|  |  |
| --- | --- |
| 3GPP TS 28.908 V0.3.0 (2022-07) | |
| Technical Report | |
| 3rd Generation Partnership Project;  Technical Specification Group Services and System Aspects;  Management and orchestration;  Study on Artificial Intelligence / Machine Learning (AI/ML) management (Release 18) | |
|  | |
| *5G-logo_175px* | 3GPP-logo_web |
|  | |
| The present document has been developed within the 3rd Generation Partnership Project (3GPP TM) and may be further elaborated for the purposes of 3GPP. The present document has not been subject to any approval process by the 3GPPOrganizational Partners and shall not be implemented. This Specification is provided for future development work within 3GPPonly. The Organizational Partners accept no liability for any use of this Specification. Specifications and Reports for implementation of the 3GPP TM system should be obtained via the 3GPP Organizational Partners' Publications Offices. | |

|  |
| --- |
|  |
| ***3GPP***  Postal address  3GPP support office address  650 Route des Lucioles - Sophia Antipolis  Valbonne - FRANCE  Tel.: +33 4 92 94 42 00 Fax: +33 4 93 65 47 16  Internet  http://www.3gpp.org |
| ***Copyright Notification***  No part may be reproduced except as authorized by written permission. The copyright and the foregoing restriction extend to reproduction in all media.  © 2022, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).  All rights reserved.  UMTS™ is a Trade Mark of ETSI registered for the benefit of its members  3GPP™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners LTE™ is a Trade Mark of ETSI registered for the benefit of its Members and of the 3GPP Organizational Partners  GSM® and the GSM logo are registered and owned by the GSM Association |

Contents

Foreword 5

Introduction 6

1 Scope 7

2 References 7

3 Definitions of terms, symbols and abbreviations 7

3.1 Terms 7

3.2 Symbols 7

3.3 Abbreviations 7

4 Concepts and overview 8

4.1 Overview 8

4.2 AI/ML workflow for 5GS 8

5 Use cases, potential requirements and possible solutions 9

5.1 AI/ML model performance management 9

5.1.1 Description 9

5.1.2 Use cases 9

5.1.2.1 AI/ML model performance indicators 9

5.1.3 Potential requirements 9

5.1.4 Possible solutions 10

5.2 Event data for ML training 10

5.2.1 Description 10

5.2.2 Use cases 10

5.2.2.1 Pre-processed event data for ML training 10

5.2.3 Potential requirements 11

5.2.4 Possible solutions 11

5.3 AI/ML entity validation 11

5.3.1 Description 11

5.3.2 Use cases 11

5.3.2.1 AI/ML entity validation performance reporting 11

5.3.3 Potential requirements 11

5.3.4 Possible solutions 12

5.4 AI/ML entity testing 12

5.4.1 Description 12

5.4.2 Use cases 12

5.4.2.1 Consumer-requested AI/ML entity testing 12

5.4.2.2 Control of AI/ML entity testing 13

5.4.3 Potential requirements 13

5.4.4 Possible solutions 13

5.5 AI/ML deployment 14

5.5.1 Description 14

5.5.2 Use cases 14

5.5.2.1 AI/ML deployment information retrieved by consumer 14

5.5.3 Potential requirements 14

5.5.4 Possible solutions 14

5.5.5 Evaluation 14

5.6 AI/ML Inference History 15

5.6.1 Description 15

5.6.2 Use cases 15

5.6.2.1 Tracking AI/ML inference decision and context 15

5.6.3 Potential requirements 15

5.6.4 Possible solutions 15

5.7 AI/ML context 16

5.7.1 Description 16

5.7.2 Use cases 16

5.7.2.1 AI/ML context monitoring and reporting 16

5.7.3 Potential requirements 16

5.7.4 Possible solutions 16

5.8 AIMLEntity Capability Discovery and Mapping 16

5.8.1 Description 16

5.8.2 Use cases 17

5.8.2.1 Identifying capabilities of AI/MLEntities 17

5.8.2.2 Mapping of the capabilities of AI/MLEntities 17

5.8.3 Potential requirements 17

5.8.4 Possible solutions 18

5.9 AI/ML update management 18

5.9.1 Description 18

5.9.2 Use cases 18

5.9.2.2 AI/ML entities updating initiated by producer 18

5.9.4 Possible solutions 19

Annex X (informative): Change history 20

# Foreword

This Technical Specification has been produced by the 3rd Generation Partnership Project (3GPP).

