wide network topology. This would create more descriptive and efficient features to input into one or more AI core processing functions supporting one or more different use cases. For instance, a neural network utilizing unsupervised learning capabilities could be used to uncover recurring patterns in network data that may not otherwise be clear.
- Even within the AI core processing function, a number of different AI and ML technologies may be used together to increase the accuracy and better clarify the rationale of an AI recommendation. In this case, a specific AI task could be undertaken by many different AI modules each looking at different data/features using different models or technologies/algorithms so that the final recommendation can be associated with some probability of accuracy. Additionally, some of the AI core processing functions could be used to create context and rationale for the final recommendation, providing a historical record for later post-mortem analysis in the case of failures. Given the wealth of real-time network performance and quality metrics available in most networks, reinforcement learning systems can often be utilized to dynamically adapt to changing network and traffic conditions.
- When the AI use case involves automated actions, these actions, as they are applied to the network, may invoke configuration or service management orchestration functions. These functions may use AI technologies to implement the complex sequence of steps required to robustly deploy network changes. In many cases, rules-based approaches provide a means to execute recommended actions with some level of deterministic control to help ensure overall system integrity.

Output Adaptors and/or Formatters

AI recommendations from core AI processing functions typically require additional formatting, presentation or orchestration functions to achieve the desired output. As noted above, network/service orchestration functions may require AI functions to robustly deploy network changes. In addition, rules engines may be employed to enforce network policies that may be used to constrain any AI initiated actions to satisfy customer SLAs, as well as regulatory requirements that may be impacted by automated network changes. For example, network facilities supporting critical infrastructure (e.g., related to IoT access) emergency services or in support of public safety need to have priority regardless of traffic metrics.

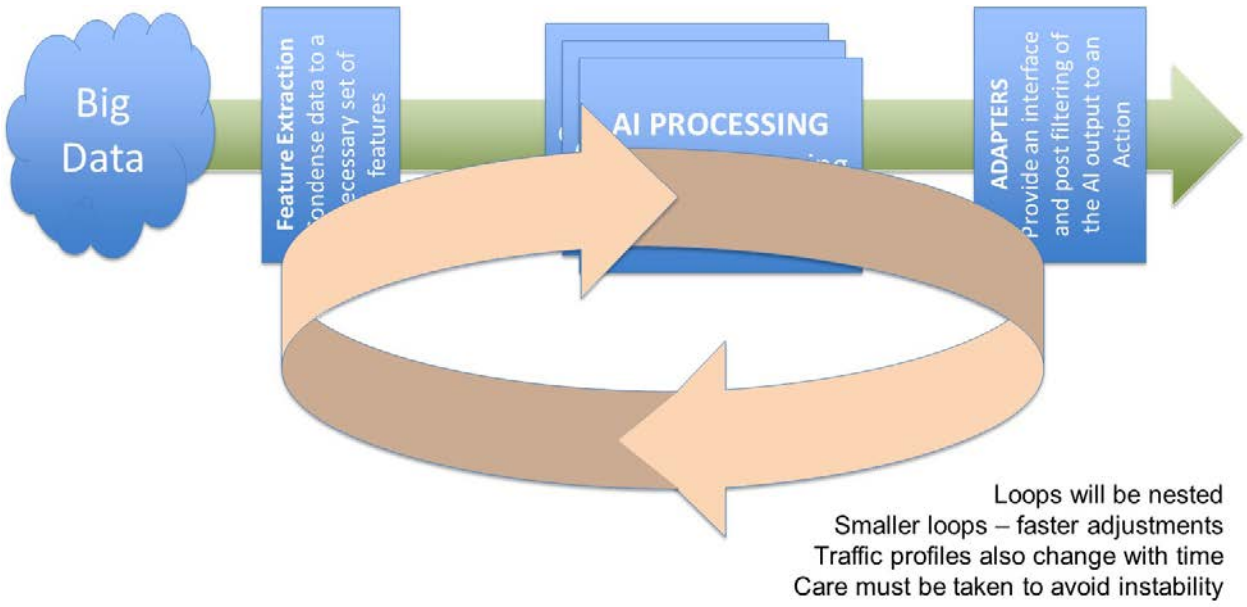

Figure 12 – Control Loops in AI Applications

In a network with multiple AI use cases and/or multiple AI processing functions within a use case, care must be taken to ensure network stability. AI network applications are often deployed in closed-loop systems. The wealth of network performance data available often means that AI functions can be deployed using reinforcement learning techniques to dynamically optimize system performance. However, multiple feedback loops can interfere with each other, causing instability, if care is not taken to ensure that the loop-time constants for nested loops are properly managed. Generally speaking, the inner/shorter loops can operate at faster speeds, while longer outer loops need to operate at longer time constants to ensure the network has stabilized before making new changes.

