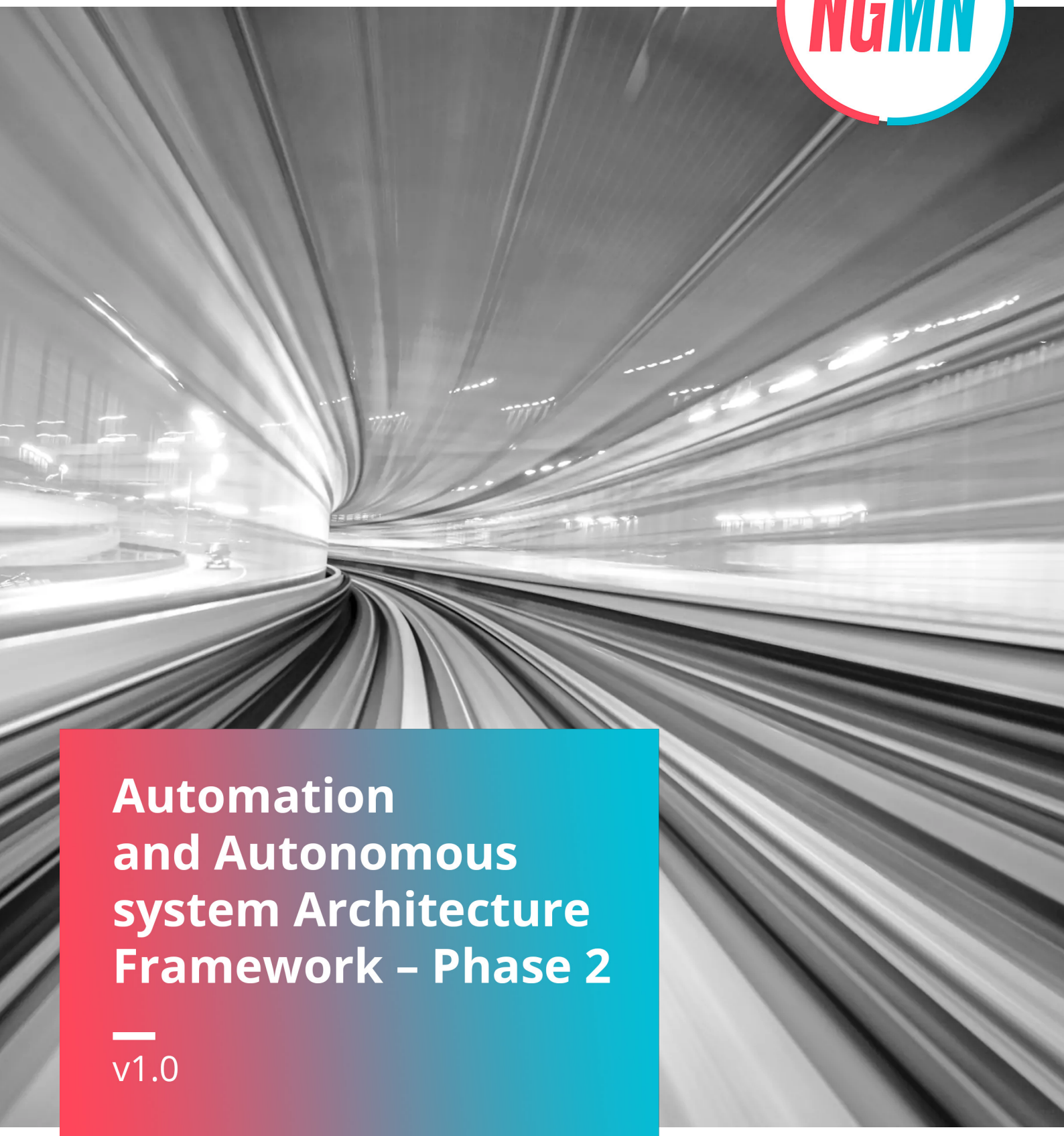

Automation and Autonomous system Architecture Framework – Phase 2

v1.0

www.ngmn.org

WE MAKE BETTER CONNECTIONS

AUTOMATION AND AUTONOMOUS SYSTEM ARCHITECTURE FRAMEWORK – PHASE 2

by NGMN Alliance

Version:	1.0
Date:	01 October 2024
Document Type:	Final Deliverable (approved)
Confidentiality Class:	Public
Project:	Network Automation and Autonomy Based on AI
Editor / Submitter:	Sebastian Thalanany (UScellular)
Project co-leads:	Lingli Deng (China Mobile), Sebastian Zechlin (Deutsche Telekom)
Contributors:	Jean Paul Pallois (Huawei), Andreas Volk (HPE), Jermin Girgis (TELUS), Tony Verspecht (Cisco), Gary Li (Intel), Yuhan Zhang (China Mobile), Amit Dass (Cisco), Luigi Licciardi (Huawei), Manchang Ju (ZTE), Sebastian Thalanany (UScellular)
Approved by / Date:	NGMN Alliance Board, 24 September 2024

ABSTRACT: SHORT INTRODUCTION AND PURPOSE OF DOCUMENT

This document leverages the foundations of an autonomous system [1], applicable to a service-based architecture [2] to further examine the forward-looking considerations for an autonomous system architecture framework, as the next generation End-to-End (E2E) system continues to evolve.

The functions and characteristics that embody the capabilities of an E2E autonomous system for network automation are examined in terms of a high-level framework. Artificial Intelligence/Machine Learning (AI/ML) models [3] of cognition and application aspects are pivotal ingredients within an autonomous system. The term "system" in this context is an abstraction, which generalises and subsumes details such as specific networks, protocols, and implementations, in terms of high-level requirements, perspectives, and insights.

Network slicing serves as a foundational building-block, for a realisation of flexible, granular, and optimised allocation of system resources, such as computing, networking, and storage, to support various deployment scenarios, and service innovation. An automation of network slicing in an E2E system imbued with autonomous system capabilities is a significant value proposition, where an autonomous system self-adaptively manages complexity in a virtualised environment, without human intervention. The application of assorted AI/ML models, together with continuous training, integration, and delivery, facilitate continuously updated autonomous system behaviours to suit diverse deployment arrangements and services.

The autonomous system architecture framework is intended to serve as guidance in the development of inter-operable and market enabling specifications, for a continuing advancement of an automated and self-adaptive 5G ecosystem of heterogeneous access, virtualisation, forward-looking service enablers, and emerging usage scenarios.

TABLE OF CONTENTS

01 INTRODUCTION

The purpose of this document is to infer and delineate a high-level framework of architectural principles and requirements. that yield network automation and autonomy based on Artificial Intelligence/Machine Learning (AI/ML), for system-wide network and operational automation, without human intervention. Considerations and requirements that build and advance the foundational aspects are examined, in terms of system-wide aspects and insights, for further understanding, usage scenarios, and specifications.

With the continuing evolution of the 5G Advanced ecosystem [2], this document advances and leverages the foundational requirements in the initial phase of the autonomous system architecture framework [1]. The advancements provide system-wide guidance and direction for standards development with respect to an articulation of interoperable capabilities, and services, associated with an autonomous system oriented network automation. It is anticipated that the 5G Advanced ecosystem evolution will be characterised, in terms of a convergence of diverse access technologies, virtualisation, hybrid clouds, together with high levels of decentralisation and distribution, to facilitate emerging services, with a customisable, user-centric experience.

With these intrinsically interdependent requirements, an accrual of the associated system-wide complexity is expected, which would need to be managed and scaled effectively, through the use of an autonomous system architecture framework for system-wide network automation.

02 EXPECTED BENEFITS AND COMMERCIAL IMPACT

The management of complexity and system performance optimisation are significant benefits, realised through an application of autonomous system constructs for realising network automation. As a result, it is anticipated that network automation will serve as a foundational catalyst for enabling service innovation, evolution, support for diverse business models, and flexible deployment.

At the same time, network automation facilitates a minimisation of operation and maintenance expenditures, as well as enabling continuous improvements in configuration, integration, upgrades, service experience, personalisation, fault mitigation and management. Commercial beneficiaries of network automation include Network Service Providers (NSPs) (e.g., operators), Service Providers (SPs) (e.g., Verticals), and users (e.g., human and machine interfaces).

An autonomous system rendered network automation framework, provides a holistic framework for a dynamic self-adaptation of the system to a given operating environment, while satisfying diverse Key Performance Indicators (KPIs), personalisation of services, and various Key Value Indicators (KVIs), associated with social, economic, and environmental aspects. This framework provides architectural considerations for enabling the requisite operational capabilities to meet the growing system and service demands that exceed human response limits, as a result of the increasing system and service complexity, which accrue with continuing technological advances (e.g., virtualisation/softwarisation, network distribution, and decentralisation).

03 DEFINITIONS

AUTONOMIC FUNCTION

A function with intelligent and cognitive attributes, within an autonomous system, which operates through closed-loop feedback of a response for a given stimulus, for an automatic and adaptable behaviour (subject to input governance policies and configuration), and is able to derive all the necessary information, through the discovery of knowledge within its environment.

AI AND ML MODEL

A model representing mathematical algorithms that learns using data and input consisting of human expertise to generate an effective and optimised decision, in the presence of dynamic change, when the model is provided with actual information of a corresponding nature for which the model was designed.

KEY PERFORMANCE INDICATOR (KPI)

This refers to a measurable metric (e.g., data rate, spectral efficiency, latency etc.) that reveals the performance of a system or entity with respect to a specific objective associated with the system or entity, for obtaining insights related to the performance of the system or entity.

KEY VALUE INDICATOR (KVI)

This refers to a metric for monitoring and validating the impact of prominent societal, economic, and environmental values [3] on emerging technologies and vice-versa, to study and shape the development direction of technologies, guided by balanced and holistic considerations.

MACHINE LEARNING MODEL

A model created by a machine through an application of learning techniques on input data. The model may be utilised to generate predictions (e.g., regression, classification, clustering etc.) on untrained or raw input data. Encapsulation of the model may be performed with software (e.g., within a virtual machine or container). The learning techniques span a broad variety of algorithms (e.g., learning of a function that maps input data into corresponding output data).

MACHINE LEARNING DATA MODEL

This pertains to a description of the data used for data handling in machine learning applications. The data model may specify the data exchanged between an ML overlay network (e.g., virtualised network) and an ML underlay network (e.g., physical network). The data model includes data structures as well as a semantic description, while collecting data from an ML underlay network, and while applying the output from the ML overlay network to the ML underlay network [4].

04 AUTOMATION AND AUTONOMOUS SYSTEM CONTEXT

This chapter provides a brief overview of an autonomous system context and background for self-adaptive automation, for guiding a consistent understanding of high-level considerations, without reference to specific implementations.

The scale and complexity of next-generation networks are expected to continue to rise, to enable higher levels of sophistication with respect to the end-to-end system and services to meet the demands of higher orders of personalised human experience, as well as higher orders of end-to-end system optimisation, resource utilisation (e.g., networking, spectrum, computing, and storage). These emerging directions demand an overlay of autonomous system constructs, consisting of decentralised and distributed intelligence, with feedback control loops, for a realisation of self-Configuring, Healing, Optimising and Protecting (CHOP) automation.

As shown in Fig. 1, the autonomic principles are applied and embedded within an end-to-end system, consisting of distributed AI/ML model oriented cognitive functions, yielding self-CHOP network automation. This self-CHOP system-wide behaviour leverages a variety of cooperative technologies to sustain a system-wide equilibrium under changing conditions, in terms of effectively satisfying the performance objectives and a personalisation of services, while dynamically and flexibly interacting, with the operating environment.

The autonomous system constructs (e.g., AI/ML model assisted feedback control loops) within the end-to-end system are expected to be layered, with more complexity and scope management towards the network core, while autonomous system constructs at the network edges (e.g., distributed radio networks) are expected to satisfy ultra-low latencies. These autonomous system constructs are anticipated to be optimised for energy efficiency, in terms of training and updating the AI/ML models. A framework of autonomous system constructs is expected to operate within a given administrative domain, as well as across cooperating administrative domains in Federated Learning (FL) arrangements [5]. In such a framework, the nature of the constituent AI/ML models is subject to human governance, through the related governance interfaces

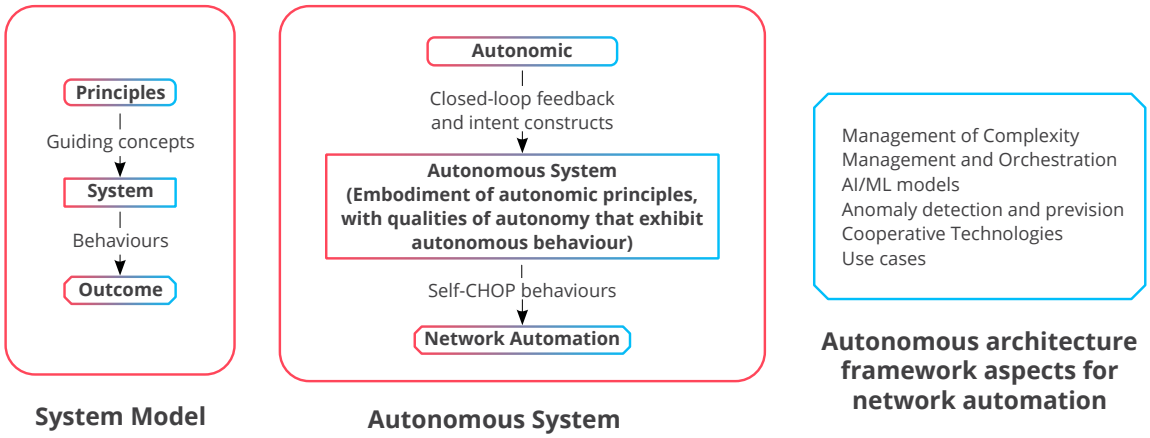

Fig. 1: Autonomous system context

to ensure the intended behavioural and performance objectives of the constituent autonomic functions, while remaining compatible with a broader social and environmental context.

The use of distributed autonomous system constructs over a service based architectural framework, is pivotal since it opens a realisation of innovative business models and diverse deployment choices, across cooperating and diverse administrative domains that allow multi-tenant and multi-service considerations. While system-wide flexibility, resource granularity, and functional portability are intrinsic characteristics of a service-based architecture, the performance demands combined with an agile cooperation demand across distributed resources, within and across domains, together with a maintenance of intended performance and service objectives, contribute to the rising levels of system-wide complexity, as the system evolves.

Consequently, the conventional and rigid arrangements of system and service management are inadequate, and thereby underscore the necessity of an overlay of a distributed autonomous system over a service-based and network sliced architectural framework. The self-CHOP attributes that are intrinsic to the behaviour of autonomous system constructs are essential for self-adaptive automation, which then demand the related studies and creation of system-wide interoperable specifications.

4.1 MANAGEMENT OF COMPLEXITY

Management of complexity, through the use of autonomous closed-loop feedback constructs is pivotal for self-CHOP network automation. This type of network automation is beyond simple, piecewise, and programmatic automation, which is limited in terms of network automation adaptation to dynamic conditions within and in the environment in which the entire system operates. Some of the main aspects and benefits, associated with the management of complexity, through the use of autonomous system constructs for network automation, from an end-to-end system perspective, include the following:

- Self-CHOP network automation of various tasks and operations, through the elimination of manual tasks, and an avoidance of human errors, while providing a consistent and reliable configuration across the network.
- Optimisation of system-wide resource allocation, such as bandwidth allocation, routing decisions, and system capacity planning.
- Performance enhancement, in terms of adaptability to service demands, reliability, availability, resource utilisation efficiency, energy efficiency, service experience etc.
- Cost reduction through an automation of system operations and maintenance, including staffing optimisation with relevant skills, fault isolation, resilience, and mitigation.
- Continuous improvement through an automation of network configuration, integration, upgrades, and service experience. This facilitates a faster deployment of new services, enables rapid testing and validation, and supports iterative development and deployment cycles.
- Personalisation and service innovation, through an adaptation of the network to satisfy individual user requirements. This allows for customised service provisioning, user-specific configurations, and tailored experiences, enhancing customer satisfaction, and driving service innovation
- Cognitive capabilities imbued by AI/ML algorithms that enable self-CHOP behaviours of the system.
- Commercial beneficiaries consist of diverse stakeholders, including Network Service Providers (NSPs), Service Providers (SPs), and end-users.