The contents of the present document are subject to continuing work within the TSG and may change following formal TSG approval. Should the TSG modify the contents of the present document, it will be re-released by the TSG with an identifying change of release date and an increase in version number as follows:

Version x.y.z

where:

x the first digit:

1 presented to TSG for information;

2 presented to TSG for approval;

3 or greater indicates TSG approved document under change control.

y the second digit is incremented for all changes of substance, i.e. technical enhancements, corrections, updates, etc.

z the third digit is incremented when editorial only changes have been incorporated in the document.

In the present document, modal verbs have the following meanings:

**shall** indicates a mandatory requirement to do something

**shall not** indicates an interdiction (prohibition) to do something

The constructions "shall" and "shall not" are confined to the context of normative provisions, and do not appear in Technical Reports.

The constructions "must" and "must not" are not used as substitutes for "shall" and "shall not". Their use is avoided insofar as possible, and they are not used in a normative context except in a direct citation from an external, referenced, non-3GPP document, or so as to maintain continuity of style when extending or modifying the provisions of such a referenced document.

**should** indicates a recommendation to do something

**should not** indicates a recommendation not to do something

**may** indicates permission to do something

**need not** indicates permission not to do something

The construction "may not" is ambiguous and is not used in normative elements. The unambiguous constructions "might not" or "shall not" are used instead, depending upon the meaning intended.

**can** indicates that something is possible

**cannot** indicates that something is impossible

The constructions "can" and "cannot" are not substitutes for "may" and "need not".

**will** indicates that something is certain or expected to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**will not** indicates that something is certain or expected not to happen as a result of action taken by an agency the behaviour of which is outside the scope of the present document

**might** indicates a likelihood that something will happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

**might not** indicates a likelihood that something will not happen as a result of action taken by some agency the behaviour of which is outside the scope of the present document

In addition:

**is** (or any other verb in the indicative mood) indicates a statement of fact

**is not** (or any other negative verb in the indicative mood) indicates a statement of fact

The constructions "is" and "is not" do not indicate requirements.

# Introduction

This clause is optional. If it exists, it shall be the second unnumbered clause.

# 1 Scope

The present document studies the Artificial Intelligence / Machine Learning (AI/ML) management capabilities and services for 5GS where AI/ML is used, including management and orchestration (e,g., MDA, see TS 28.104 [2]) and 5G networks (e.g., NWDAF, see TS 23.288 [3]).

# 2 References

The following documents contain provisions which, through reference in this text, constitute provisions of the present document.

- References are either specific (identified by date of publication, edition number, version number, etc.) or non‑specific.

- For a specific reference, subsequent revisions do not apply.

- For a non-specific reference, the latest version applies. In the case of a reference to a 3GPP document (including a GSM document), a non-specific reference implicitly refers to the latest version of that document *in the same Release as the present document*.

[1] 3GPP TR 21.905: "Vocabulary for 3GPP Specifications".

[2] 3GPP TS 28.104: "Management and orchestration; Management Data Analytics".

[3] 3GPP TS 23.288: "Architecture enhancements for 5G System (5GS) to support network data analytics services".

[4] 3GPP TS 28.105: " Artificial Intelligence / Machine Learning (AI/ML) management ".

# 3 Definitions of terms, symbols and abbreviations

## 3.1 Terms

For the purposes of the present document, the terms given in 3GPP TR 21.905 [1] ], TS 28.105 [4] and the following apply. A term defined in the present document takes precedence over the definition of the same term, if any, in 3GPP TR 21.905 [1].

**example:** text used to clarify abstract rules by applying them literally.

## 3.2 Symbols

For the purposes of the present document, the following symbols apply:

<symbol> <Explanation>

## 3.3 Abbreviations

For the purposes of the present document, the abbreviations given in 3GPP TR 21.905 [1] and the following apply. An abbreviation defined in the present document takes precedence over the definition of the same abbreviation, if any, in 3GPP TR 21.905 [1].

<ABBREVIATION> <Expansion>

# 4 Concepts and overview

## 4.1 Overview

Artificial Intelligence/Machine Learning (AI/ML) techniques are being embraced by telecommunication service providers around the world to facilitate enabling the existing and the new challenging use cases that 5G offers. AI/ML capabilities are being increasingly adopted in mobile networks as a key enabler for wide range of features and functionalities that maximise efficiency and bring intelligence and automation in various domains of the 5GS. For example, these include the Management Data Analytics (MDA) in the management and orchestration [1], the Network Data Analytics Function in the 5G core network domain [3].

