HEAVY READING
WHITE PAPER

Independent market research and
competitive analysis of next-generation
business and technology solutions for
service providers and vendors

Contextual Service Assurance in 5G:
New Requirements, New Opportunities

A Heavy Reading white paper produced on behalf of TEOCO

TEOCO

AUTHOR: GABRIEL BROWN, PRINCIPAL ANALYST, HEAVY READING

A NEW SERVICE ASSURANCE FOR 5G

The range of 5G services now under development, from virtual reality to production-critical automation to massive-scale Internet of Things (IoT), has the potential to catalyze new value chains and operating models across every industry worldwide, and in turn, underpin the long-term commercial future of network operators. Realizing these benefits will require technology that fundamentally changes how networks are designed and services are consumed.

This white paper addresses how service assurance should evolve to meet the new demands introduced with 5G. It makes the case for automated, closed-loop network management that will enable operators to meet demanding service-level agreements (SLAs) across a range of dynamic and diverse service types, introducing the concept *of contextual service assurance* as central to 5G. It uses the example of network slices sharing a common resource pool, with each slice optimized to the functional requirements of a customer or application type. It also looks ahead to how machine learning and artificial intelligence (AI) will help operators manage the vast scale and diversity of services and network functions that will characterize 5G.

New Architectures, New Domains, New Services

Service assurance is used to monitor, model and analyze network data to make sure service quality levels are achieved and maintained. Its purpose is to deliver an optimal subscriber experience, according to network policy, resource availability and commercial terms. This makes service assurance tools critical to network operators. 5G and cloud introduce new technologies and processes, which in combination generate a more dynamic networking environment, driving the need for contextual service assurance. **Figure 1** identifies the major changes, and challenges, introduced with 5G.

Figure 1: Service Assurance Challenges in 5G & NFV

Data Volume	Vast amounts of telemetry & metadata to analyze	▶	Machine learning & contextual analysis
Cloud & SDN	Virtual core; dynamic resource allocation; no physical ports	▶	Virtual probes; multi-layer root cause analysis
Network Slicing	Common resource pool, with slice optimized to functional requirements of an application	▶	Real-time slice monitoring; correlate multiple slices; scale/out
New Use Cases	New performance requirements; diverse customer expectations	▶	User-configurable services; SLA reporting via customer portals
RAN Diversity	vRAN; mmWave; numerology; MIMO; shared spectrum; energy	▶	New SON tools; closed-loop optimization

Source: Heavy Reading, TEOCO

Contextual service assurance in 5G networks refers, in the first instance, to the ability to monitor a horizontal infrastructure running multiple service types that can be dynamically requested, instantiated, scaled and terminated. This network will be characterized by extremely large data volumes – both payload and signaling – that will overwhelm traditional service assurance tools and manual oversight by engineers in the network operations center. In a cloud-based 5G network, the impact of one service type on another must be understood and managed, according to context – i.e., the importance of the service, its likely duration, performance requirements, resource availability and prevailing network conditions.

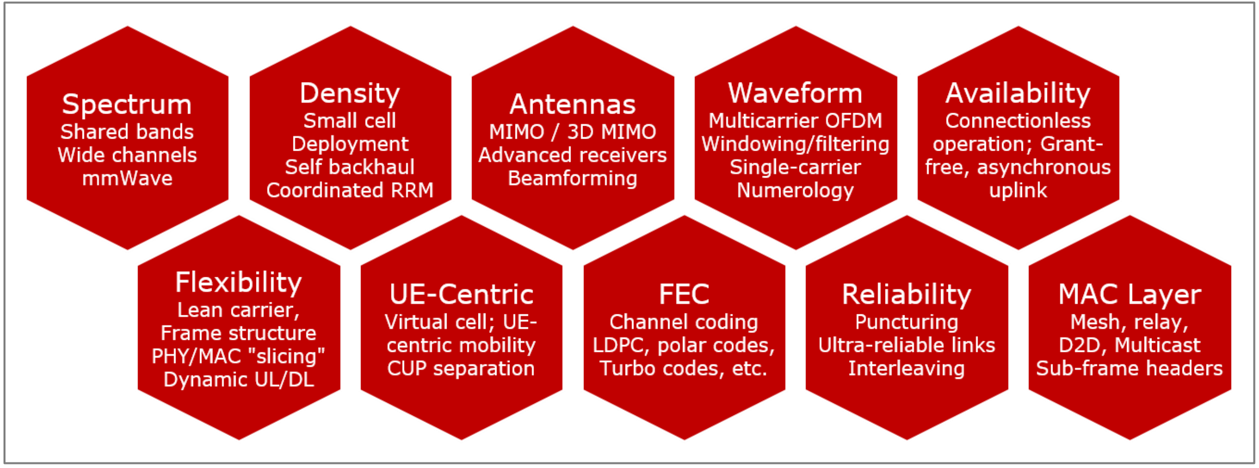
The economic case for 5G is predicated on the ability for operators to extend the network to support coexistence of various service types on an end-to-end basis, across multiple network domains, including radio access, transport and core network. Therefore, how these domains map to the underlying infrastructure, and understanding the service and network context, is critical to effective service assurance.