NSPs and SPs can achieve operational efficiency, accelerate service delivery, and

explore new business opportunities.

End-users benefit from improved service quality, faster response times, and personalised experiences.

As a result of the diverse and interdependent characteristics inherent in emerging systems, the management of complexity is also a significant requirement for sustainable system behaviours. The growth in complexity follows a continuing advancement of system features, and capabilities, which are necessary to meet the diverse demands of continuing service innovation. The management of complexity is an indispensable requirement as networks, services, and devices continue to evolve, where a self-adaptive operational optimisation of system performance, scalability, and sustainability, enable a personalised service experience.

4.2 NETWORK EVOLUTION

Network evolution refers to the continuous improvement and adaptation of networks to meet the requirements of evolving technologies and services. This includes improvements in terms of ubiquitous coverage, adaptable coverage, flexible resource allocation, diverse types of access, diverse deployment scenarios, decentralisation, and distribution, together with the emerging demands of service innovation.

The transformation of the architectural model of an emerging end-to-end system, based on a virtualised service-based architecture, is subject to continuing advancements that facilitate higher levels of granularity and flexibility to meet the rising demands of diverse latency, data rates, coverage, capacity, and the quality of service experience requirements. This naturally entails a continuous accrual of complexity, which follows this arc of increasingly sophisticated system wide capabilities.

Consequently, an effective method for managing increasing levels of complexity is offered through autonomous system constructs, astutely distributed throughout an evolving E2E system, which facilitates zero-touch [6] operation, and a sustenance of system wide equilibrium, while dynamically adapting to its operating environment, throughout the lifecycle of the system. The prominent aspects of network evolution span:

- Diverse access modalities (e.g., terrestrial, non-terrestrial access, licensed and unlicensed spectrum)
- Decentralised and distributed architectural models, with cloud-native functions
- AI/ML native functions, with virtualised and network sliced allocation of networking, computing, and storage resources.

The abstraction of the networking, computing, and storage resources, in a virtualised service-based architectural model, provides a flexible slicing [7] [8] of the E2E system resources, across the radio, transport, core, management, and application layers, as a prominent building block capability of network evolution through the next-generation. A logical partitioning of the E2E system resources into a customisable and isolated network slices provides a virtualised building block, with varying levels of granularity to be effectively leveraged by an overlay of autonomous system constructs. The autonomous system constructs imbue self-CHOP characteristics to manage complexity, while enabling an intelligent management and orchestration of the requisite network slices to suit service demands dynamically and automatically, with optimised resource utilisation.

4.2.1 NETWORK OPTIMISATION AND SCALABILITY

An enhancement of performance, and resource utilisation efficiency facilitate the realisation of network optimisation and scalability of the E2E system. This includes a self-adaptive automation of techniques, such as traffic engineering, load balancing, and diverse service quality demands. Scalability is promoted through interoperable and extensible system design capabilities that accommodate increasing traffic demands that follow emerging system demands.

Autonomous system constructs augment these attributes, through self-adaptive automation, for dynamically adjusting the requisite network parameters, based on real-time conditions, while flexibly ensuring an optimised and sustainable system performance. The AI/ML models, which are embedded within autonomous system constructs, facilitate an optimisation of network resource utilisation, such as radio channel bandwidth allocation, networking, computing, and storage allocation, traffic routing etc., ensuring both an efficient and scalable allocation of system resources.

The network slicing of resources in the E2E system promotes a requisite granularity of customisable, optimised, and scalable allocation of resources to suit the diverse demands of emerging services, driven by innovative usage scenarios, over a service-based autonomous system framework.

4.2.2 OPERATIONAL OPTIMISATION

The streamlining of the network and system operations, requires an effective management of complexity offered through an E2E application of autonomous system constructs, for a self-adaptive and zero-touch automation of the network operations, for the given system environment conditions, service requirements, and intended system performance objectives.

The network operational objectives include a satisfaction of requirements such as device provisioning, software upgrades, configuration management, billing and charging models, minimisation of faults, and an avoidance or minimisation of human intervention for system operations. An adoption of autonomous system constructs throughout the E2E system, promotes these objectives, including a reduction of time to market, improvement of sustainability and evolutionary extensions, reduction of the total cost of ownership, and advancements in the efficiency and look-ahead capabilities of system management and orchestration for effective anomaly detection and prevision.

The cognitive functional capabilities imbued within autonomous system constructs, enable a plurality of operational scenarios, to flexibly suit business and deployment objectives, which yield an E2E system transformation towards the support of innovative use cases, together with context-aware and situation-aware self-adaptive capabilities. These directions promote network programmability, energy efficiency, intelligence at the network edges for an augmented service experience, high availability, and advanced reliability, in highly disaggregated network configurations, with high levels of autonomy and zero-touch automation.

4.2.3 FULFILMENT AND ASSURANCE

The management and orchestration subsystem within the E2E system, imbued with autonomous system constructs, self-adaptively manages the cognitive network functions, embedded with requisite AI/ML models, and the system wide resource allocation. The closed-loop and self-adaptive nature of autonomous system constructs facilitate the network slice orchestration with a set of objectives. The main objectives include the provisioning of the network slice, performance assurance and fulfilment, intent realisation, automated Service Level Agreement(SLA) monitoring, alignment with expected KPIs, Quality of Service (QoS) etc. In this regard fulfilment refers to the process of delivering customisable services to customers, including service activation, configuration, and provisioning. On the other hand, assurance refers to monitoring, troubleshooting, and maintenance of service quality.

Autonomous system constructs promote intelligent and self-adaptive fulfilment and assurance processes, which are self-adaptively optimised to ensure efficient service delivery and timely fault resolution. Zero-touch automation simplifies service fulfilment processes, through an autonomous automation of service provisioning, activation, and configuration, ensuring faster service delivery and a sustainable improvement of customer satisfaction. The autonomous

system constructs, promote a proactive fault detection, anomaly detection, and performance monitoring, allowing for a rapid fault resolution and a proactive assurance of service quality

4.2.4 NETWORK AS A SERVICE (NAAS)

Network as a Service (NaaS) is an approach that provides network infrastructure and services on-demand. It allows organisations to access and use network resources, such as bandwidth, routing, and security, without owning or managing the underlying infrastructure. NaaS offers flexibility, scalability, and cost-effectiveness.

The service-based architectural framework decouples the software from the underlying hardware, together with a decoupling of the functionality from the location, promoting flexible deployment strategies to suit diverse business and deployment models. Beyond this useful abstraction enabled by virtualisation, network programmability is a significant inherent capability, where the control, data, and management planes can be independently scaled and programmed to optimise system-wide operations, and network evolution. Virtualisation consists of both Network Functions Virtualisation (NFV) [9] and Software Defined Networking (SDN) [10], which enable the following benefits:

- Autonomous framework mediated control plane and data plane
- Decoupling and programmability of control and data planes, across the network and devices
- Agnostic behaviours between the network infrastructure and the application.

Automation enables the dynamic provisioning and management of network services, allowing users to easily access and consume network resources as a service. AI/ML algorithms assist in resource allocation and scheduling, for optimising service delivery and ensuring efficient utilisation of network resources.

4.2.5 NETWORK SIMPLIFICATION

The embodiment of AI/ML modalities, together with network programmability (e.g., network slicing), promotes improvements in an autonomous system for managing complexity. This in turn yields enhancements in network reliability, stability, scalability, and sustainability, in terms of fault recovery and a faster convergence towards the intended performance targets and service experience expectations.

In other words, network evolution and adaptability to changing conditions, with respect to operational, business, and deployment aspects are simplified through the self-CHOP behaviours offered by an autonomous system. These self-CHOP behaviours offered by an autonomous system facilitate a simplification of network evolution in terms of optimising configurations, anomaly prevision/detection, and dynamically adjusting to growth/modifications, while promoting avenues for security and privacy. These directions lead to a simplification of architectural considerations, for realising a more efficient routing and resource utilisation, as well as for ensuring network adaptability to optimal information pathways, resulting in an overall network performance and efficiency enhancement.

4.3 SERVICE EVOLUTION

An application of autonomic principles facilitates a self-adaptive evolution of innovative services, aligned with configured intents in the autonomous system, as it interacts with its operating environment. At the same time, the autonomous system dynamically adjusts to meet the requisite user-centric service demands, characterised in terms of KPIs and KVIs. This facilitates a continuing enhancement of interdisciplinary services that span diverse areas of personalisation and optimisation of user experience.

Examples of a variety of innovative services, characterised in terms of user-centric Quality of Experience (QoE), and application-specific QoS, include, while not limited to the following examples of broad categories:

- Telemedicine, e-Health
- Smart agriculture
- Smart industry
- Smart city
- Sensing and communications
- Intelligent transportation

The autonomous system operates over a virtualised service-based framework, which provides flexibility in terms of decentralised and distributed system deployment arrangements. These assorted choices of deployment arrangements, together with advanced intelligent devices (e.g., extended reality, wearables), provide a framework for service evolution. Cloud-native functions render functional portability, in concert with network slicing to virtually carve out and isolate end-to-end system resources (e.g., networking, spectrum, computing, and storage resources).

4.3.1 CUSTOMISABLE AND PERSONALISED SERVICES

The automation of customisation and personalisation hinging on system intents and user-centric preferences, enable flexible and granular choices to craft service features, configurations, and experiential orientations (e.g., satisfaction of QoE, KPIs, and KVIs). The application of autonomous system constructs is essential for automatically satisfying these diverse requirements, while simultaneously minimising the potential adverse impacts of faults or other impairments in the system, and allowing for a graceful degradation, and service recovery.

These benefits span a wide variety of business and deployment models, since the distributed and cooperative autonomous system constructs have the quality of self-sovereignty, while promoting cooperation, which naturally inhibits a spread of fault impacts. The embedded AI/ML models in the autonomous system constructs, which consist of autonomic functions, perform the requisite analysis of user data and behaviour, to align with personalised recommendations, and to augment the user experience.

4.3.2 ZERO-TOUCH OPTIMISATION

The self-CHOP characteristics offered through autonomous system constructs, distributed across the end-to-end system, enable a resource optimised and zero-touch system wide automation. This enables an efficient utilisation and dynamic adaptation to the changing network demands in real-time.

The AI/ML models within the distributed autonomous system constructs, provide an intelligent analysis of the network and system data to dynamically adjust the service parameters, while optimising resource utilisation and sustaining an intended service performance.

4.3.3 EXTENDED REALITY (XR)

An Extended Reality (XR) service encompasses technologies that include Virtual Reality (VR), Augmented Reality (AR), and Mixed Reality (MR). The evolution of this class of services consists of enhancements of XR service experience, which is characterised in terms of demanding performance objectives, such as high reliability, low-latency, high bandwidth, an efficient delivery of content, sometimes under resource constrained conditions (e.g., device form factor, processing capabilities, battery life, number of radio receivers etc.). Immersive and interactive behaviours of XR services are expected to meet the required QoE (e.g., KPI, KVI)

The AI/ML models within the distributed autonomous system constructs, assist an XR service to meet the desired performance and experiential objectives that yield an attractive and immersive experience. The underlying self-adaptive system-wide automation fabric is pivotal for reliable connectivity and efficient real-time communications that promote and sustain an enhanced user-centric XR service experience.

4.4 DEVICE EVOLUTION

The advancement of device functionality, diversity, form-factors, and capabilities, served by an emerging network and system, are pivotal for an enablement of service evolution and experience. These compelling directions are an intrinsic motivator for managing the associated rising levels of complexity, both at the device and at the system level, within which the device is designed to function and operate effectively.

As the system wide complexity proportionately follows these device advancements, to suit the increasing levels of service sophistication, including AI/ML model transfer , sensing and communication capabilities etc., an application of self-CHOP autonomic principles, throughout the system is a pivotal requirement to meet the requisite KPIs.

The self-CHOP oriented automation simplifies the device management processes, such as device configuration, monitoring, and maintenance, by minimising or avoiding human intervention. This augments the consistency of device operation behaviours, while ensuring the objectives of compatibility, interoperability, privacy, and security of the devices to align with the evolving network and system requirements. The requisite AI/ML models, within the autonomous system framework cooperate to analyse device data to identify any performance degradation, detect anomalies, predict impending anomalies, optimise device configurations, and to enhance the effectiveness of device management.

05 AUTONOMOUS SYSTEM MANAGEMENT AND ORCHESTRATION

The emerging, distinct, and divergent requirements of 5G Advanced system and beyond services, spanning, the broad categories of eMBB, mIoT, and URLLC, consist of foundational design principles and technologies. The prominent design principles that are intrinsic to a service-based framework imbue the beneficial attributes of agility, and flexibility, together with distributed cloud-native functionality. These design principles leverage architectural constructs, such as network slicing, decoupling of the control plane and user plane, Software Defined Networking (SDN), Network Functions Virtualisation (NFV), distributed edge computing, connectivity convergence (e.g., terrestrial, non-terrestrial, diverse spectrum access etc.).

These evolutionary directions imply a corresponding demand for continuing improvements in system-wide performance, energy efficiency, and resource utilisation efficiency. Consequently, these aspects contribute to a corresponding rise in system wide complexity, which underscore the significance of requiring a self-adaptive system-wide capability. This type of self-adaptive automation can be realised through autonomous management and orchestration, which is anticipated to advance with an evolution of system and services, to effectively and efficiently manage both system-wide scale and complexity. System-wide self-adaptability implies a dynamic and context-aware automation of system capabilities to realise anomaly detection, intent-based behaviours, and anomaly prevision, while satisfying system performance objectives, and diverse service demands.

A harnessing of autonomic principles, to align with customisable business objectives, based on initial conditions, such as system-wide and network operational goals, intents, policies, and configuration information, promotes self-CHOP capabilities. This promotes a realisation of zero-touch automated operations, which in turn facilitate autonomous management and orchestration. The level of autonomic behaviours spans from lower levels of automation with some human intervention towards zero-touch automation, which characterises a completely autonomous system.

The self-adaptability and self-optimisation characteristics of an end-to-end system, including user equipment, demand a distribution of AI/ML modalities that yield distributed intelligence for an efficient and predictive allocation of networking, computing, and storage resources, to satisfy system performance and service demand objectives. Autonomous system management and orchestration play a significant role for the realisation of these objectives. Contextual self-adaptability of the end-to-end system leverages closed feedback loops that utilise the system inputs consisting of a monitoring and analysis of information to formulate decisions and actions to yield zero-touch automation.

This process facilitates a continuous self-adaptation of the end-to-end system for optimised behaviours, while operating in a given dynamic and changing environment. The autonomous nature of this process is constrained by intents and policies, where AI/ML modalities (e.g., discriminative, and generative artificial intelligence) serve as a catalyst for intelligent management and orchestration for a realisation of zero-touch automation. The discriminative AI/ML modalities harness the conventional

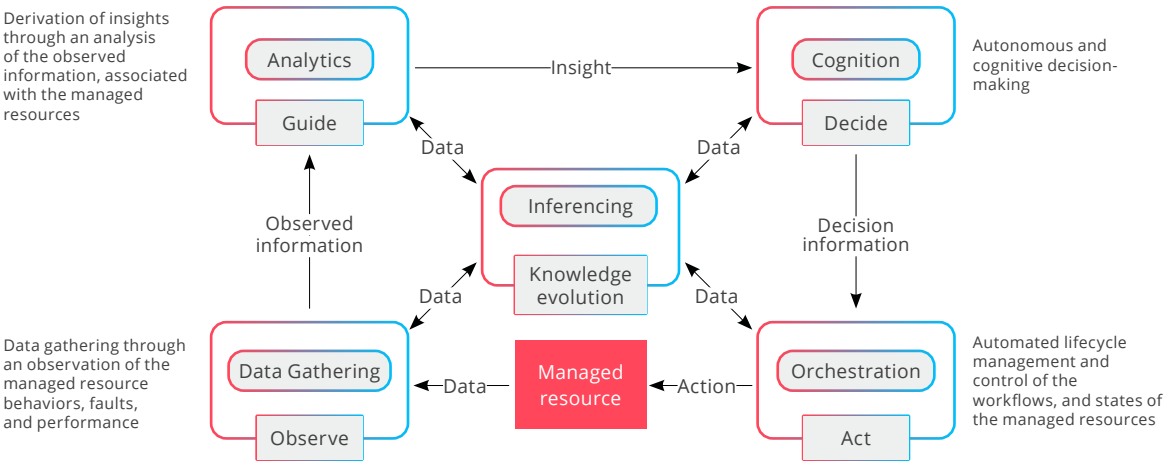

Fig. 2: Autonomous management and orchestration context

methods of AI/ML, such as supervised, unsupervised, and reinforcement learning methods, while the generative AI/ML modalities create new content, such as through the use of Large Language Models (LLMs), with appropriate relevance to a given domain (e.g., telecommunications etc.).

