The Challenge of Zero Touch and Explainable AI

Biswadeb Dutta, Andreas Krichel and Marie-Paule Odini*

HPE, Palo Alto, United States
E-mail: biswadeb.dutta@hpe.com, andreas.krichel@hpe.com,
marie-paule.odini@hpe.com
*Corresponding Author

Abstract

With ever increasing complexity and dynamicity in digital service provider networks, especially with the emergence of 5G, operators seek more automation to reduce the cost of operations, time to service and revenue of new and innovative services, and increase the efficiency of resource utilization, Complex algorithms leveraging ML (machine learning) are introduced, often with the need for frequent training as the networks evolve. Inference is then applied either in the core directly, or in the management stack to trigger actions and configuration changes automatically. This is the essence of Zero Touch. The challenge that analysts are often faced with is to trace back from the inference or prediction to the original events or symptoms that led to the triggered action, which ML model version or pipeline was used. This paper describes the challenges faced by analysts and provides some solutions.

Keywords: Zero touch, closed loop, 5G, analytics, machine learning, explainable AI.

1 Introduction

With the advent of 5G, not only does network usage and traffic increase significantly, but so does the complexity of the network and IT environments. More and more events are generated and managing 5G networks with traditional OSS systems is not possible anymore. Automation is required and AI/ML is being introduced to apply some machine learning algorithms to raw data, filter the volume of events, identify problems, and present most relevant information to upper management layers. This is called AIOps, or 'AI based Operations', as defined by Gartner [1]. However, along with AI comes more and more concern about "black box AI", on whether AI was performed, if so which algorithm was used, etc. Explainable AI, XAI, is now raised as a major ethical requirement, both in the US, by DARPA [2], and in Europe [3].

With AIOps, several AI/ML algorithms are starting to be used, such as linear regression, decision trees, naive Bayes, nearest neighbors, etc. More specific algorithms or AI/ML models may be developed to meet specific requirements. ITU-T Y.3172 [11] provides a high-level architectural framework that is generally applicable to describe the principal architectural components used to integrate machine learning in 5G and future networks. The document also describes the high-level architecture of an ML sandbox and ML pipeline. Using this architectural framework as a reference, one finds that events reaching the SINK node, after ML processing, need information regarding the input data used at the SOURCE, and the details of the ML pipeline and algorithms, to arrive at the results in the event at the SINK. This would enable the fact analysis and understanding of results at the SINK.

Similarly AI is being introduced in the Core network typically with Network Data Analytics Function (NWDAF, introduced by 3GPP [6, 7]) which collects data from different Network Functions (NF) on the service based architecture and then performs some descriptive analytics or predictions to deliver smart metrics or notifications to different consumer NF or to the management stack.

2 AI and Closed Loops in 5G Networks

As mentioned earlier, AI is introduced in the core network with the NWDAF (network data analytics function, see [7]) and a closed loop with the 5G Core Network Functions (see [15]), but also in the management stack with 3GPP MDAF (management data analytics function, see [6]). The two closed loops are depicted in Figure 1.
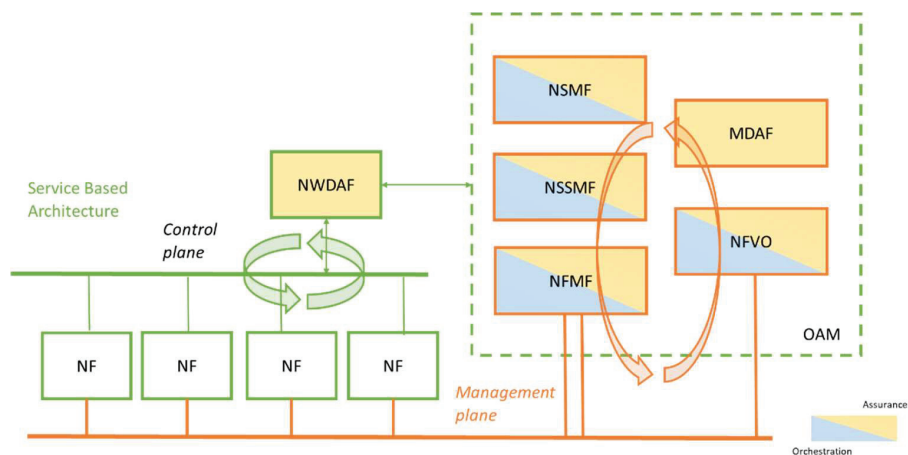
Figure 1   5G Core and OAM closed loop.