Contextual service assurance, more broadly, is multidimensional. Extending the definition beyond the network-service interaction in 5G, to include factors such as subscriber state, location-specific conditions, the prevailing performance of adjacent and underlay networks – e.g., 3G/4G radio access network (RAN) and IP transport – and predictive information on possible future performance, enables operators to build a fuller picture of network context. Using this in combination with business information, operators can use service assurance to underpin policy decisions and, as they gain confidence, closed-loop automation.

New RAN Capabilities

Although 5G defines a multi-access system architecture, and although service assurance must run end-to-end, 5G will stand or fall on the radio. A unified air interface is important to economies of scale in research and development (R&D), manufacturing, deployment and operation. **Figure 2** shows the foundation technologies that make up 5G New Radio (NR), which is the new specification being defined by 3GPP.

Figure 2: Enablers for 5G NR



Source: Heavy Reading

To be able to scale across spectrum bands and deployment scenarios, the 5G air interface is highly configurable. This has implications for service assurance at both the air interface and wider RAN architecture level. Some of the major factors to consider are:

- **Flexible Air Interface:** To scale to different channel widths, at different frequencies and in different scenarios, 5G NR uses flexible numerology with a flexible frame structure. This has many, many implications for service assurance – one important example (among many) could be PHY/MAC "slicing" to extend network slicing from the core, across the radio to the device radio.

- **Unlicensed & Shared Spectrum:** There is now substantial research that shows shared spectrum can generate higher utilization than single-use spectrum in many scenarios. Shared spectrum for 5G NR is now a 3GPP study item, ear-marked for standardization in Release 16. The ability to extend services beyond licensed bands is important for many reasons. For example, enterprises may want to deploy on-premises private networks, or an operator may want to add extra capacity. In other cases, the use of unlicensed spectrum may breach an SLA and should not be used.
- **Energy KPIs:** With the cost of electricity, efficiency is critical to network economics (innovation on "lean" radio design is needed to address this) and even more critical on battery-powered devices, notably IoT sensors. A service assurance solution should ensure that the network does not request higher power output from IoT devices, as this will greatly reduce their lifespan and degrade the economics.
- **RAN Architecture:** 5G extends virtual RAN and cloud RAN architectural concepts to offer a model for coordinated radio resource scheduling and L3 packet processing that may, for example, allow operators to tightly map application requirements to radio resources. Other emerging architectures, such as mesh, relays and multicast, will similarly impact application performance and, therefore, service assurance.

Virtualization & Cloud

Service assurance changes substantially – radically, even – when using virtualized functions. Physical appliances are designed with built-in redundancy at the line card and chassis level, and are then deployed in redundant configuration (1+1 or N+1), which helps to ensure fail-over and continuous operation of the service. This works well, but is hard to adapt to changing circumstances, and it can be expensive. Decoupling applications from hardware has major implications for fault, configuration, accounting, performance, security (FCAPS) and for service assurance more generally. This impact is multiplied when virtualized network functions (VNFs) are deployed on shared, multi-vendor cloud infrastructure.

Moreover, it is common for networks to incorporate physical and virtual functions in a hybrid model, which adds further complexity that must be addressed by the service assurance solution. For the immediate future, monitoring tools must be able to assure traditional and virtualized networks and understand the relationships between them.