Fig. 2 , illustrates the context of a closed-loop and automated decision-making process for yielding self-adaptive end-to-end system behaviours. The embedded and diverse AI/ML modalities augment the performance, scalability, resource utilisation efficiency, complexity management and fault resilience of the system, while rendering a sustainable and personalisable service experience. Within the context of Fig. 2, an illustrative representation of Machine Learning/Deep Learning (ML/DL) and update process and revision, based on changing conditions within the system as it operates in a given environment, is shown in Fig. 3.

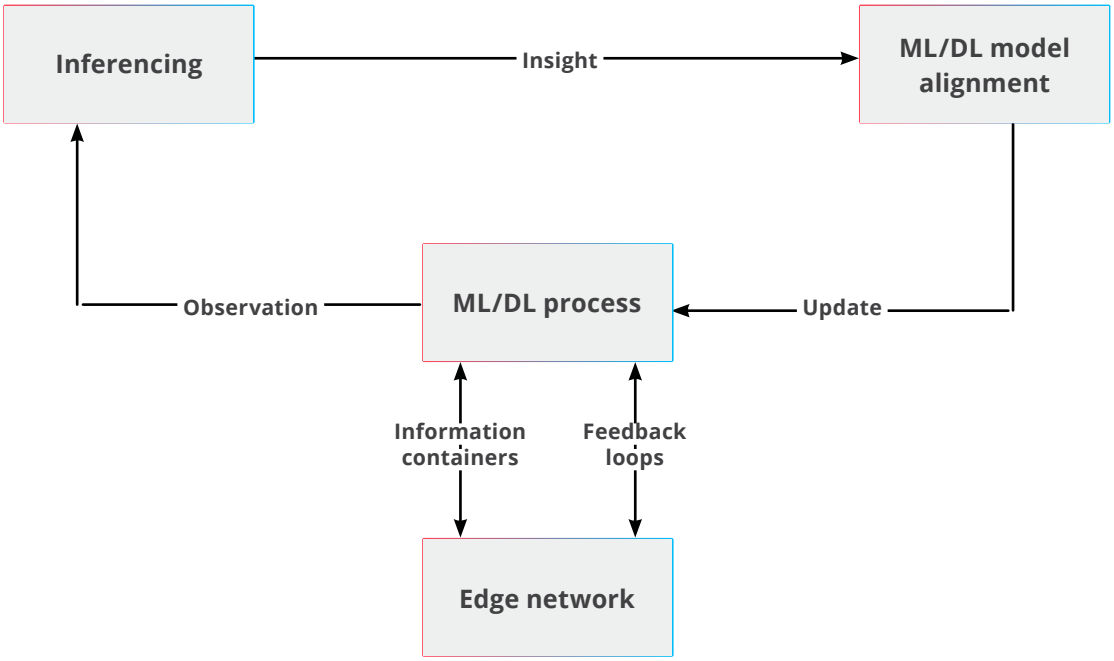

Fig. 3: Example of ML/DL update process and environment

5.1 AUTONOMOUS SYSTEM AUGMENTATION

The self-adaptive characteristics of an autonomous system, hinge on the integrity of a model representation of the system in terms of the intents, policies, business, and deployment objectives that satisfy the desired system performance and service experience goals. However, the quality of self-adaptation of the system depends on the accuracy of the closed-loop model representation that reflects the system requirements and goals. This implies that any gaps in the system goals, as a result of changing system operation conditions, or incomplete requirements, would need to be filled, so that the system model is continuously tuned to sustain self-adaptive behaviours.

Both Discriminative AI (DAI) and Generative AI (GAI) imbued functions that are utilised in an end-to-end manner across the system, augment the self-CHOP characteristics of autonomous behaviours of the system. GAI harnesses Large Language Models (LLMs) that leverage deep learning schemes to generate natural language content, based on an inference of context, structure, and semantics for a given dataset [11]. The various deep learning schemes include Recurrent Neural Networks (RNNs), transformers (e.g., GPT, BERT etc.) [12], Long Short-Term Memory (LSTM) networks etc.

Attention schemes are leveraged in a transformer, which has exhibited an efficient performance with limited training, using domain specific datasets, which implies its potential to be tuned to the requirements of a given autonomous telecommunications deployment configuration to enable self-CHOP operations.

The adaptability of LLMs contribute to their relevance and applicability to enhance autonomic behaviours for system-wide automation and resource optimisation across any domain, including complex telecommunication systems, as well as a plethora of emerging and innovative use cases. The use of Variational Auto Encoders (VAEs) [13] within an LLM facilitates an associated deep learning scheme for removing noise and anomalies in the content generation process.

Discriminative and generative intelligence, are broadly differentiated as shown below:

- **DISCRIMINATIVE:**

  ML algorithms for modelling and classification of datasets.

- **GENERATIVE:**

  Deep ML model trained on datasets for creating new content, based on GANs (Generative Adversarial Networks), VAE (Variational Auto Encoders) and Transformers.

The leveraging of both DAI and GAI is pivotal for a continuing advancement of autonomous system management and orchestration. The transformer-based models include LLMs, which are trained on large volumes of unstructured data to enable an LLM to perform a variety of tasks, including summarisation, classification, translation, prediction, content generation etc., based on a corresponding query to invoke a response from the LLM. A few commercialised examples of LLMs include Google Gemini, Open AI GPT etc.

The use of generative AI, while beneficial in terms of augmenting human creativity through automation, has challenges that impair the relevance and accuracy of generated content, resulting from hallucinations, training biases, and ethical aspects. Additional challenges with generated content pertain to potentially malicious prompts to a generative AI model that exposes confidential information, intellectual property information, and adverse impacts to privacy aspects. These challenges require to be effectively mitigated through an embodiment of curated, reliable data sets, applicable to specific business and deployment domains, with appropriate levels of security and privacy.

An integration of Large Language Models (LLMs) and Generative AI (GAI) with the autonomous system management and orchestration provides the wherewithal for an augmentation of self-CHOP characteristics of an end-to-end autonomous system for zero-touch automation. Both LLM and GAI serve as cooperative technologies for an end-to-end autonomous system, to effectively manage and scale a continuing rise in system-wide complexity, as the ecosystem continues to evolve.

5.1.1 LARGE LANGUAGE MODEL (LLM)

The advent of LLMs, which is a foundation for GAI is poised to have a profound impact on enhancing system efficiencies through autonomous automation, in next generation telecommunication systems, and across various inter-disciplinary service domains. The conventional categories of artificial intelligence modalities that are exclusively based on algorithmic methods (e.g., supervised, unsupervised, reinforcement learning, deep learning, time series etc.) are referred to as the methods of DAI which discern specific and rules-based patterns in the observed and relevant data sets for deriving inferences and optimised actions. While DAI continues to be applied successfully across various domains, in terms of classical machine learning characteristics, such as clustering, classification, regression, reducing the data set dimensionality for improved performance, and pattern recognition, it is deficient with respect to interpreting nuances in the observed data sets, which could be significant in complex and practical scenarios, such as in telecommunication systems.

The advent of GAI harnesses LLMs embedded within an autonomous system, through the use of natural language processing interfaces, where the data sets and prompts are tuned to suit an intended set of system behaviours and deployments choices [14]. The applicability of LLMs is significant across different aspects of the system, such as in the case of network design, fault diagnosis, network configuration, network security. Queries, through the use of prompts are relevant to configure and tune the network. As an example, a query for network configuration could be presented, via a prompt to the system, such as:

"There is a new 5G base station deployed and equipped with a massive MIMO 64T/64R antenna. It covers the Golf National in Saint-Quentin-en-Yvelines for the Paris 2024 Olympics, with a given identifier. Please set its engineering parameters and configure and interconnect the 5G base station appropriately"

LLMs are poised to augment autonomous automation, across diverse network aspects and operations, such as Construction, Planning, Optimisation, Maintenance, and Resource Management. The promise of GAI that utilises telecommunication specific LLMs dwells in the arena of enhancements, such as in the wireless communications system design, planning, and optimisation, such as for deriving insights and corresponding interactions for optimisation, based on an understanding of dynamic wireless link conditions, propagation characteristics, requisite signal power adjustments etc. As compared to more general purpose LLMs, the telecommunications arena is expected to utilise different sizes of language models, such as Small Language Models (SLMs) (e.g., ~ 1 Billion parameters), Medium Language Models (MLMs) (e.g., ~10 Billion parameters), which are relatively much smaller than an LLM(e.g.,~ 100 Billion parameters) which is also more complex, computationally demanding, with increased costs around its training and deployment. Consequently, SLMs and MLMs are especially suitable at the distributed network edges of an autonomous system, where the size of the network and resources are limited.

A few examples of an LLM augmented segments of an autonomous system, include:

- Service Provisioning, where network operators would be able to seek information on the identification of valuable business areas for Fixed Wireless Access (FWA) end users.
- Maximise the value of multiple indicators (e.g. KPIs) such as an improvement in energy efficiency while maintaining an overall QoE.

- Enablement of text-based or voice-based network troubleshooting queries, such as: "*What are the reasons for the poor network quality offered to cloud video streaming services?*

Examples of prominent instruments of GAI, within an autonomous system for advancing and tuning self-CHOP qualities, leveraging different types of capabilities include:

- **CHATBOT**

- Focus on conversations and information retrieval
- Capabilities to answer FAQs, provide customer support, and collect data.
- Decision-making based on responses

- **COPILOT**

- Focus on collaboration and assistance with specific tasks.
- Capabilities to generate content suggestions, translate languages, answer complex questions, and offer feedback.
- Suggestive decision-making recommendations, providing options and insights, while enabling the network operator to make substantive decisions

- **AGENTS**

- Focus on autonomous learning and actions
- Capabilities to make decisions, perform tasks, adapt to situations, and interact within a given autonomous system operating environment, with situational awareness
- Independent decision-making, based in both DAI

Specialised LLMs, utilised by GAI, provide value added capabilities for a network operator by appropriately tuning and optimising the performance of an autonomous system for zero-touch automation. While LLMs provide these benefits for general use cases, such as realising improvements in the overall customer experience, it lacks the specificity for specific use cases, such as for a Radio Access Network (RAN) and related operations that align with related performance objectives. In such specific use cases that embody the complexities of a telecommunication system, the relevant LLM (e.g., SLM, MLM) requires to be specialised as well, where it contains highly accurate and requirements aligned language models, which are trained on curated and network-specific private data, for sustaining the intended autonomous system behaviours. This requirement stems from a recognition of the following broad limitations of generic LLMs, in the context of specific uses, such as for the RAN:

- Sensitive subscriber data and network information, which cannot be sent out of the organisation for model training (e.g., generic LLM models may have implicit biases in the training data)
- Unique network settings and business objectives that are not available in the data sets used in the training of a generic LLM model (e.g., generic LLMs lack knowledge, specific to a given network deployment scenario)
- High cost of energy usage and network resources needed for uploading of huge amount of data (e.g., OAM data and log files)

A leveraging of foundation models (e.g., LLMs), by network operators, provides an initial model, obtained from the open-source community, which obviates the processing of large amounts of data from scratch, together with the high costs incurred as part of the model training. Generic LLMs are important since they have already been trained with data from standardised and general network architectures. Subsequently, these foundation models can be tuned through the use of customised network-specific data sets, to create a relevant suite of language models (e.g., SLM, MLM), for a variety of specific use cases. These newly trained language models can iteratively be trained and updated for both consistent and accurate behaviours, through prompt engineering.

These continually updated language models are fine-tuned and curated with operational network data to produce customised, internal language models to align with the specific requirements of a given network, which results in an overall advancement of the autonomous system behaviours.

5.1.2 FOUNDATION MODELS

A foundation model or a base model is one that is pre-trained on exceptionally large and generic data sets, leveraging a variety of classical, deep learning, and Transfer Learning (TL) AI/ML techniques. The scale of their applicability results in emergent characteristics, where a clear understanding of how they work is limited. As a result of the scale of the applicability of the foundation models [15], homogenisation is incentivised, where advances in a limited set of foundation models are adapted across all other foundation models. While homogenisation has benefits, there is also a risk for inheriting anomalies.

The application of a foundation model serves as an initial fabric, which is subject to further tuning, in alignment with a given deployment data set, intents, and requirements, including network elements (e.g., routers, switches, base stations etc.), within an end-to-end autonomous framework. Multiple foundation models for the different segments of an end-to-end system (e.g., core network, edge network, radio network, user equipment etc.) could be applied as an initial fabric for further tuning and adaptation, through prompt engineering, to align with deployment specific data sets, and requirements, for improved specific language model (e.g., SLM, MLM) behaviours and accuracy.

5.1.2.1 UNDERSTANDING FOUNDATION MODELS

- **PRE-TRAINING**

- **Data and Training:** Foundation models or base models are initially trained on large, diverse datasets where the knowledge and learnings are provided from different sources like device manuals, standards, release notes, administrative guides etc. This pre-training phase allows the model to learn a wide range of features and patterns from the data provided from several such sources of documentation (public and generic) to deploy a network or part of a network in a generic way, which can include anything from vendor equipment, initial network configuration and monitoring aspects.

- **Objective:** The goal is to develop a model that has a general understanding of a generic NSP network before it's fine-tuned for more specific tasks based on the associated business and operational objectives.

- **FEATURES**

- **Generalisation:** Foundation models embody general features that are not specific to any particular network deployment but are useful across a wide range of tasks to be tuned to the specific requirements of a given performed network deployment.

- **Transfer Learning:** Foundation models are designed for transfer learning, where knowledge gained during pre-training is transferred to a new task, with similar requirements, thereby reducing the associated training and computational resources, and hence enabling an improvement in the overall system efficiency.

- **Examples:** Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), and XLNet[16] are examples of foundation models in Natural Language Processing (NLP). These models are pre-trained on a target dataset, which is generic in nature, and can be fine-tuned for tasks or actions related to specific use cases and requirements.

5.1.3 MULTIFACETED LLM

The use of a multi-agent, multi-modal and multi-domain LLM enhances the behaviours of an autonomous system, through learned generalisations and understanding. This learning is leveraged to interpret and generate responses across multiple types of data inputs (e.g., text, images, graphs etc.), as well as across multiple domains, which include network segments, such as the core, transport, access, and edge, including capabilities, such as carrier aggregation and dual connectivity, for coverage and capacity enhancements.

This type of model design approach combines the versatility of multiple LLM agents having diverse multi-modal learning with extensive applicability for multi-domain models, making it significant and valuable for a wide-array of applications and services, to suit diverse deployment scenarios, for realising autonomous system enhancements, in terms of effectiveness and efficiency.

5.1.3.1 MULTI-AGENT CAPABILITIES

By leveraging the diverse capabilities and roles of individual agents within a multi-agent system, an autonomous system can tackle complex tasks through collaboration. With an optimisation of task allocation, robust reasoning is fostered through iterative queries and resolution, management of complex and layered context information, and an enhancement of memory management to support the intricate interactions within multi-agent systems.

Multi-agent system would also enable their application to an existing autonomous system to provide insights on future development and application to distributed autonomous systems (e.g., cloud, heterogeneous access etc.) to suit evolving business and deployment objectives.