The AI/ML-enabled functions in the 5GS use the AI/ML model for inference and in order to enable and facilitate the AI/ML adoption, the AI/ML model needs to be created, trained and then managed during its entire lifecycle.

To enable, facilitate and support AI/ML-capabilities in the 5GS, the following management capabilities are studied in this report:

* Validation of AI/ML model and AI/ML-enabled function
* Testing of AI/ML model and AI/ML-enabled function (before deployment)
* Deployment of AI/ML model (new or updated model) and AI/ML-enabled function
* Configuration of AI/ML-enabled function
* Performance evaluation of AI/ML-enabled function

NOTE: The AI/ML model training capability is specified in TS 28.105 [4].

## 4.2 AI/ML workflow for 5GS

AI/ML techniques are widely used in 5GS (including 5GC, NG-RAN and management system), and the generic workflow of the operational steps in the lifecycle of an AI/ML entity, is described below.

The workflow involves the training phases, inference phases, and the operation steps for each phase. These are briefly described below:

**Training phase:**

**AI/ML Training:** Learning by the Machine from the training data to generate the (new or updated) AI/ML entity (see TS 28.105 [4]) that could be used for inference. The AI/ML Training may also include the validation of the generated AI/ML entity to evaluate the performance variance of the AI/ML entity when performing on the training data and validation data. If the validation result does not meet the expectation (e.g., the variance is not acceptable), the AI/ML entity needs to be re-trained. This is the initial step of the workflow. The AI/ML Training MnS is specified in TS 28.105 [4].

**AI/ML Testing:** Testing of the validated AI/ML entity with testing data to evaluate the performance of the trained AI/ML entity for selection for inference. When the performance of the trained AI/ML entity meets the expectations on both training data and validation data, the AI/ML entity is finally tested to evaluate the performance on testing data. If the testing result meets the expectation, the AI/ML entity may be counted as a candidate for use towards the intended use case or task, otherwise the AI/ML entity may need to be further (re)trained. In some cases, the AI/ML entity may need to be verified which is the special case of testing to check whether it works in the AI-enabled function or the target node in other cases, this step may be skipped, for instance in case the input data and output data, data types and formats, have been unchanged from the last AI/ML entity.

**Inference phase:**

**AI/ML Inference:** performing the inference using the AI/ML entity.

Each operational step in the workflow is supported by the specific AI/ML management capabilities to including:

**- Management for training phase**

AI/ML training control: allowing the consumer to trigger and manage the model training/retraining based on the performance evaluation results observed by the model performance monitoring (performance data and/or feedback). For example, if the model performance decreases, the AI/ML performance management capability may trigger the AI/ML training to start retraining.

AI/ML testing management: allowing the consumer to initiate the AI/ML entity test and receive the testing results for a trained AI/ML model.

**- Management for inference phase**

AI/ML inference activation/deactivation: allowing the consumer to activate/deactivate the inference using AI/ML entity.

AI/ML inference monitoring: allowing the consumer to monitor and evaluate the inference performance of an AI/ML entity.

Editor’s note: more details and/or clarifications to the management tasks can be added later.

# 5 Use cases, potential requirements and possible solutions

## 5.1 AI/ML model performance management

### 5.1.1 Description

During AI/ML model training, test and deployment, the AI/ML model performance evaluation and management is needed. The related performance indicators need to be collected and analyzed. The purpose of AI/ML performance management is to find the problem, figure out what the problem is, and fix it in time to make sure the model can be trained, tested, and deployed healthy.

### 5.1.2 Use cases

#### 5.1.2.1 AI/ML model performance indicators

The AI/ML model performance indicators related to AI/ML model training, test and deployment need to be defined. The indicators mainly include three aspects:

Resource-related indicators: the performance indicators of the system that the model trains or deploys.

Model-related indicators: performance indicators of the model itself.

Service-related indicators: the running state indicators of the launched model.

These indicators need to be precisely defined. For different service, some indicators can be selected for evaluation. For example, resource-related indicators and model-related indicators may be selected in the training phase, while service-related indicators may be selected in the deployment phase. The AI/ML MnS producer should first determine which indicators are needed and then use these indicators for evaluation.

### 5.1.3 Potential requirements

**REQ-MODEL\_PERF-CON-1**  The AI/ML MnS producer should have a capability to define performance indicators of AI/ML model and select some indicators based on the service.