The network functions virtualization (NFV) model moves some of the responsibility for failover to the cloud platform. The intent is that redundancy and resiliency are inherent to the NFVI (a.k.a. the NFV cloud platform) so that operators no longer need to spend as much on redundant equipment that may only be used rarely. A second order effect is that this allows for, and is enabled by, simpler, stateless, "cloud-native" VNF designs.

The NFVI must monitor the performance of the infrastructure to support scale out of VNFs and associated virtual machines or containers, according to demand, or to recover in the event of failure. This references the models used by hyper-scale cloud providers, which have been shown to offer high-reliability services using low-reliability components in highly changeable demand scenarios. Service assurance tools, therefore, may need insight into the performance of the underlying technology platforms – for example, to identify cache utilization rates in server hardware.

Monitoring the performance of the NFVI is the job of the VIM (e.g., OpenStack) working as, or in combination with, a "resource orchestrator." These functions work in conjunction with a VNF manager to ensure VNFs have the compute, storage and networking resources they need to meet the performance targets required by the service.

The development of monitoring tools has lagged the NFV vision somewhat. However, with the emergence of virtual probes, streaming telemetry data and new analytics frameworks, real progress is now being made. This is critical, given that several 5G network functions will be hosted on cloud infrastructure.

NETWORK SLICING: CONTEXT DRIVES DECISIONS

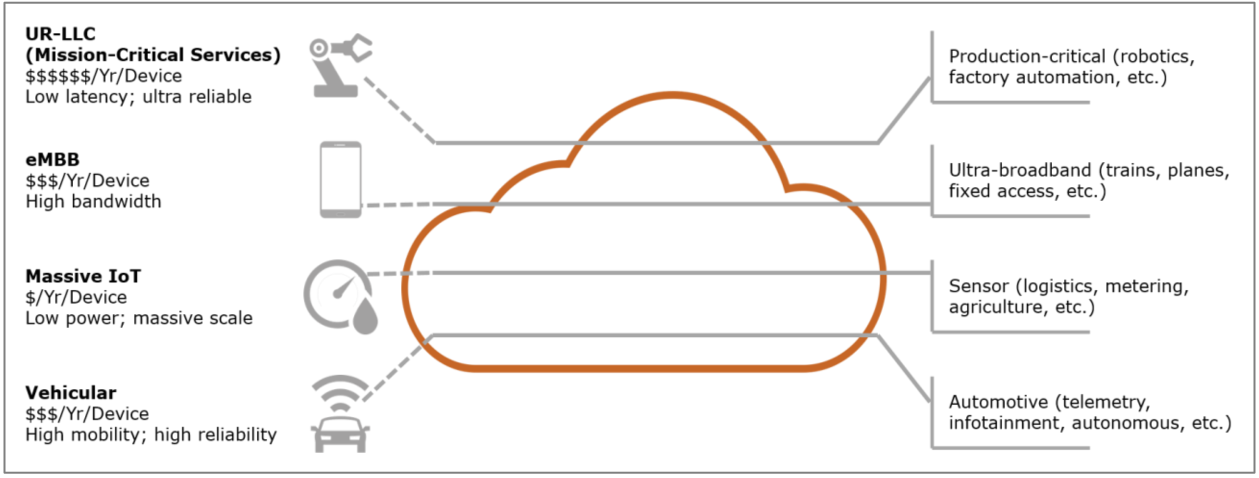
Network slicing is the feature of 5G that has garnered the most attention, and has the most commercial promise. This is the ability to support diverse services on a common network platform, with each slice optimized to the specific functional requirements of a customer or application. Taken to its logical conclusion, 5G network slices can be thought of as the network adapting itself, in software, to the needs of the application.

A useful way to define a network slice is using an adapted version of the Next Generation Mobile Network Initiative (NGMN) definition, as follows: "A set of network functions instantiated to form a complete logical network that meet the performance requirements of a service type(s)."