The structure of multi-agent systems can be categorised into various types, based on the functionality of each agent and their associated interactions, which is characterised as follows:

- **Equi-Level structure:** In this type of agent structure, different agents operate at the same hierarchical level, where each agent has its role and strategy, but neither holds a hierarchical advantage over another. The agents in such systems can have the same, neutral, or opposing objectives. Agents with the same objectives collaborate towards a common goal without any centralised mediation. The emphasis is on collective decision-making and shared responsibilities. With opposing objectives, the agents negotiate to reach a consensus collectively, or to achieve some conclusion or resolution.
- **Hierarchical structure:** In this type of agent structure, the agents have a hierarchy, where there is a leader agent with one or multiple followers. The role of the leader agent is to plan or guide, while the follower agents respond or execute based on the guidance from the leader agent. Hierarchical structures are prevalent in scenarios, where a centralised coordination is required. In this type of structure, agents make decisions in a sequential order, where the leader agent is the first to generate an output (e.g., instructions or workflow), and a follower agent takes subsequent action, as instructed by the leader agent.
- **Hybrid structure:** In this type of structure, both equi-level and hierarchical structures are accommodated in the same multi-agent system. As a whole the multi-agent system may be arranged in different combinations of equi-level or hierarchical structures to manage complex tasks. These tasks may be broken down into smaller tasks or workflows to form multi-agent subsystems, which may harness other agents to assist with these smaller tasks or workflows.

5.1.3.2 MULTI-MODAL CAPABILITIES

Multi-modal capabilities of models are associated with their ability to process and interpret information from various modalities or types of data. These modalities of data include text, real-time topological data, performance data, and

other types of sensory data. Models with multi-modal capabilities are designed to learn and understand complex relationships among different types of data, which enables them to capture a more comprehensive understanding and improved inferences from the data being analysed.

- **Input processing:** The model can process various types of inputs, including text, such as configuration files, log files, hardware/software data, fault management/performance management graphs. The models could include specialised components or sub-models, which are trained to extract features and understand the content within each modality or type of data being examined. For instance, it might use Convolutional Neural Networks (CNNs) for image analysis, Recurrent Neural Networks (RNNs) [17] or transformers for textual, device and network configuration data.
- **Integration of modalities:** The model integrates the information derived from the different modalities of data after processing the data. The processing could involve a combining of features or representations in a way that allows the model to leverage cross-modal data to enhance the model's understanding of both the context and context of the data, leading to an improvement of interpretation and inference.

5.1.3.3 MULTI-DOMAIN CAPABILITIES

A multi-domain model is designed to process and analyse large volumes of data that span different network segments (e.g., core, transport, radio, edge networks), which may also be associated with different administrative domains. Such multi-domain models harness LLMs that extract information, in terms of configurations, insights, trends in behaviours, and generate relevant reports or feedback, which could be utilised to take decisions, within an autonomous system.

The application of a multi-domain model is particularly useful for intent based network behaviours, as well as for deriving business intelligence, network prevision, and a correlation of observations, through a synthesis of information from diverse sources. The beneficial characteristics of a multi-domain model include:

- **Diverse domain knowledge:** The model has improved inferencing characteristics, realised through the learning acquired from being trained on data pertaining to a wide range of domains. This extensive training enables it to understand and generate more predictable, accurate, and relevant outputs.
- **Adaptability and contextualisation:** The model can adapt its responses or content generation, based on domain-specific content input. It is also versatile in terms of understanding diverse types of data inputs, while adapting and adjusting appropriately to domain-specific terminology, requirements, expectations, and context.

5.2 INTELLIGENT SYSTEM ADAPTABILITY

With a closer integration and processing at the edge, the cloud is enabled to manage larger and more complex tasks with broader scope. This collaborative workflow reduces latencies, saves energy, and gives end users a more seamless experience. It promotes a more powerful and efficient way to distribute GAI workloads, and the related paradigm is expected to continue to grow, together with an edge to cloud integration that will continue to evolve with more capabilities, to manage both complexity and scale, effectively and efficiently.

Prominent aspects be considered for an intelligent management and orchestration of a decentralised and distributed autonomous system, with self-CHOP characteristics include:

- **Automation:** Automating repetitive and routine tasks, such as resource allocation, scaling, or maintenance operations.
- **Adaptability:** The ability to adjust to changing network conditions, user demands, or external factors without manual intervention.
- **Predictive Analysis:** Using historical data and real-time metrics to predict future network conditions, potential failures, or resource requirements.
- **Healing:** The ability to detect faults or failures and automatically take corrective actions to restore normal operations.
- **Optimisation:** Continuously analysing network performance and adjusting to ensure optimal efficiency, speed, and cost-effectiveness.

5.3 SYSTEM OPERATION OPTIMISATION

Autonomic principles that characterise an autonomous system are essential for enabling self-CHOP oriented automation, while also enabling the optimisation of end-to-end system operations. The self-CHOP behaviours of an autonomous system are realised through relevant feedback control loops and AI/ML embedded functionality (e.g., cloud-native functions), distributed throughout the system, which promote both adaptable and predictable system behaviours, promoting the potential for an attractive, sustainable, and evolutionary direction for business, deployment, and operational transformation.

5.3.1 OPERATIONAL TRANSFORMATION

The benefits that accrue from an operational transformation that embodies autonomous self-CHOP automation, include:

- **Efficiency:** Zero-touch operations, facilitating system-wide responses that exceed human response limits, while enhancing system-wide resource utilisation, enabling humans to engage in creative and forward-looking system design.
- **Scalability:** Effective handling of increasing volumes of tasks and data, while managing complexity, enabling businesses to avoid a corresponding linear increase in costs.
- **Consistency:** Uniform system-wide behaviours that align with programmed targets, while adapting to changes within the system and its operating environment, while avoiding/minimising behavioural variability, in terms of both quantitative and qualitative measures, in the absence of human intervention.
- **Error reduction:** An avoidance minimisation of system-wide operational errors, which promote a reduction of human intervention induced operational errors.
- **Proactive behaviour:** Self-adaptive system operations to effectively detect and predict anomalies before system behaviours deteriorate or outages occur, such as through rerouting of network traffic, isolation of fault conditions, preservation of service experience, maximising system availability etc.

5.3.2 IMPLEMENTATION DIRECTIONS

The prominent implementation directions towards operational optimisation may be characterised broadly in terms of the following aspects:

- **Objective definition:** Establishment of specific operational tasks to be automated based on AI/ML models.
- **Data collection:** Gathering of relevant information for training AI/ML models, where the relevance and the quality of information underscores the quality of the AI/ML models for effectively realising system-wide self-CHOP automation behaviours.
- **Model development:** Training and validation of AI/ML models using the collected data, where the model could

be based on classification, regression, clustering etc., or be model-free, such as in reinforcement learning, which learns dynamically, based on specific behavioural targets.

- **Integration:** Embedding of AI/ML models within the operational system or the setting up of Application Programming Interface (API) (e.g., REST etc.) calls to AI/ML models hosted on a server or a cloud platform for enabling self-CHOP automation in the operational system.
- **Scripting and workflow design:** Leveraging of AI/ML model inferences from automated system operations, as inputs into designing scripts and workflows for both anomaly detection and prevision, which could also be automated through other AI/ML models, for automated lifecycle management (e.g., alerts, alarms etc.)
- **Testing:** The testing of AI/ML model driven actions realise expected system-wide self-CHOP automation behaviours in a controlled environment, which is necessary to ensure expected AI/ML model driven system operation behaviours in a production or a deployment environment.
- **Deployment:** The rolling out of self-CHOP automated system operations in a deployment environment, while monitoring the alignment of system operations with expected behavioural targets.
- **Monitoring:** The self-CHOP automated system operations require continuous monitoring of system operations, with inferences provided through feedback loops between the system and its operating environment, for a continuous improvement and refinement of AI/ML models embedded within the system.
- **Iteration:** A continuous update/revision of the system embedded with AI/ML models, which are leveraged for self-CHOP automation of system operations to adapt efficiently and effectively to changes in the operating environment or in the associated business model

5.3.3 OTHER CONSIDERATIONS

A variety of other broad considerations in terms of leveraging self-CHOP system operation optimisation, realised through a system-wide application of autonomous system constructs, include the following:

- **Transparency:** Clarity and understanding for the stakeholders, in terms of the AI/ML model embedded system-wide self-CHOP automation, with respect to impacts on critical business operations.
- **Fallback mechanisms:** Design of mechanisms to manage situations, where AI/ML model embedded system wide operational behaviours deviate below configured confidence thresholds, with respect to expected behavioural targets
- **Ethical aspects:** An establishment of checks and balances, in terms of AI/ML model embedded system operation behaviours and related decision-making processes, where there are potentially adverse impacts to human well-being, societal harmony, threat to life etc.
- **Continuous learning:** An integration of AI/ML model driven self-CHOP automation of system wide operations, while providing significant enhancements, through continuous learning, in terms of managing complexity, scalability, operational and resource utilisation efficiency. This necessitates the architecting of a well-crafted strategy that encompasses both technical and ethical implications, in a holistic and harmonious manner.

06 COOPERATIVE TECHNOLOGIES

To support and enhance the behaviours of an autonomous system, as the evolution of next-generation systems continue beyond 5G Advanced, a few examples of cooperative technologies that embody DAI and GAI are considered. These cooperative technologies leverage access to higher levels of system-wide programmability (e.g., cloud-native functions, IPv6 segment routing [18] etc.) and resource granularity (e.g., decentralised and distributed spectrum, networking, computing, and storage resources), across the system that is available for network slicing. These cooperative technologies serve as enabling capabilities for intelligent connectivity and services, across cyber and physical interfaces for continuing advancements in an autonomous system, towards higher levels of autonomy that characterise self-CHOP behaviours for zero-touch automation.

6.1 NETWORK SLICING

A flexible allocation of networking, computing, and storage resources is realised through network slicing, which is a foundational building block of a service-based architecture that leverages virtualised functions, operating over a shared infrastructure. This functional building block facilitates the creation and instantiation of diverse and multiple logically isolated composition of networking, computing, and storage allocations, to suit the QoS and KPI demands of a variety of services.

The prominent benefits of network slicing include:

- Multiple network functions (e.g., xNF) sharing the same infrastructure for cost optimisation.
- Logical isolation of allocated resources, which in turn facilitates service isolation and adaptability to suit SLAs (Service Level Agreements).
- Flexible and dynamic management, creation, modification, and deletion of networking, computing, and storage resources

With these system-wide benefits, network slicing is foundational in emerging next-generation systems, from an end-to-end perspective. The networking, computing, and storage resource segments, associated with an end-to-end network slice, span the core, edge, transport, and radio networks, together with user equipment. These characteristics of a network slice enable the design of multiple logical networks, with distinct capabilities to suit the requirements of a given service, over a shared physical infrastructure. This attractive flexibility enabled by network slicing, promotes an enablement of a plurality of business models, deployment models, and Verticals to deliver different types of communication services (e.g., video streaming, IoT, URLLC etc.) to consumers, with diverse QoS demands.

The adaptability and the flexibility of network slicing to support the demands of emerging services follows a corresponding rise in complexity. This rise in complexity is an impediment in terms of realising the benefits of network slicing, which are pivotal for realising the enormous opportunities of return on investments, through a satisfaction of service evolution, service lifecycle, and rapid time-to-market demands.

Cognitive and autonomous system capabilities mitigate rising complexity, for the realisation of evolutionary connectivity and service opportunities, through system-wide automation, without human intervention. Autonomous network automation obviates the need for human intervention for system operation, while optimising system performance and operating expenses, together with resource allocation efficiency, energy efficiency, and a satisfaction of system and network KPIs and KVIs.

The management and orchestration of end-to-end network slices, and the lifecycle of network slice instances, including their preparation, commissioning, operation, and decommissioning, are critical for adequately supporting the demands of emerging services. The AI/ML embedded capabilities of cognition and self-CHOP, provide the foundations for customisable zero-touch automation that include the planning of resource capacity, and predicting the required network slice resources, in order to deliver dynamic and self-adaptive network slicing, while establishing a corresponding SLA assurance.

6.2 MACHINE LEARNING OPERATIONS (MLOPS) - STREAMLINING OF ML MODELS

MLOps process provides a valuable approach for the creation, and the maintenance of the quality of the ML models, throughout the lifecycle of ML model deployment and operation. The MLOps process is collaborative, in terms of connecting ML algorithms with business, and operations teams, to accelerate the development and production of ML models, through the practice of Continuous Integration/ Continuous Deployment/Continuous Training (CI/CD/CT).

This helps with an effective lifecycle management of ML models to promote their applicability in an efficient, risk-reduced, and scalable manner. An alignment and compliance of ML models with policies and regulations, promotes an enhanced transparency, together with any drift checks, between expected and actual behaviours of an ML model, or in other words, the consistency or reproducibility of an ML model. The ML models are embedded within a larger software system, which facilitates both access to, and a monitoring of the ML models The MLOps process can be viewed as an extension of the DevOps practices for a rapid deployment f ML models [19]

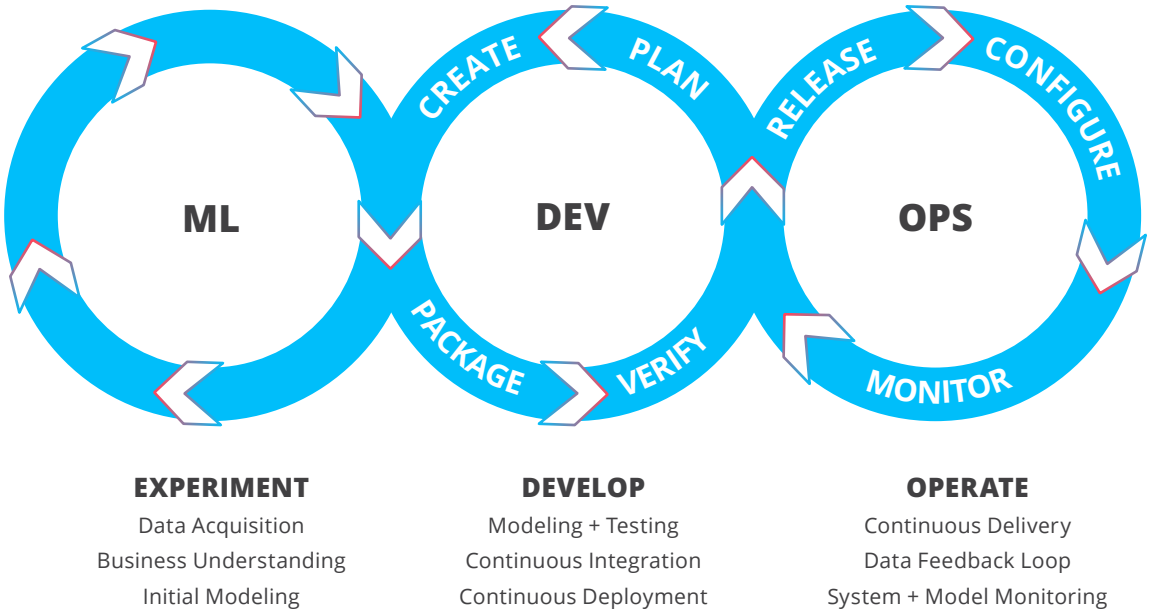

Fig. 4: View of the MLOps process

The MLOps process, consists of the following main functions:

- **Project design:** Includes requirements collection, scenario design, data usability checks etc.
- **Model development:** Includes data engineering, model engineering, evaluation, and verification etc.
- **Model operation:** Includes deployment, CI/CD/CT workflow, monitoring, scheduling etc.

An ML project cycle evolves with the development of AI, and the emergence of MLOps in the industry to complete the ML project lifecycle. MLOps builds and maintains the ML pipelines, around the CI/CD/CT in a full lifecycle, closed-loop system of the associated and connected pipelines. The ML project lifecycle, typically has four stages, consisting of requirements design, development, delivery, and operation, which are decomposed such as, requirements management, data engineering, model development, model delivery, and model operation, as follows:

- **Requirements management:** Feasibility analysis and articulation of technical requirements and solutions, based on business objectives and business requirements.
- **Data Engineering:** Transformation of source data into usable data and storage in a suitable location for utilisation.
- **Model development:** Model development and optimisation in an experimental environment, through the process of model training, parameter tuning, evaluation, and selection.
- **Model delivery:** The model is deployed to a relevant target environment, after the model is packaged with configuration, code, scripts, and generated deliverables.
- **Model Operations:** Provide monitoring and operational maintenance of model services in a production environment.

As the evolution of intelligent network automation progresses towards autonomous systems that embody AI/ML models and the associated algorithms, MLOps serves as an effective approach to enable the realisation of an autonomous system, including intelligent network management, intelligent operation, and intelligent network element management, This approach serves to improve the research and development efficiency of intelligent network applications through a systematic and automated lifecycle management of applications.