### 5.1.4 Possible solutions

TBD

## 5.2 Event data for ML training

### 5.2.1 Description

In analytics solutions, Performance Measurements (PMs) and Fault Reports (FRs) from various network function are collected and analytics applied on the PMs and FRs to come up with statistical insights and predictions of events from the raw data. For most algorithms, the prediction accuracy depends upon the amount of relevant historical data, motivating the need to store ever more data, which correspondingly increases the storage and processing resource requirements. However, not all recorded data is useful as the derived events, e.g. captured through analytics processes, may have loss of information OR mis-information e.g., with respect to time of the event.

### 5.2.2 Use cases

#### 5.2.2.1 Pre-processed event data for ML training

For AI/ML algorithms, a large amount of data points does not necessarily add value, e.g., if most of it includes biased data which ends up getting discarded during the pre-processing stages. Instead, the AI/ML algorithms need to have information-rich events data that is condensed but with most of it useful for the required training. For example, one could train an interference optimization solution that learns the best way to combat interference by looking at counters of handover failures correlated with signal quality. However, for most of the time in the radio network, there will be no interference events, but this cannot be determined if the events are not captured form the data. As such all the data must be kept and used for training. However, the data could also be mined for the interference event or signatures thereof. Then an equivalent interference management solution could be trained using the interference signatures or the signatures combined with only a small amount of raw data.

It is as such necessary to provide means to isolate and store the information rich events in the network, to ensure that minimizing storage and processing costs by discarding the unnecessary raw data does not compromise the ability to and still avails adequate historical information to adequately train AI/ML applications. In other words it is necessary for network functions to their management system to generate data on about the observed network events, e.g., based on the criteria set by the Operator, which events can then be stored to be used later to train AI/ML applications.

Network Function

AI/ML Trainig Function as Network events Data Consumer

Events processing

Events data

Management Function

Events processing

Network/system data

Network/system events data

Fig 2.2.1-1. Exposing and storing network events data

### 5.2.3 Potential requirements

**REQ-EVENT-DATA-1** The 3GPP management system shall enable an authorized consumer to request from the network data producer for network events corresponding to the data produced by that network data producer.

**REQ-EVENT-DATA-2** The 3GPP management system shall enable A network data producer to generate network events in place of or alongside the network data that they produce

**REQ-EVENT-DATA-3** The 3GPP management system shall enable a network events aggregator to take the events from different network entities and re-expose them in an aggregated way that eliminates duplications

### 5.2.4 Possible solutions

TBD

## 5.3 AI/ML entity validation

### 5.3.1 Description

During the AI/ML training process, the generated AI/ML entity (see TS 28.105 [4]) needs to be validated. The purpose of AI/ML validation is to evaluate the performance of the AI/ML entity when performing on the validation data, and to identify the variance of the performance on the training data and the validation data. If the variance is not acceptable, the entity would need to be tuned (re-trained) before being made available to the consumer.

The training data and validation are normally split from the same data set with a certain ratio in terms of the quantity of the data examples, therefore they have the same pattern.

### 5.3.2 Use cases

#### 5.3.2.1 AI/ML entity validation performance reporting

In the AI/ML training, the AI/ML entity is generated based on the learning from the training data, and validated using validation data. The performance of the AI/ML entity has tight dependency on the data (i.e., training data) from which the AI/ML entity is generated. Therefore, an AI/ML entity performing well on the training data may not perform well on other data. If the performance of AI/ML entity is not good enough as result of AI/ML validation, the AI/ML entity will be tuned (re-trained) and validated again. The process of AI/ML entity generation and validation is repeated by the AI/ML training function, until the performance of the AI/ML entity meets the expectation on both training data and validation data. The producer in the end selects one or more AI/ML entities with the best level performance on both training data and validation data as the result of the AI/ML training, and reports to the consumer. The performance of each selected AI/ML entity on both training data and validation data also need to be reported.

The performance result of the validation may also be impacted by the ratio of the training data and validation data. Consumer needs to be aware of the ratio of training data and validation data, besides the performance score on each data set, in order to be confident about the performance of AI/ML entity.

### 5.3.3 Potential requirements

**REQ-MODEL\_VLD-CON-1** The AI/ML Training MnS producer should have a capability to validate the AI/ML entities during the training process, and report the performance of the AI/ML entities on both the training data and validation data to the authorized consumer.

**REQ-MODEL\_VLD-CON-2** The AI/ML Training MnS producer should have a capability to report the ratio (in terms of the quantity of the data examples) of the training data and validation data used for training of an AI/ML entity during the training process.