A network slice consists of three layers: service instance layer, network slice instance layer and resource layer. A slice is typically made up of sub-network instances to create an end-to-end service, which in a mobile network includes RAN, core and service platforms. Contextual service assurance should operate across these layers and end-to-end. Slices can be fine-grained, at the per-user or per-service level, or more coarse-grained, at a company or industry level (e.g., a connected car slice or meter-reading slice).

The concept is outlined in **Figure 3**. In practice, coarser-grained slices are likely to come to market first, with increasing granularity over time.

Figure 3: Network Slicing Concept


UR-LLC
(Mission-Critical Services)
$$$$$$/Yr/Device
Low latency; ultra reliable
eMBB
$$$/Yr/Device
High bandwidth
Massive IoT
$/Yr/Device
Low power; massive scale
Vehicular
$$$/Yr/Device
High mobility; high reliability
Production-critical (robotics, factory automation, etc.)
Ultra-broadband (trains, planes, fixed access, etc.)
Sensor (logistics, metering, agriculture, etc.)
Automotive (telemetry, infotainment, autonomous, etc.)

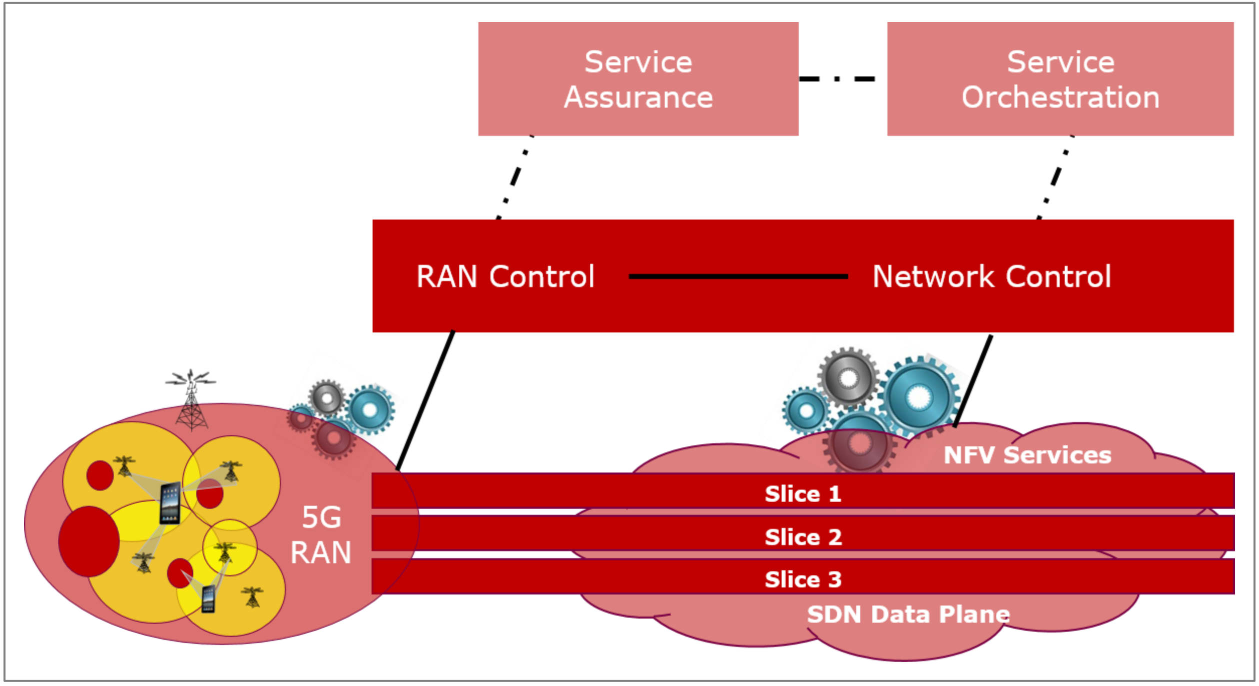
Source: Heavy Reading

End-to-End Network Slices

To meet demanding performance requirements in a 5G mobile network, slices should run end-to-end from the radio to the core network and the application logic. This requires service and resource orchestration across the main functional domains. Each of these domains, however, has its own control system. In the transport network, software-defined networking (SDN) may be used for traffic segmentation and to apply policy; in the RAN, packet scheduling may be managed by a radio controller or the base station itself; and in the cloud-based core network, an NFV orchestrator will be used.