The arrows indicate the interaction between Assurance/Analytics functions and the Orchestration/Configuration functions. They are typically called "closed loop" representing the automated detection and resolution of problems without a human operator involved (also known a "zero-touch").

With 5G closed loops can exist within the network (before any problem escalates into the management plain), and at the management level (the OAM).

Each of these closed loops is based on modelling, policies, workflows and APIs, as further illustrated in Figure 2. ZSM [4] is one example of standards defined for closed loop.

2.1 Closed Loop Principles

Closed loop as defined in the industry by ETSI ZSM [4] or 3GPP [6] as a set of building blocks starting with monitoring (also called observability) that typically collects data from the managed resource, then data is sent to be analysed to an analytics function, the output is a decision that triggers some action, that is sent to an execution entity, also called the consumer of the analytics function. The execution entity will perform some operations on the managed entity and the process will continue in circle, i.e. closed loop, based on this new updated environment. We can say, a service consumer's intent is fulfilled in full autonomy of the manged service, enabled by a closed loop, i.e. "zero-touch" for the operator.
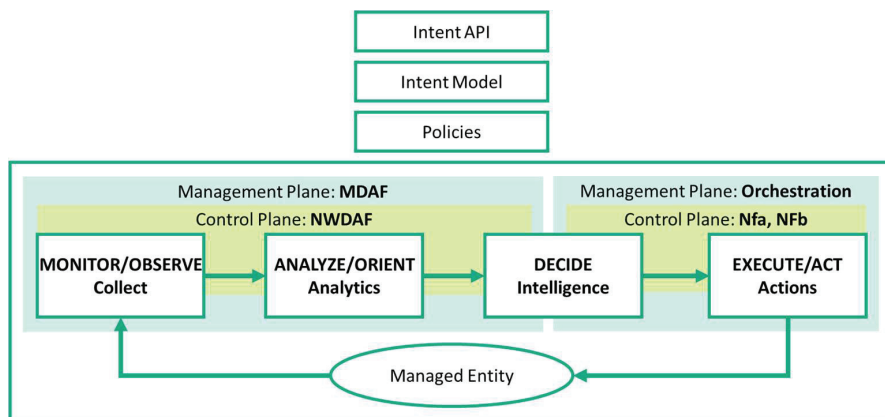
Figure 2   Closed loop principle, examples in control plane and management plane.

3GPP suggests analytics functions on different levels: the NWDAF at control plane level [7], the MDAF at management level [6].

Figure 2 illustrates the concept and maps the specific functions to the building blocks of the closed loop.

An example of a "fast" closed loop in the network, is the NWDAF in the 5G core [7]. The NWDAF collects data from different NFs, analyses and provides either smart metrics, like UE mobility information that is passed on to another NF or group of NFs that will decide and execute some operation. So it avoids larger data processing at management level, although it is limited to a specific scope of the managed functions. In some other use cases though, NWDAF itself might play a role in the decision process. For instance, if a threshold is defined by a policy, the NWDAF could 'decide' that a slice is congested, and let another entity, namely the OAM, execute some slice life cycle management operation to fix the issue.

The most typical other example and covering the wider scope of managed entities, including services, is the closed loop set up in the management stack between Assurance, Orchestration and the Managed Entity (service or resource). Here the Assurance MDAF collects data, such as faults or metrics from the 5G Core, and makes some decision based on an anomaly detection, or prediction, and sends the information to orchestration for some policy update or configuration change on the 5G Core managed entity. An Intent Based Service Orchestration (cf. [5, 9]) may finally decide *how* a higher service level decision (expressed as intent) is translated to lower service or resource control level.