MLOps has been identified in the industry [20] as a prominent enabler of an autonomous system, which is pivotal for enabling the various aspects of an autonomous system for network automation, such as:

- Management of the research and development assets of autonomous network applications, including data, AI/ML models, algorithms etc.
- Standardisation of the development process of AI/ML models with a specific focus on application demand analysis and to establish an effective application value analysis system.
- Realisation of a scalable (e.g., large scale) and automatic deployment of autonomous applications.
- Continuous monitoring of the effectiveness of autonomous network applications to avoid the risk of AI/ML model degradation, caused by data drift, while supporting iterative training, and a continuous optimisation of the related AI/ML models.
- Standardisation of the workflow of autonomous network applications from model development to model delivery and operation, while also improving the quality and efficiency of model delivery.

6.3 APPLICABILITY OF GENERATIVE AI

The rising complexity and interdependence, across the various segments (e.g., applications, core network, transport network, radio network, and the user equipment) of an end-to-end system, demands an integration of GAI with the self-organising framework of an autonomous system. This would facilitate an augmentation of the autonomous system, in terms of flexible and dynamic deployment arrangements, with respect to an optimised utilisation of networking, computing, and storage resources.

With this contextual backdrop, it is evident that GAI is poised to play a pivotal role within the autonomous system for the realisation of a sustainable zero-touch end-to-end system, based on self-CHOP behaviours, hinging on a given intent ensemble. LLMs are an integral aspect of GAI, which focusses on system inferred text generation. GAI augments the capabilities of an autonomous system, to enable a compelling Quality of Experience (QoE) in an extensible, efficient, and scalable manner.

Ensemble hybrid models in GAI[21] involve a combination of multiple generative models (foundation models) or techniques to enhance the quality and integrity of the generated content output. Each foundation model, within an ensemble hybrid model in GAI may focus on a different aspect of the data or use different techniques. The hybrid aspect of a generative model consists of both a statistical analysis component and a symbolic or semantic component to interpret meaning and insight from an input data set. The ensemble aspect of a generative model consists of two or more learning models, such as regression models, neural networks etc., to improve the accuracy of predictions, based on an input data set. An ensemble hybrid model in GAI combines their constituent predictions through techniques like averaging, voting, or stacking to produce a final prediction, represented by the generated content output. Consequently, ensemble hybrid models in GAI are applicable within GAI to improve the integrity, quality, and diversity of the generated content, while avoiding or mitigating the impact of inaccurate inferences.

An augmentation of the autonomous system leverages the inferencing capability within GAI, through the use of LLMs, which describe the characteristics, context, and requirements of a given system, in terms of domain-specific knowledge, associated with the system[22]. This enables improvements in a dynamic adaptation of the autonomous system, as changes within the system and its operating environment occur, while preserving the service experience, optimisation of resource utilisation, and the expected end-to-end system KPI and KVI targets.

An example of GAI, leveraging a digital twin, which is a virtual representation of a physical system, for augmenting the behaviours of an autonomous system is depicted in Fig. 5.

System-wide contextual learning is reflected in the associated LLM, to tune the system behaviors to align with the KPI and KVI objectives, over the lifecycle of the system. With these adaptive capabilities, the LLM complements autonomous system constructs, by facilitating agile, flexible, and adaptable workflows, to appropriately support the changing conditions (e.g., traffic load, resource constraints, anomaly detection, anomaly prevision etc.), within the system, and its operating environment. The LLM within an autonomous system, which may also leverage a digital twin, which serves at different and flexible levels of granularity, to facilitate and orchestrate the necessary functions required of a task or a workflow input, within an end-to-end system.

Large volumes of data associated with the system are leveraged for training a corresponding LLM, to understand and generate inferences from this data, based on the specific data patterns and relationships. This enables GAI to predict corresponding responses that can be utilised to advance autonomous system behaviours. In this context GAI and LLM, function in a complementary fashion to promote and advance the operation of an autonomous system, while adapting and optimising the end-to-end system behaviour. This is accomplished through a continuous update of the domain-specific knowledge, utilising the building blocks of AI/ML models subjected to the CI/CD/CT process. There are a variety of LLMs that leverage natural language processing techniques for describing the characteristics of a given system, such as through queries, response retrieval, sentiment evaluation, semantic analysis etc. Examples of LLMs include Midjourney [23] LaMDA [24], ChatGPT [25] etc.

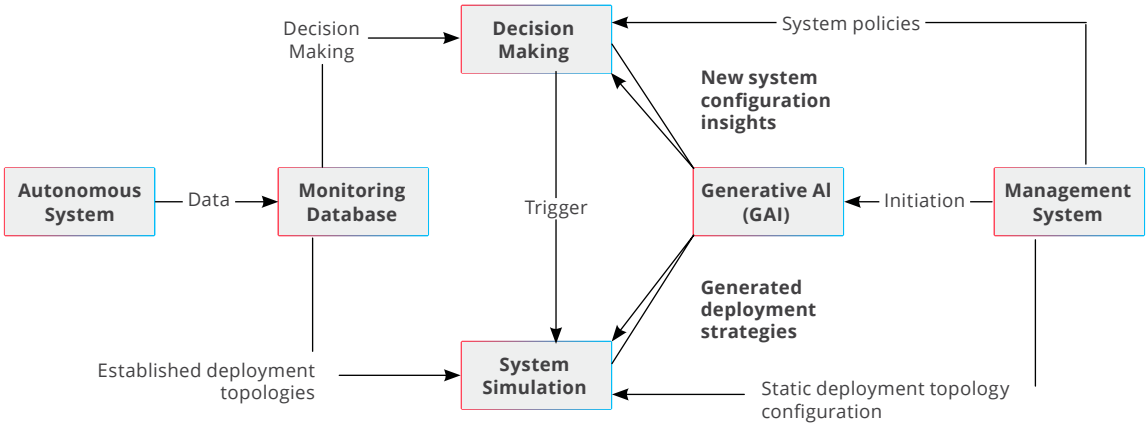

Fig. 5 : Generative AI leveraging a digital twin for autonomous system

6.3.1 AUTONOMOUS SYSTEM ADVANCEMENT - LARGE LANGUAGE MODEL (LLM)

The self-CHOP attributes of an autonomous system hinge on the effectiveness and the fidelity of the underlying algorithms and the quality of available information, in concert with the relevant feedback control loops between the autonomous system and its operating environment to maintain system wide stability, under dynamic conditions. This implies the need for a choice-driven approach, where the relevant LLMs are aligned with specific considerations (e.g., optimum trade-off between energy savings versus QoE) and are domain-specific (e.g., telecommunications). Domain-specific models are based on generic foundation models, together with a continuous integration, deployment, and training of the underlying domain-specific models.

Among the many usage scenarios of LLMs, a prominent usage scenario involves a complementary use of APIs, such as intent-oriented APIs that enable network administrators to interrogate the network segments in an end-to-end system, with chatbots or plain text, obtain a corresponding response from the network segment. Audio assistants can also leverage LLMs to deliver rapid, precise, and complete answers to problems observed within the autonomous system, which self-adaptively corrects the anomaly and establishes system-wide equilibrium. Examples of interrogation include broad questions, such as:

- "Do I have a problem in my network?"
- "Can you show me graphically, where in the system on my 5G network is the downlink speed less than 155Mbps?"
- "What is the most adversely impacted coverage area, or KPI in my network?

A significant requirement for a sustainable self-adaptive equilibrium in an autonomous system is the definition and establishment of requisite open interfaces (e.g., 3GPP, TMF, CAMARA), since the machine-to-machine interactions serve the need to exploit, question, or get answers from the different LLMs, running in different network elements, and network segments (e.g., core, transport, edge, radio, OAM etc.)

The human interfaces provided by LLMs provide the network administrators with convenient interactive capabilities, for interacting with an autonomous system, which exploits the relevant technical documentation, whether it is vendor-specific or standardised. This organically promotes higher efficiencies, in terms of optimising system-wide operations.

6.3.2 NETWORK SERVICE PROVIDER (NSP) TUNED INTELLIGENCE

The autonomous system is required to adhere to the design requirements, and performance objectives that an NSP expects of the system, which would typically be multifaceted, with interdependencies and inter-relationships, across the objectives that underpin the system behaviour. As

complexity rises, the degree of sophistication required, within an autonomous system, to meet the emerging demands of continuing service evolution, the underlying AI/ML models are likely to incur gaps that may result in inadequacies that adversely impact autonomous system behaviour in terms of satisfying the intended objectives. These gaps may consist of an insufficient capture of system requirements, system analysis, or inadequate stakeholder information.

The gaps are likely to accrue with rising complexity, and dynamic shifts in system-wide resource utilisation (e.g., changing traffic conditions, radio network conditions, service evolution, service usage demands etc.). GAI facilitates a promising approach to effectively bridge these gaps, through the use of training data, relevant for the domain, within which the autonomous system operates. GAI leverages specific LLMs relevant for a given domain to model the objectives of the domain with high fidelity, so that semantic errors [26] are avoided in the ML models for the autonomous system, created by the associated Generative AI process. An iterative feedback loop, based on the Monitor–Analyse–Plan–Execute over shared Knowledge (MAPE-K) [27], is an useful approach for the GAI process.

6.3.3 SITUATIONAL AWARENESS

A leveraging of GAI for situational awareness provides an operational team with a deeper and rapid understanding of the associated autonomous environment, for a timely resolution of any issues, which in turn facilitates an optimised system performance. This approach combines real-time data analysis with advanced AI/ML capabilities, to yield actionable insights, while enhancing the overall operational efficiency of an autonomous system. The information associated with situational awareness consists of real-time monitoring and analysis of the end-to-end system metrics, together with the associated data logs to derive a comprehensive understanding of the current system state and recent history. GAI harnesses this process to develop useful and accurate insights to enhance the self-adaptive behaviour of the autonomous system with improved fidelity.

As an example, a brief description of a GAI process in this regard is delineated as follows:

- **Current system health monitoring**

Examples of queries, with respect to current system health monitoring, include:

- "What is the health of wireless coverage in the access environment?

- "What is the health of access experience for a given customer?

- **Data collection**

Examples of data collection, such as telemetry data, include:

- Collection of real-time telemetry data from various sources such as system logs, performance metrics, error logs, and network traffic.

- Continuous collection of data, using agents installed in each relevant environment (e.g., production, staging etc.).

- **Generative AI (GenAI) processing**
Examples of GenAI processing, include:

  > **Natural Language Processing (NLP) engine**

  - Utilisation of NLP to understand user queries

  - Parsing of a query to identify the environment, and the specific metrics of interest

  > **Data aggregation and analysis**

  - Aggregation of collected data from multiple sources (e.g., servers, databases, network devices).

  - Application of machine learning models to analyse the data, identification of patterns and anomalies.

> **Health score calculation:**

- Calculation of a composite health score based on metrics such as CPU usage, memory usage, disk I/O, network latency, error rates, and user experience data.
- Usage of statistical models and historical data to benchmark current performance relative to expected performance.

> **Generative response:**

- Generation of a detailed natural language response, summarising the health status.

- **Inclusion of key metrics, current issues, potential risks, and suggested actions, together with historical Issue analysis**

  Examples of historical issue analysis, include:

> **Historical query examples:**

- "What were issues with access on a specific date?"
- "Were there any agent framework issues on a specific date?"

> **Data collection logs:**

- Retrieval of archived logs, incident reports, and performance data for a specified date and environment.
- Utilisation of a centralised logging system (e.g., ELK stack, Splunk) to store historical data.

> **GAI processing:**

- NLP Engine:
- Understand the user query and extract the data associated with the environment, and specific components of interest.

> **Data mining:**

- Execution of data mining on historical logs to identify error patterns, incident reports, and system alerts.
- Use of clustering algorithms to group similar issues and identify common root causes.

> **Incident correlation:**

- Correlation of incidents across different logs and metrics to provide a comprehensive view of what happened.
- Identification of sequences of events that led to issues using causal inference techniques.

> **Report generation:**

- Generation of a detailed report that summarises the issues that occurred, including timelines, affected components, root causes, and resolutions.
- Provision of insights into recurring issues and recommendation for preventive measures.

- **Agent framework monitoring**

Examples of agent framework monitoring, within GAI, include:

> **Specific Query:**

- "Any agent framework issues in a specific version?"

> **Agent Diagnostics:**

- Enablement of diagnostic capabilities in the environment.
- Collection of detailed diagnostic data from agents, including health status, performance metrics, and error logs.

◊ MONITORING OF GAI PROCESSES

Examples of a monitoring GAI processes, include:

» **NLP engine:**

- Parsing of the query to focus on an agent framework behaviour, within the specified development environment

» **Generation of insights:**

- Generation of insights based on the analysis, highlighting specific issues, their impacts, and possible causes.
- Recommendation of actions to resolve the identified issues and to improve the stability and performance of an agent framework.

» **Diagnostics enablement:**

- Enablement of diagnostic logging and monitoring for all relevant environments.
- Configuration of agents and monitoring tools to collect comprehensive data, including performance metrics, error logs, and system events.
- Implementation of a robust data collection and storage infrastructure to support real-time and historical analysis.

» **Diagnostic data analysis:**

- Analysis of diagnostic data to identify any performance issues, errors, or failures within the agent framework.
- Usage of anomaly detection algorithms to pinpoint unusual behaviours. or deviations from expected performance.

6.3.4 AUTO-DISCOVERY ANALYSIS FOR CROSS-LAYER CORRELATION

Auto-discovery analysis involves an automated identification of relationship and dependencies between the different layers of an autonomous system. This process reveals an understanding of the interconnected and interdependent nature of the various resources of an autonomous system, while also identifying the root causes of anomalies that may span multiple layers of an autonomous system.

Using temporal correlations and domain knowledge, the symptoms and root causes can be effectively grouped. An exemplification of this process includes:

◊ TEMPORAL CORRELATION AND RELATIONSHIP ANALYSIS

Temporal correlation and relationship analysis, together with examples, are shown below:

» **Temporal correlation:**

- Temporal correlation refers to an identification of patterns or events that occur in a time sequence across different system layers. These correlations help in an understanding of how events in one layer affect another over time.

» **Example of temporal correlation**

Scenario: Observation of a spike in the occurrence of database query times.

Observation: The spike in query times is noted at 3:00 PM.

Related Events:

› At 2:55 PM,
a sudden increase in network latency is observed.

› At 2:50 PM,
a high CPU usage alert is triggered on the application server.

› At 2:45 PM,
a deployment of a new application version is recorded.

» **Relationship analysis:**

- Analysis of relationships in terms of potential causal links between correlated events across different system layers.

» **Example of relationship analysis**

Scenario: Analysis based on the example of a temporal correlation shown above.

Inference:

- The congestion in the network could have led to increased network latency.
- The deployment of the new application version might have caused high CPU usage on the application server.
- The increased network latency might have caused the spike in application response time.

◊ DISCOVERING TOPOLOGY RELATIONSHIPS

This entails a discovery of topological relationships, which reveals an understanding of the physical and logical connections between various system components based on correlation analysis.

» Example of discovering topology relationships

Scenario: A microservices architecture with multiple topologically interconnected services.

Topology Discovery:

› Service A (front-end) depends on Service B (API).

› Service B depends on Service C (database).

› Temporal correlation analysis shows that an issue in Service C leads to failures in Service B and subsequently Service A.

› This correlation helps in mapping out the dependencies and understanding the topology of the microservices.

6.3.4.1 AN ILLUSTRATIVE EXAMPLE

An illustrative example of an observed application downtime, with related analysis in terms of correlation, relationships, root cause, and discovery is shown below:

◊ TEMPORAL CORRELATION:

» Observation: Application downtime observed at 4:00 PM.

» Related events:

› At 3:55 PM, a high memory usage alert on the web server.

› At 3:50 PM, increased error rates on the API server.

› At 3:45 PM, a database deadlock detected.

◊ LIKELY RELATIONSHIPS:

» Analysis:

- The database deadlock at 3:45 PM caused API server errors at 3:50 PM.
- The API server errors led to high memory usage on the web server at 3:55 PM.
- The high memory usage resulted in application downtime at 4:00 PM.

