**3GPP SA Meeting #105 *SP-24xxxx***

Malborne, Australia, 10 - 13 September 2024

**Source: SA5**

**Title: pCR to TR 22.850 SA study on AI/ML Consistency Alignment**

**Document for: Approval**

**Agenda Item: 7**

**3GPP TSG-SA5 Meeting #156 *S5-244584d1***

Maastricht, Netherlands, 19 - 23 August 2024

**Source: NEC, Intel, Deutsche Telekom**

**Title: potential pCR to TR 22.850 SA study on AI/ML Consistency Alignment**

**Document for: Discussion & endorsement**

**Agenda Item: 5.2 Technical issues at SA5 level**

# 1 Decision/action requested

***The group is asked to discuss and approve the proposed input***

# 2 References

[1] [SP-240970](https://www.3gpp.org/ftp/TSG_SA/TSG_SA/TSGS_104_Shanghai_2024-06/Docs/SP-240970.zip), Study on 3GPP AI/ML Consistency Alignment

[2] 3GPP TR 22.850; Study on 3GPP AI/ML Consistency Alignment

[3] 3GPP [TS 28.105](https://www.3gpp.org/ftp/Specs/archive/28_series/28.105/28105-i40.zip) Management and orchestration; Artificial Intelligence/ Machine Learning (AI/ML) management

[4] [SP-240954](https://www.3gpp.org/ftp/TSG_SA/TSG_SA/TSGS_104_Shanghai_2024-06/Docs/SP-240954.zip), Input to the TSG SA level discussion on potential study on AI/ML E2E framework

[5] [SP-201084](https://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGs_90E_Electronic/Docs/SP-201084.zip), latest terms of reference for SA WG5 - Management, Orchestration and Charging

# 3 Rationale

The discussions during SA#104 led to the approval of the new SID on 3GPP AI/ML Consistency Alignment [1]. A draft skeleton for the study TR 22.850 is being discussed [2].

The study aims to achieve consistent terminology and an E2E framework, particularly from the lifecycle management (LCM) perspective for AI/ML models, and the cross-domain collection and storage of training/inference data.

This study will investigate ongoing AI/ML work within TSG CT, TSG RAN, and TSG SA Working Groups, identifying instances of potential misalignment and/or inconsistencies.

* Based on the discussion conducted prior and during the SA-level study proposal, it is probably fair to conclude at this stage that SA5 is among the leading 3GPP WGs that have managed to develop and accomplish stable AI/ML related specifications for Rel-18 [3,4].
* The specifications [2] focused on the management and orchestration aspects of AI/ML features and capabilities in the 5G system addressing various aspects including e.g. terminology and ML model LCM. These components form an integral part of the of SA study on 3GPP AI/ML Consistency Alignment.
* In the context of the 3GPP system, lifecycle management (LCM) of network functions (NFs), features, entities, and ML models primarily involves management tasks. These tasks fall under the scope of SA5's Terms of Reference (ToR) [5]. SA5 has already completed numerous specifications addressing LCM, including the recently finalized Release-18 specification on AI/ML management [3]. This extensive work highlights SA5's expertise, foundational achievements and responsibilities to AI/ML lifecycle management within 3GPP.

Therefore, SA5 believes that an input summarising relevant work accomplished by SA5 on AI/ML management, including AI/ML-related terms and LCM would provide valuable initial input for the SA study TR 22.850 on 3GPP AI/ML Consistency Alignment. This input would help ensure that the study is grounded in the comprehensive and established work already completed by SA5, promoting consistency and alignment across 3GPP AI/ML initiatives.

This pCR propose an input to TR 22.850 Study on 3GPP AI/ML Consistency Alignment capturing SA5’s normative work accomplishments on the topic of AI/ML management that directly relevant to the scope and objective the SA study.

# 4 Detailed proposal

***1st change***

2 References

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

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

[x1] [SP-201084](https://www.3gpp.org/ftp/tsg_sa/TSG_SA/TSGs_90E_Electronic/Docs/SP-201084.zip), latest terms of reference for SA WG5 - Management, Orchestration and Charging

[x2] 3GPP [TS 28.105](https://www.3gpp.org/ftp/Specs/archive/28_series/28.105/28105-i40.zip) Management and orchestration; Artificial Intelligence/ Machine Learning (AI/ML) management

[x3] 3GPP [TR 28.908](https://www.3gpp.org/ftp/Specs/archive/28_series/28.908/28908-i00.zip): “Study on Artificial Intelligence/Machine Learning (AI/ ML) management”

[x4] [SP-231780](https://www.3gpp.org/ftp/TSG_SA/TSG_SA/TSGS_102_Edinburgh_2023-12/Docs/SP-231780.zip), New SID: Study on AI/ML management - phase 2

[x5] 3GPP TR 28.858; AI/ML management – phase 2

***Next change***

5 AI/ML related activities in all Working Groups

5.1 General

This clause will investigate and identify AI/ML related activities of all working groups of Rel-18 features and Rel-19 studies and work items in TSG CT, TSG RAN and TSG SA Working Groups.