2.2 Analytics Engine principles

Each analytics component of the above closed loop can be enriched with AI analytics and different AI models that are being trained with initial data sets, then deployed as inference engine to process the incoming monitored data. But these engine can also be updated, with new algorithms, or updated training data sets that generate new trained models. Consequently, each inference engine is characterized by the specific trained model or pipeline of trained models it is using at a given time.
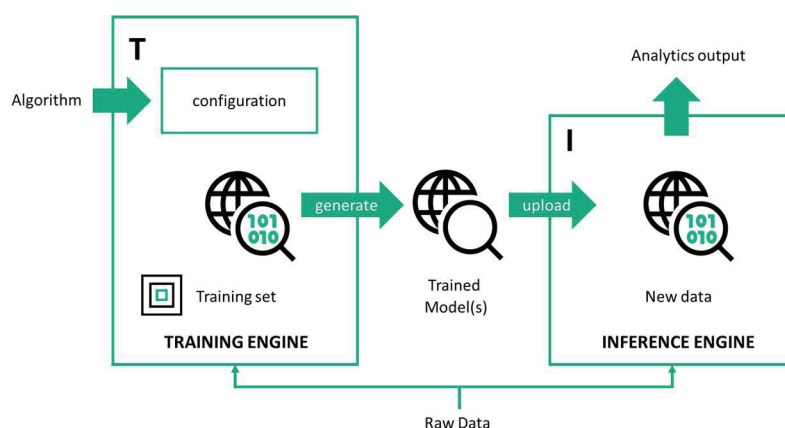

Figure 3   Analytics engine principles.

3 AI Closed Loops and Explainable AI

3.1 Sequence of Closed Loops in a 5G Managed Network

ITU-T Y.3174 [12] identifies the principal data collection, processing and output generation requirements in ML applications (or "ML pipelines"). ITU-T Y.3176 [13] describes the high- level requirements, and associated architectural components, for integration of an ML marketplace into the management of current generation networks. Within the context of these recommendations, it is possible to foresee multiple closed loops being executed at different levels in a network for example NWDAF within the control plane or MDAF with an orchestrator in the management plane, applying different models. The sequence of closed loops results in the network being updated constantly. As the managed entity is constantly evolving, this requires the data sets to be updated regularly and the models to be re-trained to better meet the
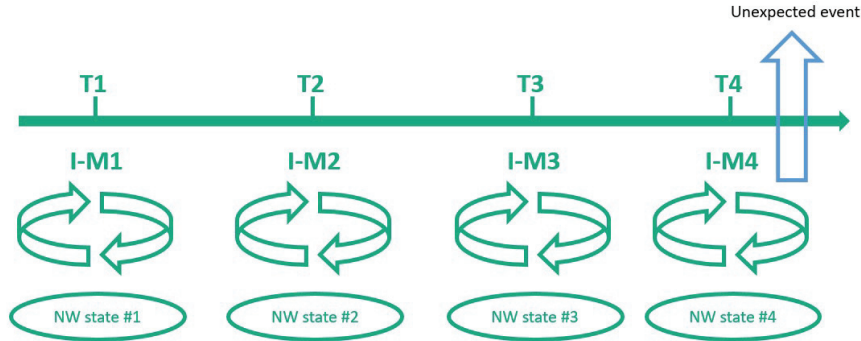
Figure 4   Sequence of closed loops with evolving models in a managed 5G network.

requirements of the updated network. Newer algorithms, better performing or addressing some new issues may be deployed. ITU-T Y.3174 [12] calls such engine Machine learning function orchestrator (MLFO). Consequently the network at a given time T4 is the result of a series of closed loop actions based on a series of evolving inference models I-M1, I-M2, I-M3, I-M4 that triggered changes in the network state from state#1 to state#4 as described in Figure 4.

If an unexpected event was to occur at T4, it may be required to trace back what happened in the network. Some external network audit may request the same. Customers, government entities or regulators may also request some tracing and history, at least for some segments of the network, i.e. a PNI-NPN private network (Public Network Integrated Non-Public Network as defined by 3GPP, see [15] for general concepts of network sharing).