3.2 AI Technology Development and Management

The inherent complexity of AI solutions has led to the creation of many different open source and commercial software libraries and frameworks to enable AI developers to more efficiently build solutions. These frameworks provide APIs and software libraries to allow developers to leverage common technology functions. Some of the frameworks also include development process tools to better manage AI projects.

In network operator-based telecommunications systems, AI/ML systems place unique management requirements on the existing network management and operations environment. To address these challenges, AT&T developed and open-sourced a platform called Acumos. This platform employs the open source collaboration model to achieve flexibility and create a highly adaptable industry-wide framework for building, training, integrating and deploying ML solutions.

To simplify the development process and make it understandable to a wide audience of developers, Acumos breaks the development flow into four distinct steps:

- First, models are onboarded to an Acumos-compliant platform and packaged as distinct microservices with a component blueprint describing the microservice API and dependencies.
- Second, the model is packaged into a training application that can be deployed to an appropriate training environment from which data can be acquired and cached for later retraining, if needed.
- Third, a reference to the trained model, called a predictor, is published into a catalog that can be shared across a community where other developers can find it, discuss it, review it and create a full solution by using the predictor and "chaining" to other components employing the predictor. This enables them to make decisions while providing many of the more conventional capabilities that are required to act on them.
- Finally, the entire solution is packed into a Docker container that can be deployed to an appropriate runtime environment where it can be executed.

Some of the components in this packaged solution are used to access data and implement functionality. Others can be used to transfer new data records to the Acumos platform where they can be added to the cached datasets and used for additional training in a continuous learning process.

Figure 13 illustrates these four distinct stages of the development and deployment process.