### 5.3.4 Possible solutions

TBD

5.4.5 Evaluation

TBD

## 5.4 AI/ML entity testing

### 5.4.1 Description

After an AI/ML Entity is trained, validation is done to ensure the training process is completed successfully.. However, even when validation is conducted successfully during development, it is necessary to test and check if the AI/ML Entity is working correctly under certain runtime contexts or using certain testing data set. Testing may involve interaction with third parties (besides the developer of the AI/ML training function) , e.g. the operators may use the AI/ML training function or third-party systems/functions that may rely on the results computed by the AI/ML entity for testing.

After completing the AI/ML entity training, and when the performance of the trained AI/ML entity meets the expectations on both training and validation data, the AI/ML entity is made available to the consumer(s) via the AI/ML training report (see AIMLTrainingReport IOC in TS 28.105 [4]). Before applying the AI/ML entity to the target inference function or AI/ML-enabled function, the AI/ML training MnS producer may need to allow the consumer to evaluate the performance of the AI/ML entity via the AI/ML testing process using the consumer’s provided testing data. The testing data have the same pattern as the input part of the training data.

For these reasons, provision of AI/ML testing and its control need to be standardized to enable the multi-vendor interaction among the different systems. If the testing performance is not acceptable or does not meet the pre-defined requirements, the consumer may request the AI/ML training producer to re-train the AI/ML entity with specific training data and/or performance requirements.

### 5.4.2 Use cases

#### 5.4.2.1 Consumer-requested AI/ML entity testing

After receiving an AI/ML training report about a trained AI/ML entity from the AI/ML Training MnS producer, the consumer may request the testing MnS producer to test the AI/ML entity before applying it to the target inference function. In the AI/ML testing request, the consumer provides the testing data which have the same pattern as the input part of the training data.

Any AI/ML entity needs to be tested with specific inputs and features that are applicable to the use case and the applicable deployment environment.

The AI/ML testing MnS producer performs the AI/ML testing using the consumer’s provided testing data. The AI/ML testing is to conduct inference on the tested AI/ML entity using the testing data as the inference inputs and produce the inference output for each testing dataset example.

The AI/ML testing MnS producer may be the same as or different from the AI/ML Training MnS producer.

After completing the AI/ML testing, the AI/ML testing MnS producer provides the testing report indicating the success or failure of the AI/ML testing to the consumer. For a successful AI/ML testing, the testing report contains the testing results, i.e., the inference output for each testing dataset example.

The AI/ML testing MnS producer needs to have the capabilities and provide the services needed to enable the consumer to request testing and receive results on the testing of a specific AI/ML entity or of an application or function that contains an AI/ML entity.

To achieve the desired outcomes, any AI/ML entity needs to be tested with the appropriate testing data, which can reflect the current status of the network where the AI/ML entity is expected to be deployed. Correspondingly, the AI/ML testing MnS producer needs to support the required management services to test the AI/ML entities.

#### 5.4.2.2 Control of AI/ML entity testing

Given a testing capability as provided by a given AI/ML testing MnS producer, a consumer (e.g., an operator) may wish to control and manage that testing process capability. For example, the operator may wish to define policies on how frequent testing for a given AI/ML entity may be executed. Correspondingly, the 3GPP management system needs to provide the capability to allow the AI/ML entity testing to be configured.



Figure 5.4.2.2-1: AI/ML entity testing and control

### 5.4.3 Potential requirements

**REQ-AI/MLTEST-1** The AI/ML testing MnS producer should have a capability for an authorized consumer to request the testing of a specific AI/ML entity.

**REQ-AI/MLTEST-2** The AI/ML testing MnS producer should have a capability to create a testing process instance per the testing request for an authorized consumer.

**REQ-AI/MLTEST-3** The AI/ML testing MnS producer should have a capability to report to an authorized consumer the results of a specific instance of AI/ML testing process with the result of a successful AI/ML entity testing containing the inference output for each testing data example.

**REQ-AI/MLTEST-4** The AI/ML testing MnS producer should have a capability for an authorized consumer (e.g. the operator) to configure or modify an instance of AI/ML testing process.

**REQ-AI/MLTEST-5** The AI/ML testing MnS producer should have a capability to test a specific AI/ML entity using specific data specified by the consumer or using data at a location address specified by the consumer or using data with specific characteristics defined by the consumer.

**REQ-AI/MLTEST-6** The AI/ML testing MnS producer should have a capability to test a specific AI/ML entity for a specified expected runtime context as may be stated by the consumer.