To realize the vision of zero-touch network service management – with no human involvement – each of these domains must be coordinated from a services layer perspective, as shown in **Figure 4**. This also requires the service assurance to "talk" across domains. New topologies, such as edge computing, that may emerge to support ultra-reliable and/or ultra-low-latency services will further force inter-domain coordination.

Figure 4: Service Assurance for 5G Network Slices



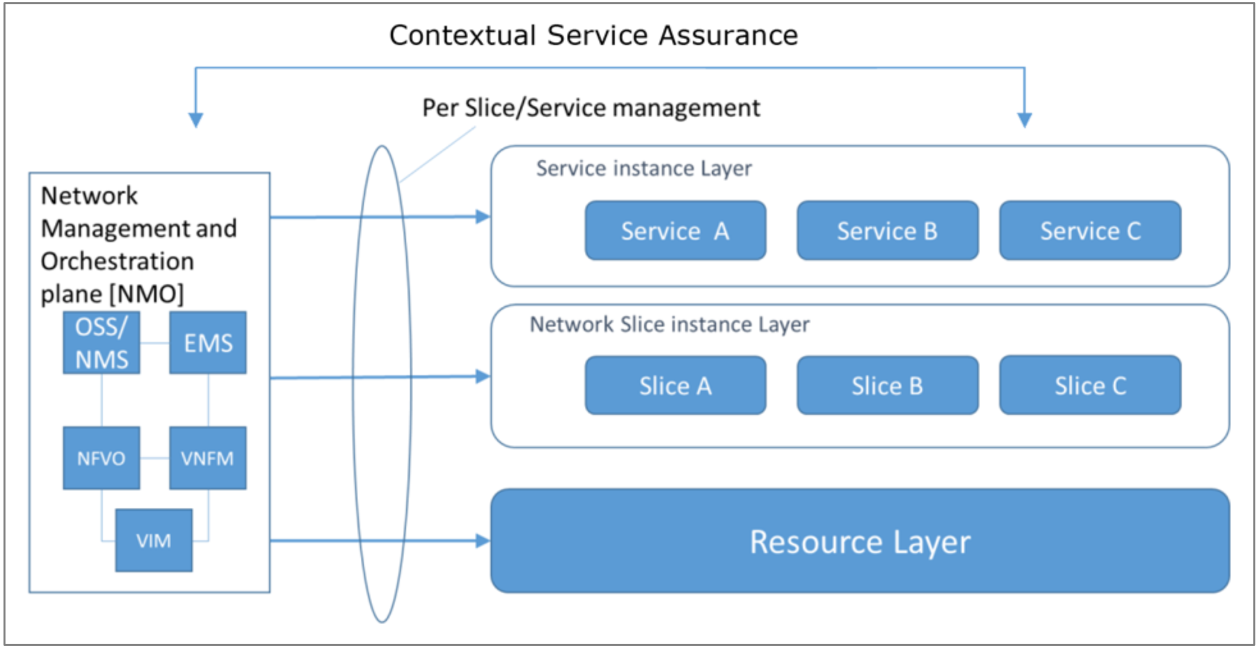
Source: Heavy Reading

SLAs & OLAs per Slice & per Customer

A network slice is a self-contained virtual network. Each slice must meet SLA and operating level agreement (OLA) thresholds determined by the provider and customer. Therefore, service management and assurance should apply on a per slice basis. Lifecycle management is also needed to manage the onboarding and scale-out and scale-in processes, particularly where slices are relatively short-lived or where the slice flexes in terms of resource requirement or load (e.g., due to a special event).

The challenge is that multiple slices use the same resource layer, as shown in **Figure 5**, and may compete for resources predictably or in ways that are unforeseen or subject to external events. This is where contextual information is critical. Service assurance tools must be able to correlate service experience with the underlying resource use to ensure a slice is not consuming more than it needs or that sufficient capacity is available for scale-out if the service is close to reaching a capacity or performance threshold. Different services will have different thresholds (e.g., a low-latency or ultra-reliable service will have hard thresholds that cannot be breached, whereas an IoT service may be able to flex with demand). Understanding this interplay between slices competing for resources and the service experience is at the heart of contextual service assurance in 5G.