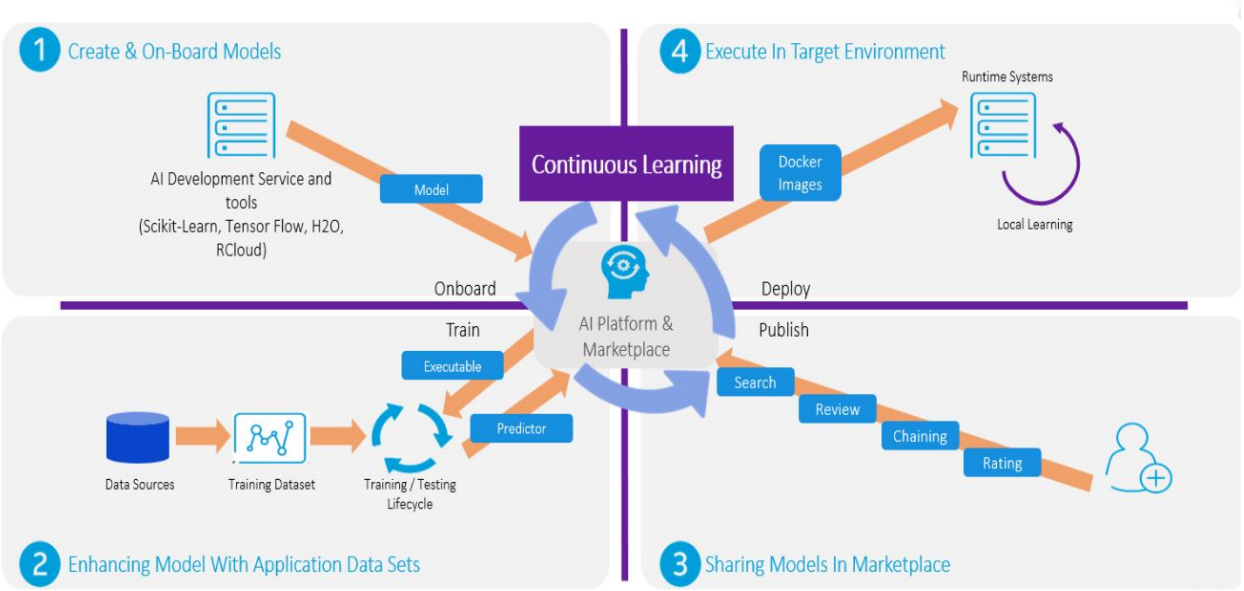

Figure 13 – Acumos AI Management Architecture

The reasons for approaching the problem in this way are to separate the very specialized functions of model design and data management from the complexities of service development and application lifecycle. Keeping these aspects of AI development distinct makes the process quicker, more reliable and open to a much larger community of users.

Step 1 – Creating and Onboarding New AI Models

As illustrated in the upper left quadrant of figure 13, Acumos does not include a specialized platform for developing AI solutions aimed at data scientists. There are already many excellent development tools for building neural nets, classifiers, clustering algorithms and other types of AI components. But these tools do not make it easy to integrate with other components. Either the tools are tied to a particular execution platform, such as a specific cloud service, or they are very specialized to the needs of the data scientist and difficult for the average software developer to use. Each tool is implemented around some language and a specific set of compatible libraries. All these things narrow the audience for any tool and limit access to previous work by requiring compatible components to adhere to some standard.

Acumos is a way to harmonize solutions across the full range of existing and future AI tools and technologies. Initially, Acumos supports such toolkits as SciKit Learn, TensorFlow, H2O and RCloud, and various programing languages. A portable Acumos

library packages the products of all these tools in such a way that they are all interoperable. By leveraging an open framework, for which all the source code is readily available and adaptable, the expectation is that the number of compatible SDKs and languages will grow over time as the Acumos community grows. By wrapping programs in a container and making each toolkit and language interoperable with the others, the range of available, compatible solutions that Acumos can use is large and will grow further over time.

For the foreseeable future, it is likely that no single, all-purpose AI hardware platform or development kit will dominate software design for ML solutions. Each problem and each approach is likely to require a different set of tools. Therefore, a platform that harmonizes many solutions by packaging them into interoperable microservices will be the best way to make sure that solutions will be useful for any target development community. Furthermore, because the field is changing rapidly, the future will include a wide array of tools that do not currently exist and probably have not yet been conceived. So any attempt to standardize on one approach or toolkit at this stage is doomed to failure. Instead, a tool that focuses on interoperability will help to keep today's solutions relevant for future applications.

Step 2 – Training the Basic AI Model

What is common across different AI platforms and tools is this new model of software development in which code must be enhanced through training. ML enables a wide variety of AI technologies, and about 60 percent of the 2016 AI investment has gone to the ML development (McKinsey Global Institute, 2017). The application of ML to business problems depends on the ability to train software by providing specialized data sets that represent a variety of conditions for the model to identify and act upon. The ability to quickly and reliably train software on any dataset is a component of every ML problem.

Because all ML is data-driven, algorithms based on ML technologies are substantially reusable in ways that prior generations of software were not. Conventional software tools are designed to address the problems of editing, inspecting and analyzing code to simplify debugging and reprogramming.

In ML problems, the process is more about taking a good basic algorithm and applying it to a new dataset to solve a new problem without developing any code. But to train the algorithm, it must be executed in an environment where the training data are available.

The ability to securely exchange, retrain and license a general-purpose solution is a key part of a commercial ML platform. Although entirely academic problems can be trained in a relatively isolated development environment into which the data are securely imported, real-world training requires moving the model to an environment where data can be conveniently and rapidly acquired. In other cases, such as robotics, crowd sourced datasets or mobile applications, for example, the data may only be available in some very specific runtime environment.

Acumos takes these packaged models and moves them from a secure development machine to a secure runtime and exposes the model to training data without the need for a developer to change the model in any way. This is done using custom training clients, data access and data caching tools, making it easy to assemble a specialized training application for each ML model. Acumos provides the training and testing interface needed to turn a basic model into a predictor that has been trained to perform a specific function.