◊ SYMPTOM GROUPING AND ROOT CAUSE:

» Grouped Symptoms:

› Database deadlock, API server errors, high memory usage, application downtime.

Root Cause:

› Identified as the database deadlock.

◊ TOPOLOGY RELATIONSHIPS:

» Discovery:

- The web server depends on the API server.
- The API server depends on the database.
- The discovered topology highlights the critical path and dependencies among components.

By leveraging auto-discovery analysis with temporal correlations, domain expertise, and policy enforcement, organisations can effectively identify and understand cross-layer dependencies and root causes. This approach not only enhances situational awareness but also improves incident resolution and the system reliability of an autonomous system.

6.4 SLA MANAGEMENT

SLAs are a list of objectives and accountabilities contractually agreed upon between a service provider and an end-user covering the conditions by which a service is delivered over the life cycle of the contract, facilitated intelligently by an autonomous system for self-adaptation, to suit dynamic changes to an SLA. From a service provider perspective there are two types of SLAs:

- **Service focused:** SLAs are the defined delivery conditions negotiated between a service provider and a large group of customers, and so the same SLA applies for all customers in that group.
- **Customer focused:** SLAs are the defined delivery conditions negotiated between a service provider and a customer to address customer specific requirements. e.g., a trading company requiring unique service delivery conditions, such as extremely low latency etc.

SLAs can cover a variety of measurable metrics such as:

- **Technical:** Latency, availability, uptime, bandwidth, jitter etc.
- **Operational:** Response time, restoration time or Mean Time To Repair (MTTR) etc.

Enterprises, public sector, healthcare, and others large customer segments rely heavily on a service provider's ability to deliver services that conform to mutually agreed conditions and non-compliance with an SLA would incur agreed penalties.

In simple terms, SLA management involves people, processes and systems that assist in ensuring that the services are delivered in compliance with agreed conditions stipulated in the service contract between the service provider and the customer and includes both proactive and reactive activities.

- Proactive activities involve performance management capabilities that constantly monitor service performance. When service delivery conditions are at risk of being impacted adversely, such conditions trigger corrective actions and customer reporting.
- Reactive activities involve the customer reaching out to a helpdesk, the creation of a trouble ticket, the involvement of higher tier support entities when needed for root cause analysis, as well as a restoration of the service and customer reporting.

6.4.1 AUTONOMOUS SLA MANAGEMENT

Appropriate AI/ML algorithms assist many aspects of SLA Management allowing a service provider to deliver higher quality

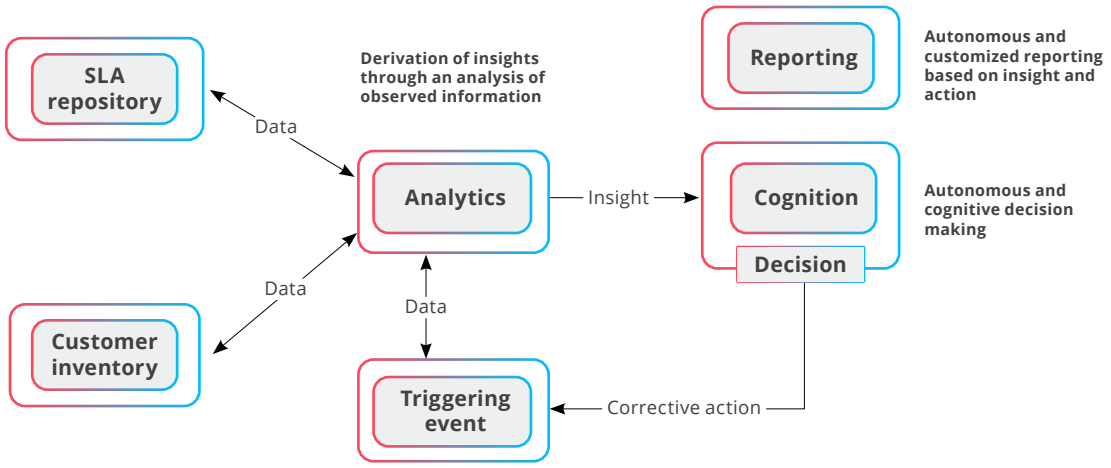

Fig. 6: Autonomous SLA management

services and differentiated offerings. Service providers typically have an SLA repository and a customer inventory system that provides a listing of customers and their corresponding SLA, associated with their service delivery contract.

During the life-cycle management of a service delivered from a service provider to the customer, requisite AI/ML models facilitate a cognitive, adaptive, and automated assistance in terms of correlating a service triggering event (outage or degradation) with a corresponding contractually agreed SLA between the service provider and the customer.

The following is an exemplification sequence of lifecycle management, based on autonomous SLA management, illustrated in Fig. 6:

- A triggering service event occurs with the associated data as input to an AI/ML trained analytics (e.g., NWDAF) engine.
- The analytics engine derives the relevant information from the provided data to query the SLA repository and customer inventory, to identify the customer and the contractually agreed service delivery conditions.
- This information serves as input for a cognitive AL/ML engine that determines an appropriate action that is necessary to update or restore the service back to its normal operating conditions.
- The cognition AL/ML engine provides insights to the reporting system, such as Customer Relationship Management (CRM), which adds this information to its logs and reports for access by the customer. Tier 1 helpdesk also gets the same information so they can keep their conversation with the customer aligned with the information reported to the customer, either proactively or reactively.

6.5 AI/ML MODELS TO SUIT END-TO-END SYSTEM REQUIREMENTS

The cognitive capabilities within an autonomous system are realised through the use of appropriate embedded AI/ML models, operating with closed-loop feedback [1], The training of these AI/ML models occur through iterative process loops, until a convergence towards expected functional behaviours is established. The levels of cognition, realised in a closed-loop feedback arrangement, may be categorised in terms of the following:

- Closed-loop feedback system to meet a target objective based on a fixed data set, which is used to train the AI/ML models, implying limited adaptability.
- Closed-loop feedback system to meet a target objective, under dynamic conditions, where the AI/ML models are learning and training continuously, implying broad adaptability.

These are broad categories of closed-loop feedback arrangements are differentiated in terms of the degree of adaptability to dynamic conditions, which reflect the various layers of artificial intelligence as depicted in Fig. 7.

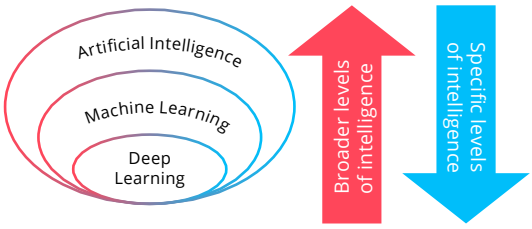

Fig. 7: Layers of intelligence

Autonomous systems leverage both the various layers of intelligence and self-CHOP adaptation of behaviours, through closed-loop constructs, to satisfy system-wide objectives, under dynamic conditions, such as in connectivity and service, for widespread automation, without human mediation. The application, update and training of the models associated with layered intelligence, depicted in Fig. 7, include machine learning and deep learning intelligence models.

Machine learning models are updated through the process of continuous integration, continuous delivery, and continuous training for enhancing cognition, knowledge, and actions, based on dynamic data and experience, drawing from interdisciplinary fields (e.g., control system, neuroscience etc.). Deep learning is a category within machine learning that harnesses artificial neural networks, akin to a hierarchical arrangement of neurons in the human brain.

6.5.1 ELEMENTS OF COGNITION AND ADAPTABILITY

From a training perspective an AI/ML model is subjected to multiple cycles or epochs in a process loop, using a complete dataset [28]. The number of cycles or epochs for an expected AI/ML model behaviour is dependent on the size of a dataset and the required granularity or accuracy of the AI/ML model behaviour. In this context, the number of cycles or epochs of training to which an AI/ML is subjected, is referred to as a "Process Loop".

Dynamic cognition and adaptability underscore the essence of an autonomous system. There are two building block primitives for AI/ML model verification and validation that require consideration:

- Model training through an iterative process loop for alignment with a target objective
- Model drift detection for any deviations in behaviour, during model operation

A two layer approach is considered, leveraging the primitives for AI/ML model verification and validation, where the layer 1 process loop aspects are associated with a realisation of autonomous Network Function (NF) behaviours, in terms of the following:

- Configuration and re-configuration to satisfy the performance requirements, leveraging data inputs
- Presentation of the most relevant information to the OSS (Operations Support System)
- Execution of predictive maintenance tasks.

In Fig. 8, within the context of shared infrastructure resources, Virtual Infrastructure Manager (VIM), and a collection of measurements, associated with an NF, the NF is continuously integrated, delivered, and trained to establish an intended autonomous behaviour (self-CHOP), for a given system and its operating environment. This is accomplished through an NF awareness of the underlying infrastructure, and its neighbouring NFs.

Any number of NFs, each with its own Element Manager (EM), generate a unique collection of measurements, which are used to determine the associated KPI, for alignment with a target objective, where the layer 1 process may be labelled as a resource layer process for the AI/ML model within a given NF.

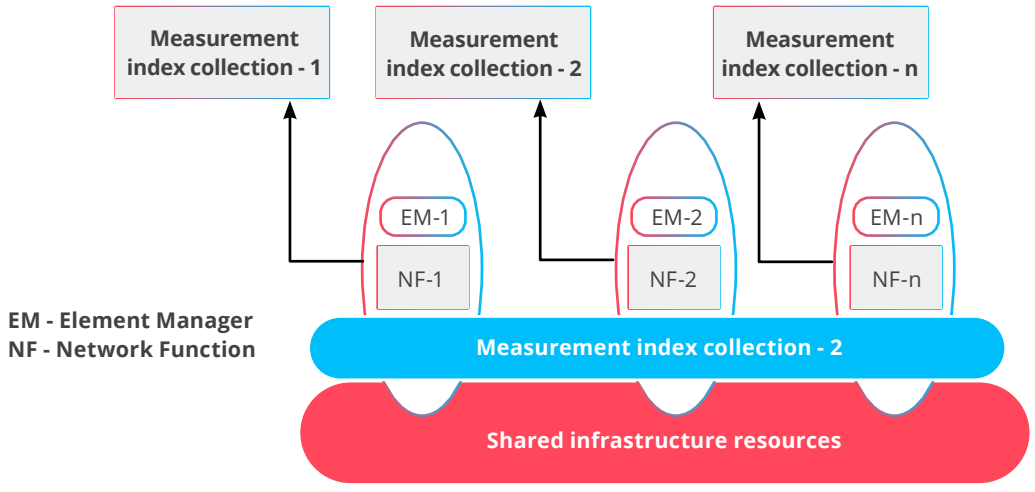

Fig. 8: Process loop for NF AI/ML model updates

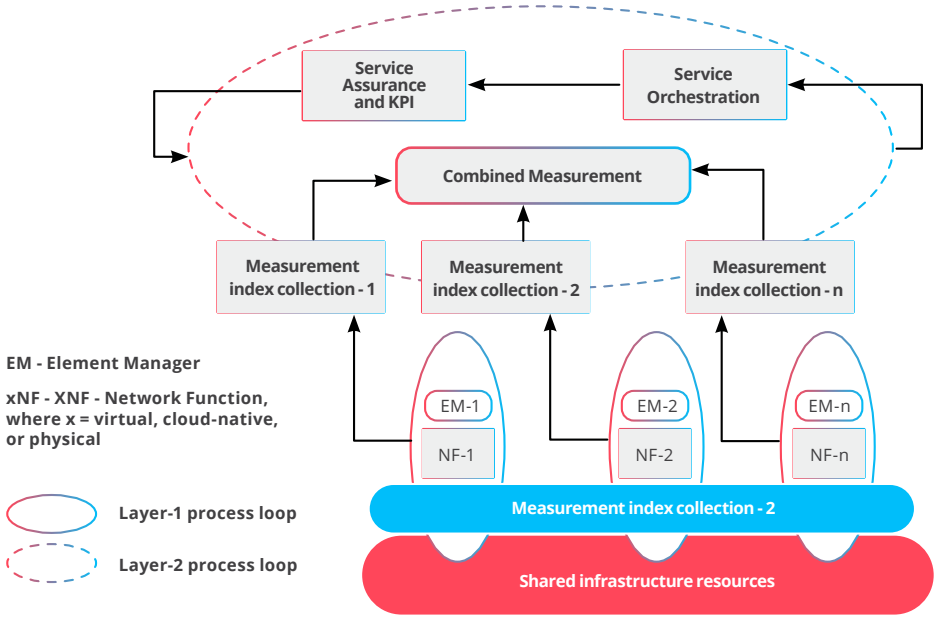

Fig. 9: Layer-1 and Layer-2 process loops for ML model updates

The layer 2 process loop leverages the layer 1 process loop for a combined collection of measurements, obtained from the AI/ML models embedded in each NF for incorporating service management, which consists of both service assurance and service orchestration, as depicted in Fig. 9. The microservices associated with a given service may span multiple NFs. There may be more than a single layer 2 process loop, for example one for each service. The layer 2 process loop may be labelled as a service layer process for the AI/ML model within a given NF.

After an ML model, embedded in the NF, is verified, and validated through the distinct lifecycles of the layer 1 and layer 2 loop processes, the NF is ready to be utilised in a closed-loop architecture, for the realisation of autonomous behaviours that render zero-touch automation. The NFs participate in the composition of a service, where the alignment of the service, with the overall service objectives, is ensured by the service orchestrator, which orchestrates the behaviour of the NFs.

The adaptability of autonomous behaviours of a system or a service follows an “intent” to satisfy performance objectives, in terms of relevant KPIs, and a system or service behaviour. Service assurance and service orchestration deploy and manage the NFs, as well as the VIM, with an appropriate configuration of the underlying shared infrastructure, via resource orchestration.

6.5.2 AI/ML MODEL TRAINING ASPECTS

The various types of AI/ML models that underpin a virtual or cloud native NF (xNF), within an autonomous system, require to be continuously trained, integrated, and delivered, to adapt to dynamically changing conditions within the system and the environment within which it operates.

The continuous cycle of training, integration, and delivery is especially significant for preserving the integrity and relevance of the xNFs and microservices, which in turn mitigates system performance degradation. This is a significant requirement for the system behaviours to sustain compliance with the intended performance objectives, and zero-touch automation, provided by the autonomous system.

The various AI/ML models and their application in the autonomous system that are relevant for managing system wide complexity to satisfy the diverse demands of both the system and services, as they continuously evolve, require to be streamlined in order to sustain the efficiency of managing the overall system complexity. The different and challenging aspects of ML model training may be broadly categorised in terms of the following:

- Complexity: The decentralised, distributed, and disaggregated nature of next generation systems compounds the complexity associated with enabling automation. The levels of complexity increase with increasing levels of interdependence, across the constituents that compose these evolving systems. For example, automating tasks at the infrastructure layer, such as firmware/ driver updates, configurations etc., are likely to impact the virtualised management layer, and in turn impact the vNF/cNF (virtualised NF/cloud-native NF), which execute as part of the virtualised management layer. Furthermore, since emerging systems have various cloud-native topologies (e.g., hybrid cloud etc.), which are realised through microservices, a change in a microservice in one part of the network (e.g., edge network) may have implications in another part of the network (e.g., core network). The scope of such types of impact hinges on the scope of a given service, across the system, where a complex service may contain a large number of microservices.

- Speed of change: The categories of evolving services, such as IoT, enhanced mobile broadband, URLLC etc., provided by a service provider, are rendered over shared, distributed, and interconnected networks. These arrangements of networks have impacts on workload automation, and on the utilisation of the underlying shared infrastructure. As the required speed of updates and adoption of these arrangements increases, to suit associated business and market changes, there is a proportional increase in the speed of updates and adoption of the embedded AI/ML models in the system, to preserve an intended system-wide behaviour.

- Stack utilisation: The utilisation of the stack, for example in a massive MIMO configuration, catering to a variety of patterns of usage, such as the number of users, link conditions, content types, specific service conditions etc., change over time, resulting in a potential degradation of the performance of the associated AI/ML models, in terms of realising sustainable automation. The rate of AI/ML model degradation over time, would be a function of the rate of stack utilisation changes in the system, which would then imply that an appropriate cadence is required for AI/ML model training, integration, and delivery.