3.2 AI/ML Model Metadata to Facilitate 5G Explainable AI

With increasing complexity and dynamicity of emerging networks, the quantum of management data being generated is exploding. The approach from the beginning has been to extract pieces of information from individual events, logs or measurements, correlate or assemble them together in some fashion, and then create a summary event for the consumption of the human being. Since the volume, velocity, dimensionality and complexity of the data has increased, simple techniques to correlate data do not work anymore. These challenges serve as ideal catalysts for the application of machine learning. With time, as more advanced machine learning algorithms and techniques are applied, it becomes difficult for the end user to explain the algorithmic approach used to arrive at the end result, especially in situations where a series of updated models over a period of time have been applied.

The solution to address this problem being proposed here is to include a dictionary that provides a repository for capturing the taxonomy of all AI/ML models used in the management system. The dictionary could be implemented as a service with a RESTful interface wrapped around a database. Since Directed Acyclic Graphs (DAGs) and Factor Graphs (FGs) provide an extremely good mechanism to represent the taxonomy of AI/ML models, we propose using a graph database, to capture the DAG or FG natively as a graph. The dictionary service would provide a repository for the meta-models representing the AI/ML algorithms used by the management system. Further, the system will need a model or algorithm instance repository (MIR or AIR), that will be used to persist explicit instances of the AI/ML algorithms used. An instance in the MIR/AIR will also include the input variables, output variables, and other features of the specific instance of the AI/ML model used. For example, if an instance of the deep learning Convolutional Neural Network (CNN) [14] is used for classification of time series (Figure 5 below) of observable measurements, performance and traffic KPIs, representing different problems in call handling in the 5G core network, the dictionary service will hold a graph representing the CNN. The MIR/AIR will hold a graph of the CNN indicating the sequence of input, convolutional, pooling, fully connected and output layers of the CNN. As this is generally the case, the CNN instance might have several convolutional and pooling layers. Further, vertices of the graph describing the input layer of the CNN will include attributes describing the semantics associated with the input variables, and associated weights. The attributes of the edges connecting
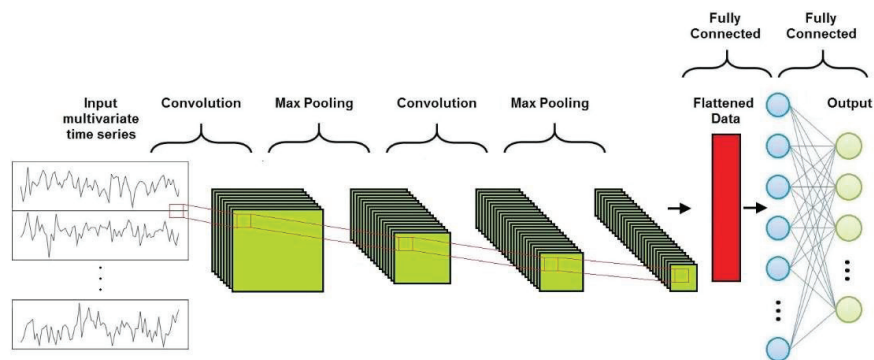
Figure 5   Example of a CNN used for time series classification in problem identification. The result of the classification generates an event that is used by an upper level management system for triage.

the convolutional layer to the pooling layer of neurons could indicate the activation function used for pooling or down sampling, the stride of the filter and the padding used. The edges and vertices representing the other layers of the CNN could be assigned attributes in a similar fashion. The vertices in the output layer of the CNN could be connected through edges to a root node representing the specific instance of the CNN. The label assigned to the root node of the CNN graph can then be used in the resulting event post application of the CNN processing of the input multivariate time series. The performance metrics used to characterize the classification results of the CNN could also be referenced in the root node representing the CNN.