Step 3 – Publishing Models to a Catalog

Data-driven ML models can be trained to recognize a pattern and to classify patterns reliably, but they cannot learn to do different things, once trained. Therefore, the third goal of a common AI development platform is making it easy to connect AI algorithms with different adapters that can apply the knowledge acquired in a training process to specific applications. For instance, when building a corporate access control system, it may be a simple matter to train a ML algorithm to recognize an employee of a company from personnel records. But before that model can be used to open and close doors, an adapter is required to actuate a lock when a known employee approaches the building entrance.

To build a working system, it is necessary that developers locate components that implement specific functions and evaluate what those components do and how they interface to external systems and data sources. Acumos does this by creating catalogs of useful functionality that can be searched and explored, reviewed and rated. Once a model is identified in a catalog, it can be acquired from the original developer and employed as a component in a complete solution. It is the catalog that connects a component developer with a population of potential users and facilitates that acquisition, transfer and updating of ML models.

Acumos provides a design studio that makes it easy to chain together a series of components by connecting data sources to model-based predictors and then connecting those predictors to adapters that operate equipment or provide other common functions like filling in spreadsheets or performing any developer-defined steps to create a simple and efficient decision tree. Such a design tool is not well suited to developing the actual ML algorithm. It is not intended to handle loops and recursive operations, but it is a tool that makes it extremely easy to integrate well-defined and independent microservices into powerful IT solutions.

Step 4 - Deploying Finished Solutions to a Runtime Environment

The final step in building a working AI solution, as shown in Figure 13, is to deploy a working application into a runtime system. In general, AI solutions will not run best in a centralized environment. For instance, it is hard to imagine an autonomous vehicle making an HTTP request to a central brain to make real-time decisions such as avoiding pedestrians. In most cases, a good AI solution will have to be packaged and delivered to a runtime system where it will actually be used. Acumos packages solutions into Docker image files, which can then be deployed into any Docker environment and managed through a set of container management tools, such as Kubernetes. Docker containers are a useful tool for deploying software, but other packaging and management systems exist today and no doubt others will proliferate over time. The basic design of packaging solutions and deploying them will be adapted to any container-like mechanism.

Acumos provides tools to package any set of components, including predictors, adapters and other microservices, as needed, to any runtime environment and to create a compatible, deployable image file. Such image files can be deployed to Azure, AWS or other popular cloud services, to any corporate data center or to any real-time environment as long as it supports Docker or other supported lifecycle management tool. The adaptive and programmable deployment interface allows Acumos to package and transfer runtime bundles to a wide array of external systems.

3.3 Network Data Collection and Analytics

Data Exposure Needs for AI

AI is data-driven. As such, AI technologies have a need for increasingly more sophisticated and granular data exposure capabilities with well-defined data definitions.

By exposing more data features, AI can more easily differentiate data patterns to more accurately characterize these patterns and thus provide better recommendations and decisions.

3GPP Data Exposure Capabilities

Data collection and analytics functionality is not entirely new for 3GPP. However, with the current focus on 5G, the topic has gotten a new impetus within 3GPP. In its specification on 5G System Architecture (TS 23.501), 3GPP has defined a new operator managed network function called the network data analytics function (NWDAF). In Release 15, the scope of NWDAF functionality was limited to providing analytics on slice-level load information to policy and slice selection functions in the 5G core network. Furthermore, how the NWDAF acquires the data to be analyzed was not defined.

For Release 16, a new study on the topic has been started to enhance this functionality by first studying and then specifying how to collect data from the network and how to feed data analytics back into the network functions for their use. New data analytics use cases are being looked at. Discussions on the topic are being captured in Technical Report TR 23.791. The results of the study are expected to result in normative specifications. This would increase the role of use of data analytics within the 3GPP networks. As of publication of this document, the study is still ongoing. This section provides a snapshot of status of work as of August 2018.

Figure 14 shows the proposed interaction between the NWDAF and various other network entities. The possible sources of input data include other network functions (NFs), application functions (AFs), OAM systems and network data repositories. The key recipient of results of analytics are NFs and AFs. However, use of the analytics output by OAM systems and its storage in data repositories is not ruled out.