Figure 5: Per-Slice/Service Management


Contextual Service Assurance
Per Slice/Service management
Network Management and Orchestration plane [NMO]
OSS/ NMS
EMS
NFVO
VNFM
VIM
Service instance Layer
Service A
Service B
Service C
Network Slice instance Layer
Slice A
Slice B
Slice C
Resource Layer

Source: 5G Americas, Heavy Reading

SERVICE ASSURANCE & CLOSED-LOOP AUTOMATION

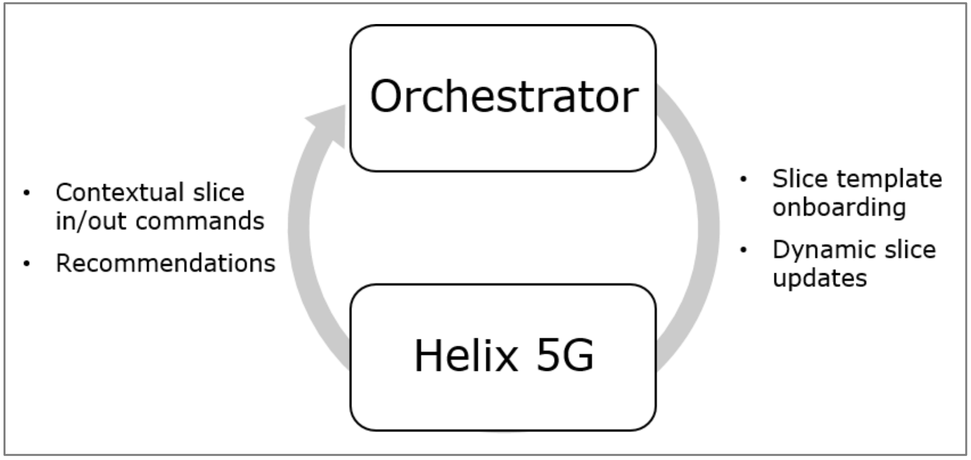
For advanced 5G services to be attractive, or even economically viable, operational complexity must not outweigh the associated benefits and revenue. The drive for operational simplicity and service agility is therefore also a driver for greater automation in 5G networks.

Service Assurance & Actuation

Automation depends on feedback loops between the service experience, the network configuration and resource utilization. To make changes to the network configuration (to "actuate") is the role of domain orchestrators, which configure VNFs and other appliances in the processing path (routers, etc.). These orchestrators would ideally use common application programming interfaces (APIs), such as those developed by ONAP, ETSI NFV and the TM Forum, as well as standardized protocols and data models, such as NETCONF/YANG. This desire to actuate network functions using orchestrators, without the need for configuration by human operators, is driving work across the industry on initiatives such as intent-based networking and the proposed new ETSI Special Interest Group on automation.

There are several challenges to automating actuation of network functions. A major one is how to make orchestrators aware of the state of the end-to-end network service and what role each domain is contributing to a satisfactory or unsatisfactory performance. A solution is to use service assurance to provide context information to the domain orchestrators. This concept is shown in **Figure 6** as a closed-loop relationship that allows for continuous optimization of services and resources.

Figure 6: Using Service Assurance for Closed-Loop Automation in 5G



Source: TEOCO

Service assurance tools combine many data sources and analyze it to derive a view of service performance. Data sources can include probes, network elements, orchestrators, element management systems, analytics platforms and more. The more real-time this view, the faster the orchestrator can react and, in turn, enable hysteresis-based scale-in and scale-out mechanisms for different network services (a.k.a. "slices") on shared resources. This form of closed-loop automation, and the operational efficiency gains and agility associated with it, depend on an authoritative, real-time view of the network and service state provided by assurance tools.

MACHINE LEARNING & AI

Machine learning is a subfield of computer science that enables computers to learn without being explicitly programmed; one way it is useful is to identify functional relationships mathematically between data that are not easily spotted by humans. Given that networks generate enormous volumes of telemetry data that is far beyond human comprehension, it is a reasonable to investigate how machine learning might be useful to network operators. These techniques can be thought of as an extension of today's performance monitoring, big data analytics, orchestration and automation tools.