A continuous re-training the embedded AI/ ML models, within the system, is required for sustainable automation. The re-training of these AI/ML models is far more elaborate, relative to a prediction of usage patterns or trends of utilisation. As a result, this re-training process is both cost and resource intensive on-site, in terms of frequent AI/ML model training in a cloud environment.

Some considerations to ameliorate the barriers, such as cost and resource demands for a frequent training of AI/ML models, consist of the following:

- Separation of resource from service processes: While the layer 2 service processes contain the overarching ingredient for service automation, they rely on the efficacy of the underlying xNFs. The AI/ML models associated with these underlying xNFs get trained as part of the layer 1 resource processes, whereas the layer 2 processes incorporate the latest updates of the xNFs. This implies that the layer 1 resource process is effectively decoupled from the layer 2 service process, thereby simplifying the overall tuning process for AI/ML model training, integration, and delivery, for a sustainable system-wide automation.

- Segmentation within the resource and service processes: For a further simplification of the layer 1 resource process and the layer 2 service process, it is advantageous to modularise each of these processes. For example, in the case of a layer 1 resource processes, the network types may be segmented in terms of macro-cells, small-cells, various types of radio-network disaggregation, edge network, transport network, core network, frequency bands, domains etc.), and layer 2 service processes.

Leveraging these approaches for continuously re-training the embedded AI/ML models enable an effective scaling for tuning the

AI/ML models to sustain system-wide automation, while compliant with system-wide performance objectives. The embedded AI/ML models that are in an acceptable state of behaviour and performance within the system are excluded from the re-training process, while others that are in a suboptimal state of behaviour and performance are re-trained as needed. Consequently, the efficiency of the processes for AI/ML model training, integration, and delivery is optimised. Some of the pre-requisite considerations for this approach consist of the following:

- KPIs: Measurements are required for an AI/ML model to ascertain whether the related automation performance has degraded to some inferred threshold that warrants re-training. This determination hinges on a breakeven analysis, where the performance of some constituents of the overall system may not have an immediate impact on the overall quality of automation. Other constituents in the overall system may have a more immediate or weighted impact on the overall quality of automation, which would be reflected by stricter KPIs for these constituents. The KPIs associated with the constituent of the system, are likely to shift and be adjusted, based on the introduction of diverse services to be supported by the system. These KPIs will both affect and influence the drift in the quality of automation, with respect to an intended quality of automation in the system.
- Drift detection: A drift in the quality of automation in the system, is described in terms of two levels, namely, a) AI/ML model drift, and b) Data drift. The former depends on the algorithm used, while the latter depends on the characteristics of the data. Both these levels of drift trigger the enforcement of AI/ML model training.
- Service assurance alignment: An alignment of the KPIs, associated with the constituents of a system, in terms of how to modify them, weight them etc., with service assurance is required for an effective and efficient assurance of services, across related domains. This ensures that the KPIs are set properly in the system.
- Overall model governance: With the layered processes for re-training the AI/ML models, within the constituents of an autonomous system, the re-training process of AI/ML models have an impact on the quality of automation. Hence, when an AI/ML model is trained, integrated, and delivered to the constituents within the system, the quality of automation requires to be monitored and measured. This is done by drift detection associated with the overall quality of automation, through a comparison with the intent associated with a supported service.

6.6 KEY PERFORMANCE INDICATOR (KPI) FOR NETWORK AUTOMATION

Autonomous system-oriented network automation leverages AI/ML algorithms, which hinge on data collection, model training, model inference, validation, and action. The performance of an autonomous system in this regard requires to be estimated for both end-to-end system advancement and optimisation, with respect to network automation.

A KPI performance indicator (Time T), which quantifies network automation, indicates the speed of the network to automatically adapt to its configuration and parameter setting dynamically to suit the objectives of the network and supported services, through continuous change. For example, changes in the system may occur as a result of adding/removing network components, hardware/software upgrade, traffic flow variability, configuration/routing updates, new service integration, network/devices failures etc.

This KPI should reflect the effectiveness of the optimisation decisions taken by the autonomous network management and operations, controllers, and orchestrators. The optimisation of network automation, realised through adaptive decision-making in the autonomous system, includes a preservation of the expected quality of network behaviors, and the quality of user experience.

As a significant ingredient of autonomous systems, AI/ML model training time may or may not the part of this KPI. From an energy efficiency perspective AI/ML training would be a natural component of the KPI for network automation. Among the various facets of autonomous network automation, this KPI would include the performance quality of an autonomous system, in terms of data collection, inference, decision making, and implementation to achieve its objective of zero-touch automation.

For example, one of the components of this KPI, measured in terms of latencies, associated with different types of processing within an autonomous system, could be categorised as follows:

- Data collection duration (tc): The duration of the data collection process, where the data collected includes configuration and performance management data, such as traffic volume, mobility patterns, energy consumption, resource allocation etc., requires to be completed with low-latency and with sufficiently high reliability, for promoting quick decision-making that yields a highly responsive and adaptive autonomous system.
- Inference time (ti): The time duration for inferencing, which leverages the latest collection of data, for predictive decision-making, within an autonomous system, with minimum latency, for a variety of appropriate actions, such as, network/ parameter change etc., should be made considering the following, among others: Operational and maintenance rules, policy control, network resource limits, and network KPI threshold values.
- Deployment time (td): The duration of the process of generating a reconfiguration instruction and a list of targeted network function services, deployment of a new configuration, and updating of the relevant AI/ML models and agents, with minimum latency
- Validation time (tv): The duration of the process of completing actions, within an autonomous system enabled network, as well as to complete the validation of network performance to meet pre-determined targets after any change to re-establish an equilibrium or a stable state.

This KPI (T= tc + ti + td + tv) represents the effective responsiveness quality of an autonomous system, imbued with appropriate AI/ML models, to enable zero-touch network automation. For instance, RAN energy efficiency features such as Sub-Frame Silence and Channel Silence will require much lower latencies for decision-making and response, relative to energy savings features, such as Active Antenna Unit (AAU) shallow and deep dormancy [29].

Another component of this KPI, is measured in terms of the energy utilisation differential before and after the application of the appropriate AI/ML models, within an autonomous system for an optimisation of zero-touch network automation. The energy demand and consumption are associated with the corresponding utilisation of networking, computing, and storage functions within an autonomous system (core, edge, transport, radio, and user equipment), which includes the management overhead for any additional and frequent data collection, analysis, model training, inference, and execution.

07 USE CASES

A plethora of emerging use cases are anticipated, through an effective management of complexity associated with evolutionary capabilities, through autonomous system capabilities, within next-generation systems (e.g., 5G advanced and beyond), A few examples that leverage the self-CHOP attributes of an autonomous system are examined.

7.1 CUSTOMER SERVICE

GAI is a particularly useful tool in the customer service domain because of its capability to mimic human-like interactions between help-seekers and computers. It can analyse real-time call discussions and customer data to provide prompts and resources to help agents or chatbot to resolve customer inquiries. It can resolve a wide range of customer inquiries, which used to be confined to a few boilerplate questions. GAI can also be harnessed to infer customer sentiments to reduce or avoid churn, provide personalised product recommendations, service adjustments, and promotional offers.

7.2 LIVE VIDEO BROADCASTING AND JOURNALISM

Network slicing serves as a foundational enabling capability within an autonomous system, which overlays a virtualised service-based architectural framework, where monetisation is accomplished through the support for emerging usage scenarios, such as the live broadcasting of events, over a 5G Stand-Alone (5G SA) network configuration [30].

Live video broadcasting and journalism that conveys event related content in real-time is realisable for both amateurs and professional journalist subscribers, over an autonomous system with self-adaptive behaviours to optimise the service experience

On demand and dynamic network slicing, in concert with APIs, in a 5G and beyond standalone autonomous systems, enable a reliable transmission and reception of high-definition video streams, without requiring cumbersome equipment (e.g., satellite vans). The separation of Quality of service On Demand (QoD) traffic from other types of data traffic is accomplished via a dedicated QoD network slice for high-definition video streams with the requisite Quality of Service (QoS).

A QoD network slice is created dynamically, while an eMBB slice is used for regular 5G subscribers. Within a QoD network slice the attributes of an associated QoS are characterised in terms of autonomous system specific QoS parameters (e.g., 5QI, in the case of a 5G system), to differentiate between Standard-Definition (SD) and the High-Definition (HD) videos). These directions enable the convenience and efficiency of live event reporting.

7.3 AUTOMATION USING EDGE AI/ML

Service providers aspire to render an enhanced customer experience, while being price-competitive and labour efficient. This drives a tremendous new opportunity for edge AI, leveraging insights from the enormous amount of data generated at the network edge.

Automation utilising AI/ML modalities at the network edge of an autonomous system, promotes a variety of use cases, which include frictionless checkout, fraud detection, theft prevention, and inventory management. High bandwidth and low latency network connections, coupled with AI/ML analytics, at a network edge are required to meet stringent use cases, performance requirements, as well as to deliver cost-effective solutions. Hybrid AI/ML modalities with a balanced combination of on-premises analytics and aggregation, together with appropriate AI/ML model training in the cloud are significant directions for rendering an enhanced service experience.

7.4 CELL OPTIMISATION WITH AI/ML

Virtualisation of functions and resources in the RAN promotes an advanced granularity for resource allocation, which leads to opportunities for an optimisation of resource utilisation. This flexibility is beneficial for an autonomous system service assurance across the RAN segment, within an end-to-end system.

In an emerging decentralised and distributed architecture of the RAN, each cell has its own underlying physical infrastructure, which is shared by a variety of NFs and workloads, where each workload renders one or more services provided by the RAN. With the dynamically changing traffic conditions, together with the services requested by the User Equipment (UE) entities (e.g., smart phones, fixed wireless modems, sensors, actuators, self-driving vehicles etc.) connected to a given cell or aggregation site, autonomous capabilities for an automated self-adaptation of the associated workloads are a pivotal requirement.

Autonomic functions imbued with AI/ML modalities, provide a disaggregated RAN with a cognitively optimised self-CHOP capabilities, to effectively support emerging and innovative services, with demanding QoS constraints (e.g., URLLC, immersive XR etc.) located at the edges of the network, to provide an adequate and sustainable service experience, in terms of the following broad objectives:

- Performance optimisation of workloads.
- Enhanced efficiency of autonomous usage and management of underlying resources.

The challenges are around a sharing of resources on the same platform, across all workloads, which are likely to have diverse and potentially conflicting requirements. Hence a multi-faceted approach as described in section 7.5.2 would be needed to harmonise the various AI/ML models.

7.5 NETWORK ENERGY SAVING

The variations in wireless traffic in terms of temporal and spatial changes, resulting from user mobility patterns, together with diverse patterns of usage, pertaining to unique styles of life and work, provide opportunities to selectively utilise sleep modes, across the network resources, to optimise the reduction of power consumption. Autonomic principles within an autonomous system provide the wherewithal to intelligently detect opportunities, and to select sleep modes for optimising system-wide energy consumption. For example, in the RAN segment, the resources that could be intelligently targeted to be turned off opportunistically, to optimise energy consumption at different levels of granularity such as symbol, carrier, transceiver, cell, site etc.

GAI appropriately tuned to deployment specific scenarios, through prompt engineering, provide enabling capabilities to optimise the trade-off between relevant KPIs, such as QoE and energy savings. This trade-off is accomplished through the use of GAI to learn the normal behaviour of a given network deployment scenario, to generate predictions of expected network traffic patterns in real-time, thereby triggering and ending sleep modes, across network resources more accurately and rapidly.

GAI facilitates a detection of more instances of lower traffic or non-traffic windows, relative to a conventional algorithmic or rules-based prediction, since it utilises both LLMs and AI/ML to continuously infer from learned historical traffic patterns, specific to a given deployment scenario, as well as traffic patterns immediately preceding a predicted output or action, relevant to the latest state of the system and network.

7.6 RENEWABLE ENERGY INTEGRATION

An autonomous system can leverage various AI/ML modalities to manage and integrate renewable energy sources for use in a next-generation system. Smart grid integration enables the utilisation of renewable energy

sources, which allows for an optimisation of energy usage from a mix of energy sources (e.g., solar, geothermal, fossil fuel, nuclear etc.).

A self-adaptive utilisation of different energy sources based on dynamic energy demands, enables an autonomous system to adjust the energy sources for the system to align periods of highly intensive energy demands to automatically select renewable energy sources. The availability of these choices enables an autonomous system to operate effectively and efficiently, while primarily harnessing renewable energy sources for reduced costs, and promoting environmental sustainability.

7.7 SUSTAINABLE HARDWARE

Within a next-generation network infrastructure, an autonomous system provides an advancement of hardware sustainability through the use of AI/ML modalities for lifecycle management, which allows for an optimised utilisation of the network hardware. This is ensured through timely software upgrades, and recycling towards a circular economy, which reduces adverse environmental impacts.

Additionally, the different AI/ML modalities within an autonomous system can be used to evaluate and select energy-efficient hardware for network upgrades and expansion, further contributing to sustainability and efficiency in network operations, while conforming to forward-looking business and deployment objectives.

7.8 NETWORK PLANNING

Network planning and radio coverage are complex and necessitates a leveraging of specific planning algorithms. Traditional radio planning tools are used and developed based on the fundamental laws of physics and related equations (e.g. electromagnetic wave properties). The effectiveness of network planning is based on an understanding of the underlying characteristics of the related network configuration, in terms of wave propagation, wave lengths, wave reflection etc. The usage of GAI and a scenario-based large language model (e.g. a dedicated LLM model for radio network planning) is additionally based on end-user experience targets (e.g., latency, jitter, bandwidth etc.).

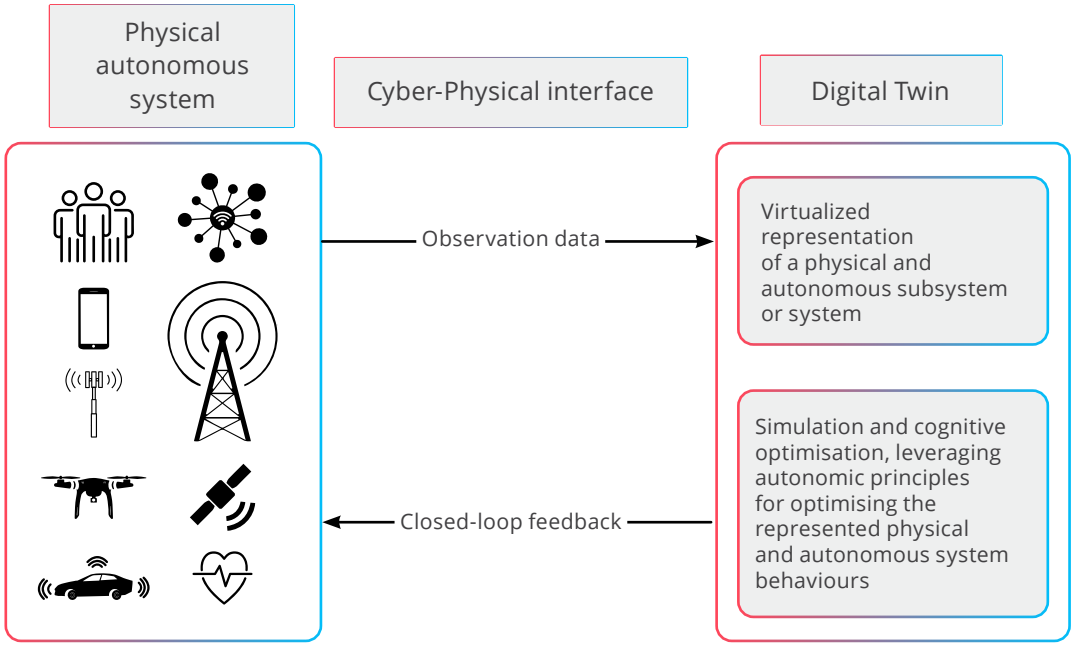

Fig. 10: Digital twin representation of a physical autonomous system

Leveraging crowdsourcing information can accelerate and complement the existing planning toolsets and processes. Additionally, LLMs can be used to continuously learn from previous and ongoing network deployments.

By connecting multiple complex AI/ML models across network planning and operations with LLMs, GAI can contribute significantly towards effective and efficient self-adaptive network planning. Constrained by parameters from various sources such as topology, network configuration, regulatory restrictions, performance requirements, spectrum allocation, budget allocation, and business objectives, GAI can generate optimal network planning, for maximising coverage, performance, and capacity.