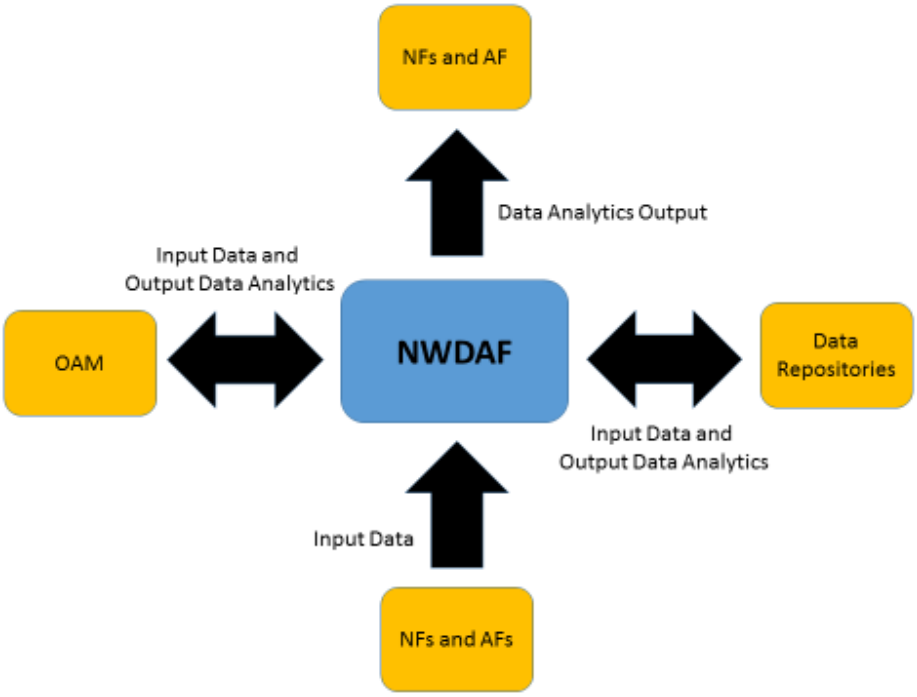

Figure 14 – 3GPP NWDAF Data Collection and Analytics Architecture

As part of the study, 3GPP has documented several use cases. Based on these use cases, key issues are being defined. Solutions for these key issues are currently being solicited from contributing companies. At the end of the study, agreed solutions will lead to a normative specification.

Based on the use cases, the following is a summary of two categories of key issues identified so far.

Key issues requiring a general solution applicable to any use case:

- Interactions of NWDAF with NFs and AFs.
  - How to expose data analytics information to NFs and AFs.
  - How to collect data for analytics from NFs and AFs.
- Interactions of NWDAF with OAM systems.

Key issues requiring a solution specific to use cases:

- Use of NWDAF for assistance in traffic routing e.g. based on location analytics, congestion analytics, etc.

- Use of NWDAF for network performance prediction.
- Use of NWDAF for QoS provisioning and adjustment.
- Use of NWDAF in selection of NF instances.
- Use of NWDAF in managing background data transfer.
- Use of NWDAF in management of massive IoT infrastructure.
- Use of NWDAF in customizing mobility management.

The initial set of solutions proposed so far address some of these key issues. These solutions can be summarized as:

- Release 15 service-based interfaces that use subscribe/publish and request/response can be used for interactions between an NWDAF and NFs/AFs.
- Service experience data can be provided to the NWDAF by the AF, and the NWDAF in turn can provision the policy function with new QoS information so that QoS can be adjusted for the service.
- NWDAF can collect UE mobility-related information, such as from OAM systems. It provides output of analytics on mobility data to the PCF, which upon further processing sends it to AMF. The AMF uses this information for managing thr registration area for the UE and possibly for paging the UE.
- NWDAF collects paging failure information from NFs by subscribing to the event. It analyzes the information to help predict paging failures, for example, at certain areas and/or at certain times and informs the network if the likelihood of paging failure exceeds a threshold.

Additional solutions are expected to be added in later half of 2018, with the study targeted for completion by the end of 2018. While the NWDAF is expected to use ML/ AI algorithms internally for analytics purposes, these algorithms are not going to be specified by 3GPP and will be left to implementation.

3GPP RAN-Related Capabilities

In addition to core network efforts related the NWDAF, 3GPP is embarking on a study of RAN-centric data collection and utilization for the 5G New Radio (NR) and LTE. This work will investigate the uses and benefits of RAN-centric data utilization, providing potential enhancements to a variety of SON features and other RAN optimizations while considering new 5G capabilities.