The above description has been provided to serve as an example of how the meta-model of the CNN in the dictionary service, and a specific instance of the CNN in the MIR/AIR, could be used to annotate the resulting event with a tag or label. The user or analyst of the end result could then use the tag or label in the event to obtain an explanation of how the event had been generated through the application of a CNN.
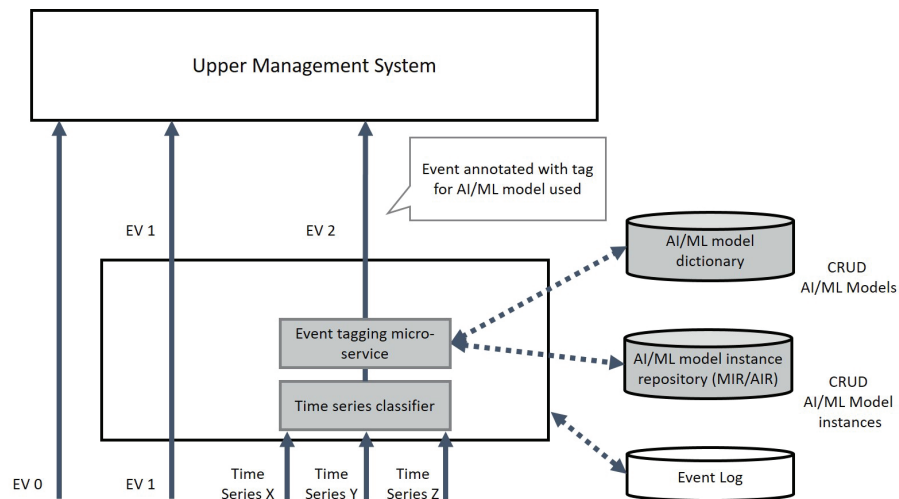
Figure 6   Schematic showing the architectural components of a system used to annotate events with information regarding the AI/ML model used to arrive at a result.

With such an annotation in the resultant event, it is now simpler for the analyst to retrieve the details from the AI/ML model dictionary and associated instance repository, namely the AI Repository (AIR) and Model Instance

Repository (MIR), reverse engineer and come up with an explanation for the result. In summary, to facilitate explainable AI, the solution being proposed is to include some metadata in the interface between the Analysis and Decision phases to convey information on the Inference engine and the version of the model or pipeline of models being used. This metadata is represented as:

I-M1, or I-(M1, M2), or I-P1 with P1 being a pipeline of models

Referring to the notation used in ITU-T Y.3172, I-Mx would correspond to metadata of the entire ML pipeline shown in Figure 6, which consists of a sequence of logical processing nodes. We are considering a further generalization of what is shown in Figure 6 of ITU-T Y.3172, where one can include the metadata of a sequence or hierarchy of ML pipelines.

For example, more sophisticated analytics models could be represented by a factor graph [10], or by a Directed Acyclic Graph (DAG) representing the pipeline of analytics models being executed. Figure 5 shows a very simple example of a Directed Acyclic Graph (DAG) where an NWDAF is predicting the NF Load or Slice Load. In this example expanding on the Figure 2, the NWDAF would receive NF load events, analyse them, and generate an NF load alert event notification including some metadata referring to the DAG described in Figure 7.
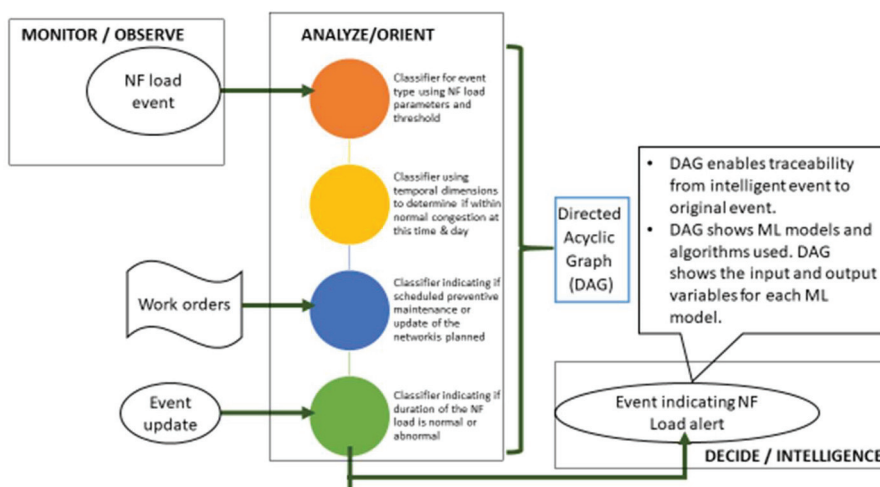

Figure 7 NWDAF NF load analytics example with directed acyclic graph.