**REQ-AI/MLTEST-8** The AI/ML testing MnS producer should support a capability for an authorized consumer to define the reporting characteristics related to a specific instance of AI/ML testing request.

**REQ-AI/MLTEST-9** The AI/ML testing MnS producer should support a capability for an authorized consumer to manage the AI/ML testing request, including suspending, resuming, canceling the request, or adjusting the desired runtime context of the testing.

### 5.4.4 Possible solutions

TBD

## 5.5 AI/ML deployment

### 5.5.1 Description

AI/ML deployment refers to the process of making an AI/ML-enabled function available in the operational environments, where it could start adding value by conducting inference (e.g., prediction). After trained AI/ML-enabled function meets the performance criteria, the AI/ML-enabled function could be deployed in 3GPP system via a software installation or via a configuration management procedure and subsequently activated.

After an AI/ML-enabled function is deployed and activated in the production environment, the data fed to the AI/ML-enabled function may change to the level where it is different from the data used in the initial prior training of the respective AI/MLEntity. To improve model performance with the changing data, the AI/ML-enabled function or the entity therein may need to be retrained and redeployed within the production environment.

### 5.5.2 Use cases

#### 5.5.2.1 AI/ML deployment information retrieved by consumer

After the AI/ML Entity is trained by the AI/ML training (AIMLT) MnS producer during development, the AI/ML-enabled function needs to be deployed to conduct inference in the operational environment.

Once the AI/ML-enabled function has been installed in the operational environment, as different consumers need to know the available information of AI/ML-enabled functions and to determine the next appropriate action, the MnS consumer may request the MnS producer to retrieve the deployment information of AI/ML-enabled function.

The general deployment information used to describe an AI/ML-enabled function may include:

- Resource information, which describes the static parameters of AI/ML Entity (e.g. aIMLEntityVersion, aIMLEntityId, trainingContext, see TS 28.105 [4])

- Management information, which describes the information model that is used for AI/ML Entity lifecycle management (e.g. activation flag, status, create time, last update time).

- Capability information, which describes the capability information (e.g. inference type, performance metrics).

### 5.5.3 Potential requirements

**REQ-MODEL\_DPL-CON-1** The MnS producer responsible for AI/ML management shall have a capability to retrieve deployment information of the AI/ML-enabled function and inform an authorized consumer about the information.

**REQ-MODEL\_DPL-CON-2** The MnS producer responsible for AI/ML management shall have a capability to allow an authorized consumer to retrive deployment information of the AI/ML-enabled function.

### 5.5.4 Possible solutions

TBD

### 5.5.5 Evaluation

TBD

## 5.6 AI/ML Inference History

### 5.6.1 Description

For different automation requirements, network and management automation functions (e.g. gNB, MDAS, SON) may apply Machine Learning functionality to make the appropriate inferences in different contexts. Depending on the contexts, the AI/MLEntity may take different decisions at inference with different outcomes. The history of such inference decisions and the context within which they are taken may be of interest to different consumers.

### 5.6.2 Use cases

#### 5.6.2.1 Tracking AI/ML inference decision and context

The AI/ML-enabled function may take different decisions at inference in different contexts and with different outcomes. The selected decisions may need to be tracked for future reference, e.g. to evaluate the appropriateness/ effectiveness of the decisions for those contexts or to evaluate degradations in the AI/MLEntity's decision-making capability. For this, the network not only needs to have the required inference capabilities but needs also to have the means to track and enable usage of the history of the inferences made by the ML applications.



**Figure 5.6.2.1-1: Example use and control of MLInferenceHistory Request and reporting in the management plane**

### 5.6.3 Potential requirements

**REQ-MLHIST-1** The producer of ML inference history should have a capability allowing an authorized consumer to request the inference history of a specific AI/MLEntity.

**REQ-MLHIST-3** The producer of ML inference history should support a capability for an authorized consumer (e.g. the function/entity that generated the Request for ML inference history) to define the reporting characteristics related to a specific instance of ML inference history or the reporting thereof.

### 5.6.4 Possible solutions

TBD

## 5.7 AI/ML context

### 5.7.1 Description

AIMLContext represents the status and conditions related to the AIMLEntity (cf. TS 28.105[4]). This may include the network context as defined in TS 28.104 [2] as well as other conditions that may be applicable to the AIMLEntity but are not part of network characteristics e.g. the time of day, season of the year. As part of AI/ML model performance management there is identification of the problem that the AI/ML model is facing. As described in TS 28.104[2], the differences in the network context, i.e., network status, under which data is collected to produce analytics, significantly affect the produced analytics. Similarly, the changes in the AI/ML context, e.g., the characteristics of the data related to the network status and conditions used for AI/ML model training, testing and deployment may affect the AI/ML entity performance, thus may represent a problem for the AI/ML entity. Thus management capabilities are needed to enable awareness of the AI/ML context in terms of the identification as well as monitoring and reporting of changes in AI/ML context as part of the identification of the problem that the AI/ML entity is facing.