Using Machine Learning for Pattern Identification

The vast data sets generated by networks overwhelm human pattern recognition. The solution has been to use analytics tools to process this data, identify patterns and provide higher-level views to human operators that allow them to act to achieve an outcome. These tools are essential to modern networks and to service assurance. Automation takes this a step further by reducing the need for human oversight; however, even in this case, analytics tools use human-coded algorithms. The opportunity is to use machine learning to reveal new insights in large data sets and, in turn, new ways to operate networks efficiently and reliably. In this sense, machine learning adds a layer of "intelligence" to network management and automation toolsets already being developed by the industry, enabling them to make better decisions.

As discussed, closed-loop automation is a key objective of network operators. Risk averse operators of mission critical network infrastructure are, however, reluctant to trust automation tools without evidence that they are reliable. This creates a challenge for the use of machine learning because, as seen in other domains, the technology is valuable precisely because it reveals unfamiliar patterns and solutions. For example, Google's DeepMind AI beat the world's best Go players by using unconventional strategies. If machine learning proposes radically unconventional changes to nationally critical service provider networks, it will take time for operators to trust the technology in closed-loop automation. Sandbox environments for testing and development will be needed.

Use Cases for Machine Learning in Networking

There are several use cases either proposed or under testing for machine learning in operator networks. Some examples include:

- **AI-Assisted Network Automation:** In 2016, Japanese carrier KDDI claimed the world's first successful AI-assisted automated network operation. The AI-based monitoring system provided advance detection of potential failure caused by, for example, software bugs, with greater than 90 percent precision. It was able to maintain service continuity by migrating functions to new NFV platforms at a different site. This occurred five times faster than using conventional techniques, said KDDI.
- **Data Center Operations:** This Google initiative to train neural networks to improve power usage effectiveness (PUE) in its data centers has become one of the canonical examples of machine learning in network contexts. The system was able to consistently achieve a 40 percent reduction in the amount of energy used for cooling, which equates to a 15 percent reduction in overall PUE overhead.
- **Industrial IoT Security:** Deutsche Telekom is developing machine learning software to identify threats and anomalies to IoT devices, with a view to stopping malware incursions before they reach their target. It has showcased security solutions that use machine learning to protect industrial control systems against malware, such as the infamous Stuxnet worm.
- **Natural Language Processing for Customer Service:** Swisscom has an in-house machine learning lab working on natural language processing to improve customer service and customer experience. It believes that the technology could lower the cost base associated with customer care using sentiment analysis for enterprise and consumer users to detect whether customers are unhappy, assess the nature of the problem, and route the customer to the appropriate agent.
- **AI-Augmented Optimization:** Vodafone used machine learning in a centralized self-organizing network (C-SON) trial to identify the optimal settings for voice over LTE services across 450 mobile cells chosen at random. This task would have taken an engineer around 2.5 months to do manually, according to Vodafone. The algorithm completed the task in four hours.

- **Customer Sentiment Analysis:** Australian operator Telstra has made a strategic commitment to machine learning and is applying it across its business. The first applications are in user-experience sentiment analysis to gauge upsell potential and customer support requirements; marketing campaign analysis, to gain quicker feedback on what works and what doesn't; and security threat analysis for managed services customers.

CONCLUSION & SUMMARY

The wide range of 5G services under development are vital to the long-term economic health of the mobile networking industry. The ability to support these diverse service types on a common infrastructure and technology platform is fundamental. 5G introduces many new technologies and processes in the radio, in the cloud core and in the software-defined transport network, which generate a more dynamic networking environment, driving the need for contextual service assurance. Network slicing shows how competing services can benefit from cross domain orchestration and end-to-end assurance.

Contextual service assurance and orchestration can be used to support automated, closed-loop network optimization that will enable operators to meet SLAs across service types. In combination with business information, operators can use multidimensional service assurance to underpin network policy and subscriber policy decisions.

Although often considered a pure research topic, operators are already making use of machine learning (albeit in a limited way) to manage and interpret vast data sets. Over time, as research progresses, it may be possible to create new solutions to complex multi-dimensional problems using machine learning and, in so doing, to add another layer of intelligence to today's state-of-the-art service assurance solutions.