Additionally, 3GPP is identifying the impact of necessary standards needed for data collection and utilization for the defined use cases and scenarios, including:

- Identification of relevant measurement quantities, events and faults for collection and utilization.
- New procedures for configuration and collection of UE measurements, RAN node measurements and signaling procedures for distributed and central analysis.
- New procedures and information exchange required for the different use cases.

If deemed necessary, 3GPP will also investigate the benefits and feasibility of introducing a logical entity/function for RAN-centric data collection and utilization.

3.4 Distributed AI and Online Learning

The distributed nature of AI for networking lends itself naturally to distributed AI solutions. Edge computation enables both low-latency, high-value responses by executing and responding rapidly to local data without the need for real-time communications to data centers and cloud-based AI servers. This AI-enabled edge can also dramatically reduce network bandwidth by enabling long-distance communication with centralized servers for selected training purposes, features of interest and anomalies rather than high-bandwidth raw data streams. Some examples of edge-based AI include:

- **AI-based agents on the device:** Smartphone-based agents will increasingly understand who you are, what you want, when you want it, how you want it done and execute tasks upon your authority. Specific applications include tasks such as purchase recommendations to automatically managing connected home conditions prior to your arrival.
- **User authentication:** Security technology combined with ML, biometrics and user behavior can supplement current authentication techniques to provide a more secure experience. For example, smartphones can capture subtle attributes of user behavior to better identify the user. Alternatively, facial recognition can be used to better authenticate access to your bank account.
- **Emotion recognition:** Smartphones can use sensors to detect, analyze and respond to people's emotional states and moods. For example, car manufacturers could use an embedded camera to understand a driver's physical condition or gauge fatigue levels to increase safety.

- **Natural-language applications:** Continuous training and deep learning will improve the accuracy of speech recognition. This might enable smartphone to use context while better understanding the user's specific intentions.
- **Augmented reality (AR) and AI vision applications.**
- **Content detection and filtering:** Restricted or unwanted content can be automatically detected, flagged and alarmed (via notifications).
- **Photography:** Smartphones cameras could automatically produce better photos based on a user's individual preferences by adjusting area specific exposure, color temperature and other factors.
- **Audio triggered event detection:** The smartphone's microphone could be used to monitor surrounding sounds and via AI, trained to alert the user based on specific audio triggers (e.g., last boarding call for Flight XYZ).
- **Device management:** ML can improve device performance and battery life by disabling used applications and dynamically managing notification services.

As on-device processing capabilities increase, these device-centric AI applications are becoming more prevalent. In the past, cloud processing was required for the above referenced edge applications. Increasingly, new devices such as smartphones and drones are now equipped to run compute-intensive AI operations.

Edge AI processing provides faster user response times (no latency impacts for accessing the cloud) and can provide increased reliability because the AI application will work even if data connection speeds are challenged. Edge computation can also be used with distributed online learning that enables the global system to rapidly learn and adapt to data streams. Leveraging online distributed learning has the potential of solving many of today's privacy concerns by never requiring centralization of the private data containing streams. At some point, this privacy preserving aspect can be enhanced with homomorphic encryption and differential privacy techniques, thus eliminating, from an infrastructure perspective, the possibility of privacy-related catastrophic compromises.

4. Network Requirements in Support of AI

4.1 Required Capabilities

The use cases outlined in section 2 point to two broad needs regarding network capabilities: data exposure and APIs/network controls.

Almost all of the use cases assume the availability of comprehensive and granular data exposure capabilities with well-defined data definitions. Although network data exposure has been in the spotlight for several years given the industry interest in network data analytics, the use of this data for potentially near-real-time network control presents new challenges. In addition, AI/ML data exposure needs present new requirements for data granularity and timing. Specific exposure capabilities will be use-case-dependent. As such, as the network AI situation matures, it will become clearer as to specific network data exposure needs and gaps.

In addition, selected AI/ML use cases utilize network APIs and other control interfaces to enable automated network management. As with data exposure, the specific APIs and control interfaces needed depend on the use case deployment scenario.