### 5.7.2 Use cases

#### 5.7.2.1 AI/ML context monitoring and reporting

AI/ML context related to AI/ML model training, testing and deployment needs to be identified by characterizing the input data used by the AI/ML model is targeted to work. As an example, such characterization may be done based on the statistical properties of data. Monitoring of such AI/ML context serves to detect the changes and anomalies in the AI/ML context. Some anomalies may be considered as a problem that AI/ML entity is facing as it may lead to its performance degradation. Therefore, the consumer of the related AI/ML service needs to be informed about such observed AI/ML context change.

### 5.7.3 Potential requirements

**REQ-1** The producer AI/ML-related MnS including for training and inference should have a capability to identify and monitor the AI/ML context, as well as to inform the consumer about observed changes in AI/ML context ,

### 5.7.4 Possible solutions

TBD

## 5.8 AIMLEntity Capability Discovery and Mapping

### 5.8.1 Description

A network or management function that applies AI/ML to accomplish specific tasks may be considered to have one or more AI/MLEntities each having specific AIML capabilities. The AIML capabilities are either of:

* a decision-making capability which is in the form of triple <x,y,z> indicating
  + x: the object or object types for which the AI/MLEntity can undertake optimization or control
  + y: the configurable attributes on object or object types x, which the AI/MLEntity optimizes or controls to achieve the desired outcomes
  + z: the performance metrics which the AI/MLEntity optimizes through its actions
* an analysis capability which is in the form of tuple <x,z> indicating
  + x: the object or object types for which the AI/MLEntity can undertakes analysis
  + z: the network context (on object x) for which the ML app produces analysis

Different network functions may need to rely on existing AI/ML capabilities to accomplish the desired automation. However, the applicability of the ML-based solutions and the details of such ML-based solutions (i.e., which AI/MLEntities shall be applied and how) for accomplishing those automaton functionalities is not obvious. On a high-level, such ML-based solutions may be categorized into cases with or those without ML orchestration. In both cases, management services are required to identify the capabilities of the involved AI/MLEntities and to map those capabilities to the desired logic.

### 5.8.2 Use cases

#### 5.8.2.1 Identifying capabilities of AI/MLEntities

Network functions, especially network automation functions, may need to rely on AI/ML capabilities that are not internal to those Network functions to accomplish the desired automation. For example, as stated in TS 28.104, “an MDA Function may optionally be deployed as one or more AI/ML-enabled function(s) in which the relevant models are used for inference per the corresponding MDA capability.” Similarly, owing to the differences in the kinds and complexity of intents that need to be fulfilled, an intent fulfilment solution may need to employ the capabilities of existing AI/ML to fulfil the intents. In any such case, management services are required to identify the capabilities of those existing AI/MLEntities.

#### 5.8.2.2 Mapping of the capabilities of AI/MLEntities

Besides the discovery of the capabilities of AI/MLEntities, services are needed for mapping the AI/ML Entities and capabilities. In other words, instead of the consumer discovering specific capabilities, the consumer may want to know the AI/MLEntities that can be used to achieve a certain outcomes. For this, the producer should be able to inform the consumer of the set of AI/MLEntities that together achieve the consumer's automomation needs.

In the case of intents for example, the complexity of the stated intents may significantly vary - from simple intents which may be fulfilled with a call to a single AI/MLEntity to complex intents that may require an intricate orchestration of multiple AI/MLEntities. For simple intents, it may be easy to map the execution logic to the one or multiple AI/MLEntities For complex intents, it may be required to employ multiple AI/MLEntities along with a corresponding functionality that manages their inter-related execution . The usage of the AI/MLEntities requires the awareness of the capabilities of their capabilities and interrelations

Moreover, given the complexity of the required mapping to the multiple AI/MLEntities, services should be supported to provide the mapping of AI/MLEntities and capabilities.

### 5.8.3 Potential requirements

**REQ-MLCAP-1** The 3GPP Management system should have a capability for an authorized consumer to request the AI/ML MnS Producer for the capabilities of existing AI/MLEntities available within the provider of AI/ML inference.