4 Conclusion

As 5G is being deployed and more and more use cases appear, increasing the amount of traffic and the dynamicity of the network requiring more and more automation and analytics in the network, it is anticipated that closed loop analytics and mechanisms such as the one proposed here will be implemented and standardized in 3GPP or ETSI specification to meet the requirements of traceable AI and trusted zero touch networks. Our solution as described in this paper provides a mechanism to introduce explainable AI to the analytics functions specified in the 3GPP 5G closed loops triggered by NWDAF in the control plane and MDAF in the management plane.

References

[1] Gartner AIOps – https://www.gartner.com/en/information-technology/glossary/aiops-artificial-intelligence-operations

[2] XAI – Darpa – https://www.darpa.mil/program/explainable-artificial-intelligence

[3] Europe explainable AI – https://publications.jrc.ec.europa.eu/repository/bitstream/JRC119336/dpad_report.pdf

[4] ETSI ZSM002 – Zero-touch network and Service Management (ZSM); Reference Architecture.

[5] ETSI ZSM005 – Zero-touch network and Service Management (ZSM); Means of Automation.

[6] 3GPP TS 28.533 5G Management and orchestration; Architecture framework.

[7] 3GPP TS 29.520 5G System; Network Data Analytics Services.

[8] C. Benzaid and T. Taleb, "AI-Driven Zero Touch Network and Service Management in 5G and Beyond: Challenges and Research Directions," in IEEE Network, vol. 34, no. 2, pp. 186–194, March/April 2020, doi: 10.1109/MNET.001.1900252.

[9] HPE 5G Orchestration and Automation Toward Zero-touch Service Management.

[10] *Factor Graphs and the Sum-Product Algorithm*, Frank R. Kschischang, Brendan J. Frey and Hans-Andrea Loeliger, IEEE Transactions on Information Theory, Vol. 47, No. 2, February 2001.

[11] ITU-T Y.3172: Architectural framework for machine learning in future network including IMT-2020.

[12] ITU-T Y.3174: Machine learning marketplace integration in future networks including IMT-2020.

[13] ITU-T Y.3176: Machine learning marketplace integration in future networks including IMT-2020.

[14] Survey of the Recent Architectures of Deep Convolutional Neural, published in Artificial Intelligence Review, April 2020. https://arxiv.org/ftp/arxiv/papers/1901/1901.06032.pdf

[15] 3GPP TS 23.501 System architecture for the 5G System (5GS).

Biographies

Biswadeb Dutta is a lead architect and member of the product engineering organization of HPE's Communications Technology Group. Biswadeb leads the architecture and development of HPE's OSS assurance portfolio. Among several of his responsibilities, Biswadeb focuses on the application of AI and ML to address problems arising in the assurance space for telecommunications networks and digital service provider operations. Biswadeb has worked with several of the largest digital service providers across the globe and assisted them in addressing their operations challenges.

Andreas Krichel Andreas Krichel is Distinguished Technologist in the Portfolio Strategy management team for HPE's Communications Technology Group. His focus is on the 5G E2E automation across the different product families. Andreas led the architecture for HPE's flagship orchestration product Service Director. Today Andreas works active in standards such as ETSI ZSM and is lead technologist for various customer engagements around zero-touch management and 5G.

Marie-Paule Odini is Distinguished Technologist in HPE Telecom Division focused on customer innovation and emerging trends including NFV, SDN, IoT, AI, 5G and 6G. Active in industry forums and standard organization. She is Chair of GreenG and held key positions such as ETSI NFV Vice Chair, IEEE SDN Chair, Editorial board member, 5G Americas key contributor, co-chair of TIP E2E network slicing project and Next Gen Alliance Steering board member. Prior to HPE she worked in France Telecom/Orange labs.