4.2 Current Standardization Activities

Most telecommunications industry standards bodies have ongoing efforts considering standardization needs for AI/ML technologies. This work tends to segment into three distinct categories:

- Efforts in measuring and assessing AI/ML technologies to ensure that systems are reliable, unbiased, explainable and scalable.
- Efforts to ensure privacy of both the data available to AI systems as well as AI results which may correlate external data to personal information (e.g., facial recognition systems that attempt to provide insights on personal preferences).
- Efforts to ensure that systems surrounding the core AI/ML processing can provide sufficient data exposure interfaces and, where needed, relevant APIs to effect network controls.

The first category above related to the measurement and assessment of AI/ML technologies is focused on providing controlled access to select data repositories. As a result, AI developers can train complex and never-before-solved AI solutions in an expanding number of domains. In addition, efforts are underway to develop AI evaluation methodologies and standard testing protocols.

The domain of privacy is yet unresolved. Serious privacy concerns exist regarding use of facial recognition systems, recommendation systems and other data analysis and tracking systems that have the potential to publicly avail personal attributes and

preferences that many people prefer to keep private. This domain also intersects with public policy as these privacy concerns overlap with existing privacy regulations and may invoke new regulations directly targeted toward AI/ML applications.

Network applications of AI/ML will generally leverage network exposure and API capabilities. In many cases, both aspects can be worked independently of AI/ML and often serve a broader scope within the network management and operations domain.

For example, as noted in section 3.3, 3GPP is addressing the need for standardized data exposure technologies in 5G networks through the creation and definition of a new network function called NWDAF. This function is defined in 3GPP TR 23.501 with applicable use cases in TR 23.791. The NWDAF system will be critical to many 5G use cases regardless of whether AI/ML technologies are used. In addition, 3GPP is embarking on a study of RAN-centric data collection and utilization for 5G NR and LTE.

Similarly, many of the APIs and control points usable by network AI/ML systems are being worked independently as part of the network evolution toward NFV and SDN.

5. Conclusion

This report looks more closely at how AI and ML technologies can be leveraged to address the pressing challenges facing the ICT industry today. AI is generally considered to be intelligence exhibited by machines or computational systems that perceive their environment and take actions to satisfy an intent. AI applications span a wide range of options from:

- Assisted intelligence comprised of targeted/narrow expert systems which help people to perform tasks faster and more accurately to:
- Autonomous intelligence systems with fully automated decision-making processes coupled with ML to perform a narrow task without human intervention while adapting to changing conditions.

AI can be utilized to better realize automated intent-based systems within the network.

The application of AI to network systems may require fundamentally new processes at each stage of the application lifecycle. In addition, although AI systems are excellent for applying cognitive processing to complex systems, errors will occur. In network applications where high levels of reliability and service availability are required, care must

be taken to ensure protective mechanisms are in place to mitigate and manage autonomous actions driven by AI systems.

This report documents a wide variety of network-related AI use cases including:

- Network anomaly detection.
- Network security.
- RAN Optimization.
- Dynamic traffic and capacity management.
- Network resiliency and self-healing.
- AI and orchestrated management.
- AI-based subscriber insights.
- AI-assisted customer support and sales.
- AI-based content processing and management.

These use cases often require robust network data exposure capabilities and network APIs (when autonomous actions are required). AI/ML enables new ways to understand and use data, renewing the need for timely access to more unique data. This data can be categorized along four basic dimensions: traffic-based attributes, network/subscriber state, topology/location and time/history. 3GPP has initiated standards work on new data collection architectures and use cases. AI/ML-based automation will also require better network APIs (e.g., making good use of NFV/SDN infrastructure).

These use cases also expose the likelihood of multiple AI closed-loop systems interacting with each other. Loops created with AI may operate at different time scales and need to be well designed from a broad network perspective to prevent network instability associated with interacting feedback loops.

Edge computation enables both low-latency, high-value responses by using AI-driven applications to use local data without the need for real-time communications to data centers and cloud-based AI servers. An AI-enabled edge can dramatically reduce network bandwidth, decrease user response times and potentially increase application reliability. Edge computation can also be used to enable distributed online learning, which has the potential of solving many of today's privacy concerns because private data need not be sent to the centralized processing functions for ML purposes.

Finally, it seems clear given the wealth of AI/ML network centric use cases that this technology can provide significant value to network related applications, services and operations.