**REQ-MLCAP-2** The AI/ML MnS Producer shall have a capability to report to an authorized consumer the capabilities of an AI/MLEntity as a decision described as a triplet <object(s), parameters, metrics> with the entries respectively indicating : the object or object types for which the AI/MLEntity can undertake optimization or control; the configuration parameters on the stated object or object types, which the ML app optimizes or controls to achieve the desired outcomes; and the network metrics which the ML app optimizes through its actions.

**REQ-MLCAP-3** The AI/ML MnS Producer shall have a capability to report to an authorized consumer the capabilities of an AI/MLEntity as an analysis described as a tuple <object(s), characteristics> with the entries respectively indicating : the object or object types for which the ML app can undertake analysis; and the network characteristics (related to the stated object or object types) for which the ML app produces analysis.

### 5.8.4 Possible solutions

TBD

## 5.9 AI/ML update management

### 5.9.1 Description

Due to the complexity and time-varying nature of network, the AI/ML entities previously deployed may no longer be applicable to the current network after running for a period of time. Typically, the performance of a trained model may degrades over time. The AI/ML entitie needs to be updated timely to ensure the performance of inference and analysis.

### 5.9.2 Use cases

#### 5.9.2.2 AI/ML entities updating initiated by producer

The AI/ML entity updating may be initiated by the AI/ML MnS producer. In order to keep the model at a requested level, the AI/ML MnS producer may periodically conduct AI/ML retraining with new available training data. Once a new version AI/ML entity is obtained after the training is finished, it can be used to update the current AI/ML entity with this new version. In another condition, the AI/ML MnS producer may initiate AI/ML entity updating based on the running model performance. For example, if the performance of the running AI/ML model is decreased under a predefined threshold, the AI/ML MnS producer may decide to start ratraining and then update the AI/ML entity to a new version which performs better.

Diagram

Description automatically generated

5.9.2.2-1: Potential requirements

|  |  |  |
| --- | --- | --- |
| **Requirement label** | **Description** | **Related use case(s)** |
| **REQ-AIML\_UPD-CON-3** | The AI/ML MnS producer should have a capability to update the AI/ML entities and inform an authorized consumer about the update status. | AI/ML entities updating initiated by producer (clause 5.X.2.2) |

### 5.9.4 Possible solutions

TBD

Annex X (informative):  
Change history

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Change history** | | | | | | | |
| **Date** | **Meeting** | **TDoc** | **CR** | **Rev** | **Cat** | **Subject/Comment** | **New version** |
| 2022-03 | SA5#142e | n/a | - | - | - | Initial skeleton | 0.0.0 |
| 2022-04 | SA5#142e | S5-222188 | - | - | - | Add scope | 0.1.0 |
| 2022-04 | SA5#142e | S5-222477 | - | - | - | Add AI/ML management overview | 0.1.0 |
| 2022-05 | SA5#143e | S5-223576 | - | - | - | Add use case on AI/ML model performance | 0.2.0 |
| 2022-07 | SA5#144e | S5-224364 | - | - | - | pCR 28.908 Requirements on AIML context | 0.3.0 |
| 2022-07 | SA5#144e | S5-224365 | - | - | - | pCR 28.908 Add AI-ML workflow for 5GS | 0.3.0 |
| 2022-07 | SA5#144e | S5-224366 | - | - | - | pCR TR 28.908 Clarifications into the overview and terminologies updates | 0.3.0 |
| 2022-07 | SA5#144e | S5-224367 | - | - | - | pCR 28.908 Requirements on AIML Inference History | 0.3.0 |
| 2022-07 | SA5#144e | S5-224368 | - | - | - | pCR 28.908 Add use case on AI-ML entity validation | 0.3.0 |
| 2022-07 | SA5#144e | S5-224369 | - | - | - | pCR 28.908 Requirements on AIML Testing | 0.3.0 |
| 2022-07 | SA5#144e | S5-224386 | - | - | - | pCR 28.908 Requirements on AIMLEntity Capability Discovery | 0.3.0 |
| 2022-07 | SA5#144e | S5-224370 | - | - | - | pCR 28.908 Requirements on Pre-processed event data for ML training | 0.3.0 |
| 2022-07 | SA5#144e | S5-224371 | - | - | - | pCR TR 28.908 Add use case on AI-ML model update | 0.3.0 |
| 2022-07 | SA5#144e | S5-224372 | - | - | - | pCR TR 28.908 Add use case on AI-ML model deployment | 0.3.0 |