7.9 DIRECTIONS TOWARDS THE DIGITAL TWIN

The replication of a physical entity or a subsystem, within an end-to-end next-generation system, using a virtualised representation characterises the notion of a Digital Twin (DT) [31]. The benefits of leveraging a DT that represents a physical entity in the digital domain, enables the analysis and optimisation of the physical entity to advance the efficacy and integrity of the self-CHOP behaviours of a physical entity, imbued with autonomous characteristics for self-adaptive or zero-touch automation.

The behaviors of an autonomous system are simulated in a virtual representation with anomaly detection and prevision, together with a maintenance of expected KPI, through the use of AI/ML modalities within a corresponding DT. The interactions between the autonomous system and its DT occur over the relevant cyber-physical interfaces, where the virtual representation consists of cloud-native functions. This would allow the DT to be located with appropriate levels of topological proximity, to minimise latencies that could impair the required synchronisation between autonomous system and the corresponding DT. A logical representation is depicted below Fig. 10.

The use of a DT, through a virtual representation of data collected from multiple sources, as the system operates within a given environment, allows an NSP to detect potential anomalies, while also providing guidance for tuning the system towards intended performance and behavioural objectives. Consequently, the capabilities provided by a DT affords an effective management of the rising complexity and scale associated with en evolving autonomous system.

A system-wide DT leverages pertinent information collected from various sources associated with the system and its operating environment to establish a coordinated and holistic view of the entire network. The DT consisting of autonomous system constructs, imbued with discriminative and generative AI/ML modalities, enables a real-time monitoring of network operations, and reveals predictive insights, in terms of impending maintenance requirements, or fault conditions. This enables NSPs to evaluate interesting what-if-scenarios for optimising the system-wide performance, resource utilisation efficiency, and behaviours, to suit customisable deployment scenarios, together with a personalisable service experience.

08 SECURITY AND PRIVACY

Security and privacy are integral aspects of an autonomous system. The different categories of AI/ML modalities (e.g., discriminative, and generative), together with the associated data sets require to be protected by zero-trust features [32]. Considerations for security and privacy are foundational, across these prominent categories of AI/ML modalities to ensure the integrity and alignment with anticipated, sustainable, and predictable behaviours of the autonomous system, through a continuing system-wide evolution.

Security benefits accrue through an effective management of complexity in an autonomous system, as well as through higher levels of decentralisation and distribution that reduce the attack surface and minimise the scope of any adverse impact on system availability and reliability. A faster convergence of the autonomous system, through vulnerable network changes further immunises the system against threat scenarios.

The quality of system resilience to threat scenarios also includes considerations supported by Distributed Ledger Technology (DLT) [33], for a robust up or down scaling of system-wide resources and functions, while fulfilling pertinent SLA requirement to adequately fulfil the demands of rendered services. Encryption methods to satisfy privacy and confidentiality objectives, are among the essential considerations. Protection of system-wide investments and the Total Cost of Ownership (TCO), while harvesting the enormous benefits of system-wide automation are among the benefits of an autonomous system, aligned with business and deployment specific requirements.

Encryption of pertinent data, periodic audits, relevant access controls, and standards compliance, are pivotal for security and privacy of data being utilised by GAI in a variety of emerging systems and services.

From this perspective the following are broad considerations that are aligned with ensuring both security and privacy:

- Ensuring the security and privacy of training data, used in GAI, for a variety of usage scenarios (e.g., healthcare, mobility, positioning, beamforming etc.)
- Use of LLMs, with Retrieval Augmented Generation (RAG) [34] for avoiding inaccuracies, and hallucinations in the outputs, to ensure the preservation of both security and privacy.
- Harnessing LLMs trained on large amounts threat scenarios, vulnerabilities, attack patterns, anomaly detection capabilities, threat insight extraction, and attack prediction [35] data, serve as an intelligent bastion of protection for security and privacy.
- Regulatory oversight to ensure compliance with security and privacy policies.
- Ethical guidelines are also a safeguard for LLMs to sustain the robustness of defences against threats that compromise security and privacy

09 INDUSTRY GAPS, COOPERATION AND STANDARDISATION

With the continuing evolution of the 5G Advanced and beyond ecosystem both discriminative (algorithmic) and generative (LLM oriented) modalities of AI/ML are essential for a corresponding advancement of autonomous system behaviours for managing complexity, scale, and sustainability. The necessary advancements in autonomous system behaviours for yielding zero-touch automation in the presence of rising complexity and scale of next-generation systems, present associated challenges, and gaps to be addressed in the industry.

For example, some of the challenges pertain to satisfying diverse and emerging requirements, associated with various connectivity arrangements (e.g., optimising coverage and capacity), as well supporting higher levels service customisation, with different quality of service related resource allocation [36].

These challenges broadly consist of the diverse demands of evolving next-generation systems that are required to meet emerging market and industry demands (e.g., distributed, and ubiquitous connectivity, higher levels of service customisation, sophistication, and experience, improved resource, and energy utilisation etc.). The gaps in the industry and standardisation broadly consist of new and necessary capabilities, relative to previous generations, in terms of realising autonomous system behaviours, through an application of various arrangements of closed-loop feedback, imbued with appropriate AI/ML modalities.

It is anticipated that the challenges and gaps can be addressed through continuing studies, research, cooperation, consensus and standardisation of selected interfaces and build-block procedures. This in turn is expected to promote interoperability and consistent autonomous system behaviours in a multi-vendor ecosystem of autonomous systems for widespread, zero-touch network automation and autonomy.

10 LIST OF ABBREVIATIONS

AI/ML	Artificial Intelligence/Machine Learning
Autonomous (Autonomic)	Self-management characterized by self-CHOP (Configuring, Healing, Optimising, and Protecting) for cognitively adapting to environmental changes to suit a given behavioral objective or intention, realized through the principle of closed-loop feedback. (adjective)
Autonomous system	Any entity, network, system, or subsystem, characterized by autonomous or autonomic capabilities that render autonomy for the entity, implying independence of human intervention. (noun)
Automatic	Attribute of an autonomous entity which is self-adaptive, or an entity that is not autonomous, while being programmatic with limited adaptability. (adjective)
Automation	Process that embodies automatic behavior. (noun)
Autonomy	Condition or state of being autonomous, where the associated entity operates independently. (noun)
BSS	Business Support System
CI/CD/CT	Continuous Integration/Continuous Delivery/Continuous Training
cNF	Cloud Native Function
DAI	Discriminative Artificial Intelligence, using conventional AI/ML methods (e.g., supervised, unsupervised, reinforcement learning, etc.)
DE	Decision Element
DLT	Distributed Ledger Technology
eMBB	enhanced Mobile Broadband
FL	Federated Learning
GAI	Generative Artificial Intelligence, using generative AI/ML methods, consisting of LLMs that yield semantic inferencing capabilities for content generation
Intent	This refers to an abstract, prescriptive, and adaptive high-level expression of policy for system-wide (end-to-end network) operation, based on autonomous systems.
IoT	Internet of Things
KP	Knowledge Plane
KPI	Key Performance Indicator

mIoT	massive IoT
NF	Network Function
NSP	Network Service Provider. This refers to an operator of a next-generation (5G and beyond) network infrastructure that also owns the assets.
NWDAF	Network Data Analytics Function
OAM	Operation Administration and Maintenance
OSS	Operations Support System
QoD	Quality on Demand. This refers to on-demand management of bandwidth and latency for a connection.
QoE	Quality of Experience
QoS	Quality of Service
REST	Representational State Transfer, which embodies an architectural style for APIs, with the principles of platform independence, statelessness between a client and server, etc.
RL	Reinforcement Learning
Self-CHOP	Self-(Configuring, Healing, Optimising, and Protecting)
SDO	Standards Development Organisation
SP	Service Provider. This refers to a service providing entity in the context of a next-generation (5G and beyond) context (e.g., Verticals or other entities, including an NSP).
TCO	Total Cost of Ownership
TL	Transfer Learning
URLLC	Ultra-Reliable Low-Latency Communications
vNF	Virtual Network Function

11 FIGURES

12 REFERENCES

[1] NGMN, "Automation and Autonomous system Architecture Framework," v1.01, November 2022

[2] 3GPP, "System architecture for the 5G System (5GS)," TS 23.501 V18.1.0, March 2023

[3] 6G SNS, "What societal values will 6G address?" Version 1.0, May 2022

[4] ISO, "Information technology - Artificial intelligence - Artificial intelligence concepts and terminology," ISO/IEC 22989:2022

[5] McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A., "Communication-efficient learning of deep networks from decentralised data," Google Inc., 20th International Conference on Artificial Intelligence and Statistics 2017

[6] ETSI, "Zero-touch network and Service Management (ZSM); Enablers for Artificial Intelligence-based Network and Service Automation," GS ZSM 012, v1.1.1, December 2022.

[7] NGMN, "Update to Description of network slicing concept," v1.0.8, September 2016

[8] Ksentini, A., Nikaein, N., "Toward enforcing network slicing on RAN: Flexibility and resources abstraction," IEEE Communications Magazine, June 2017.

[9] ETSI, "Evolving NFV towards the next decade, " http://portal.etsi.org/NFV/NFVWhitePaper.pdf, May 2023

[10] Xia, W., Wen,T., Foh, C.H., Niyato, D., Xie, H., "A Survey on Software-Defined Networking, " IEEE Communications, 2015.

[11] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., and Polosukhin. I., "Attention is all you need." 31st Conference on Advances in Neural Information Processing Systems, 2017.

[12] Anthony, G., Casas, J., Mugellini, E., Khaled, O.A., "Overview of the Transformer-based Models for NLP Tasks," Proceedings of the Federated Conference on Computer Science and Information Systems, 2020.

[13] Bank, D., Koenigstein, N., Giryes, R., "Autoencoders." arXiv, April 2021.

[14] Xu, M., Du, H., Niyato, D., Kang, J., Xiong, Z., Mao, S., Han, Z., Jamalipour, A., Kim, D.I., Shen, X., Leung, V.C.M., Poor, V. " Unleashing the Power of Edge-Cloud Generative AI in Mobile Networks: A Survey of AIGC Services,", October 2023.

[15] Bariah, L., Zou, H., Zhao, Q., Mouhouche, B., Bader, F., Debbah, M., "Understanding telecom language models through LLMs," IEEE Global Communications Conference, 2023.

[16] Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., Le, Q.V., "XLNet: Generalized autoregressive pretraining for language understanding, "33rd Conference on Neural Information Processing Systems, 2019.

[17] Elsayed, M., Erol-Kantarci, M., "AI-enabled future wireless networks - Challenged and opportunities," IEEE Vehicular Technology Magazine, September 2019.

[18] IETF, "Segment Routing over IPv6 (SRv6) Network Programming", RFC 8986, February 2021

[19] Makinen, S., Skogstrom,H., Laassonen, E., Mikkonen, T.,"Who Needs MLOps: What Data Scientists Seek to Accomplish and How Can MLOps Help?" 2021 IEEE/ACM 1st Workshop on AI Engineering - Software Engineering for AI (WAIN)

[20] Kreuzberger, D., Kuhl, N., Hirschl, S., " Machine Learning Operations (MLOps): Overview, Definition, and Architecture," IEEE Access, February 2023.

[21] Jia, J., Liang, W., Liang, Y., "A Review of Hybrid and Ensemble in Deep Learning for Natural Language Processing," https://arxiv.org/abs/2312.05589, December 2023

[22] Mueller, M., Muller, T., Talkhestani, B.A., Marks, P., Jazdi, N., Weyrich, M., "Industrial autonomous systems: A survey on definitions, characteristics, and abilities," At Automatisierungstechnik, January 2021

[23] Midjourney, "Midjourney," https://www.midjourney.com/home, 2023.

[24] Google, "Lamda: Our breakthrough conversation technology," https://blog.google/technology/ai/lamda, 2021

[25] OpenAI, "Chatgpt," https://openai.com/blog/chatgpt," 2023.

[26] C´amara, J., Troya, J., Burgueo, L., Vallecillo, A.:" On the assessment of generative AI in modeling tasks: an experience report with ChatGPT and UML," Software and Systems Modeling, June 2023.

[27] Kephart, J.; Chess, D. "The vision of autonomic computing," Computer 2003.

[28] Sokratis, B., Demestichas, P., "Framework for Trustworthy AI/ML in B5G/6G," 1st International Conference on 6G Networking, IEEE, 2022.

[29] NGMN, "Definition of the Testing Framework for the NGMN 5G Trial and Testing Initiative Phase 2," v1.8, December 2022.

[30] Telekom, "Telekom's 5G network slicing brings mobile TV teams live to the network," September 2023.

[31] Rasheed, A., San, O., Kvamsdal, T., "Digital Twin: Values, Challenges and Enablers From a Modeling Perspective," IEEE Access, February 2020.

[32] Benzaid, C., Taleb, T., "AI-Driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions", IEEE Network, March/April 2020.

[33] ITU-T, "Distributed ledger technologies: Use cases", Technical paper, HSTP.DLT-UC, October 2019.

[34] Huang, X., Tang, Y., Li, J., Zhang, N., Shen, X., "Toward Effective Retrieval Augmented Generative Services in 6G Networks," IEEE Network, 2024.

[35] Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L., "From ChatGPT to ThreatGPT: Impact of

Generative AI in Cybersecurity and Privacy", IEEE Access, Volume 11, .2023.

[36] ETSI, "Experiential Networked Intelligence (ENI); System Architecture," GS ENI 005, v2.1.1, December 2021.

www.etsi.org/deliver/etsi_gs/NFV-IFA/001_099/027/04.04.01_60/gs_NFV-IFA027v040401p.pdf. [Accessed 17 1 2024].

[14] 3GPP, „TS 28.554 V17.12.0 (2024-01), 5G end to end Key Performance Indicators (KPI)," [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/28_series/28.554/28554-hc0.zip. [Accessed 17 1 2024].

[15] 3GPP, „TS 28.552 V17.12.0 (2024-01), Management and orchestration; 5G performance measurements," [Online]. Available: https://www.3gpp.org/ftp/Specs/archive/28_series/28.552/28552-hc0.zip. [Accessed 17 1 2024].

13 ACKNOWLEDGEMENTS

China Mobile, Lingli Deng

China Mobile, Yuhan Zhang

Cisco, Amit Dass

Cisco, Tony Verspecht

Deutsche Telekom, Sebastian Zechlin

HPE, Andreas Volk

Huawei, Jean Paul Pallois

Huawei, Luigi Licciardi

Intel, Gary Li

TELUS, Jermin Girgis

UScellular, Sebastian Thalanany

ZTE, Manchang Ju

NEXT GENERATION MOBILE NETWORKS ALLIANCE

NGMN is a forum established in 2006 by world-leading Mobile Network Operators. NGMN is a global operator-led alliance, comprising over 80 companies and organisations across operators, manufacturers, consultancies and academia.

Its objective is to guarantee that next generation network infrastructure, service platforms, and devices will fulfil the requirements of operators and, ultimately, meet end-user demands and expectations.

VISION

The vision of NGMN is to provide impactful industry guidance to achieve innovative, sustainable and affordable mobile telecommunication services for the end user with a particular focus on Mastering the Route to Disaggregation, Green Future Networks and 6G, whilst continuing to support 5G's full implementation.

MISSION

The mission of NGMN is:

- To evaluate and drive technology evolution towards the three **Strategic Focus Topics:**
  - **Mastering to the Route to Disaggregation:** Leading in the development of open, disaggregated, virtualised and cloud native solutions with a focus on the E2E Operating Model
  - **Green Future Networks:** Developing sustainable and environmentally conscious solutions
  - **6G:** Anticipating the emergence of 6G by highlighting key technological trends and societal requirements, as well as outlining use cases, requirements, and design considerations to address them.
- To define precise functional and non-functional requirements for the next generation of mobile networks
- To provide guidance to equipment developers, standardisation bodies, and collaborative partners, leading to the implementation of a cost-effective network evolution
- To serve as a platform for information exchange within the industry, addressing urgent concerns, sharing experiences, and learning from technological challenges
- To identify and eliminate obstacles hindering the successful implementation of appealing mobile services.