Editor's Note: The AI/ML related content captured in TR 21.918 ("Release 18 Description; Summary of Rel-18 Work Items") can be considered as a starting point.

5.1.1 SA5

**Outline of Activities in 3GPP SA5**

SA5 is responsible for the management, orchestration, and charging for 3GPP systems [x1], covering a wide range of OAM aspects, including operation, assurance, fulfilment, and automation. The scope of responsibilities encompasses services and functions supporting orchestration, assurance, and analytics, including Life Cycle Management (LCM).

To support and facilitate the efficient deployment and operation of AI/ML capabilities and features in the 5G system (including Management and Orchestration systems, 5GC, and NG-RAN domains), the ML model needs to be managed throughout its entire lifecycle.

**AI/ML management related activities:**

* **Release-18 Specifications:**
	+ Documented in TS 28.105 [x2].
	+ Included the following AI/ML-related terminologies and definitions.
		- ML model
		- ML model training
		- ML model initial training
		- ML model re-training
		- ML model joint training
		- ML training function
		- ML model testing
		- ML testing function
		- AI/ML inference
		- AI/ML inference function
		- AI/ML inference emulation
		- ML model deployment
	+ Provided a general framework for the life cycle management (LCM) of ML models for 5GC, NG-RAN and management domains, covering the following operational steps:
		- ML model training (including validation)
		- ML model testing
		- AI/ML emulation
		- ML model deployment
		- AI/ML inference
	+ Captured detailed use cases for each step of the LCM workflow, along with relevant requirements and solutions (stage 2 and stage 3).
* **Ongoing Release-19 Study:**
	+ Phase 2 of AI/ML management, documented in TR 28.858 [x5].
	+ Aims to investigate additional advanced use cases covering AI/ML management and operational capabilities to support different types of AI/ML technologies required for AI/ML in the 5G system, such as:
		- Federated Learning
		- Reinforcement Learning
		- Online and Offline Training
		- Distributed Learning
		- Generative AI
	+ Focuses on the sustainability aspect of AI/ML, including the evaluation of energy consumption and efficiency impacts associated with AI/ML solutions across all operational phases (training, emulation, deployment, inference).Management aspects (LCM CM and PM) of additional AI/ML functionalities defined by 3GPP SA WGs and RAN WGs.

***Next change***

5.2 AI/ML related activities in TSG SA Working Groups

5.2.X AI/ML related activities in TSG SA5

5.2.X.1 Description

Editor's Note: This clause will investigate and identify AI/ML related activities of all TSG SA working groups of Rel-18 features and Rel-19.

5.2.X.2 Terminology

Editor's Note: This clause describes AI/ML related terminology (i.e. set of definitions, acronyms)

The following terms and definitions have been developed and documented in Rel-18 TS 28.105 [x2].

***Start of quoted text (From TS 28.105)***

**ML model:** a manageable representation of an ML model algorithm.

NOTE 1: an ML model algorithm is a mathematical algorithm through which running a set of input data can generate a set of inference output.

NOTE 2: ML model algorithm is proprietary and not in scope for standardization and therefore not treated in this specification.

NOTE 3: ML model may include metadata. Metadata may include e.g. information related to the trained model, and applicable runtime context.

**ML model training:** a process performed by an ML training function to take training data, run it through an ML model algorithm, derive the associated loss and adjust the parameterization of that ML model iteratively based on the computed loss and generate the trained ML model.

**ML model initial training:** a process of an initial version of an ML model.

**ML model re-training:** a process of training a previous version of an ML model and generate a new version.

NOTE 4: a new version of a trained ML model supports the same type of inference as the previous version of the ML model, i.e., the data type of inference input and data type of inference output remain unchanged between the two versions of the ML model, but parameter values might be different for the re-trained model.

**ML model joint training:** a process of training a group of ML models.

**ML training function**: a logical function with ML model training capabilities.

**ML model testing:** a process of testing an ML model using testing data.

**ML testing function**: a logical function with ML model testing capabilities.

**AI/ML inference**: a process of running a set of input data through a trained ML model to produce set of output data, such as predictions.

NOTE 5: the inference represents the process to realize the AI capabilities by utilizing a trained ML model and other AI enablers if needed, hence the AI/ML prefix is used when referring to inference as compared to training and testing.

**AI/ML inference function**: a logical function that employs trained ML model(s) to conduct inference.

**AI/ML inference emulation**: running the inference process to evaluate the performance of an ML model in an emulation environment before deploying it into the target environment.

**ML model deployment:** a process of making a trained ML model available for use in the target environment.

***End of quoted text (From TS 28.105)***

##### 5.2.X.2.1 Observations and analyses: terminology

* Developing a unified set of terms for AI/ML specifications forms the baseline for defining unified 3GPP AI/ML framework and help towards preventing subsequent duplication and inconsistency of the work among the 3GPP WGs.
* The terms documented in TS 28.105[x2] could be considered as guideline terms, with all or at least a subset of them can serving as a baseline by the 3GPP WGs. The 3GPP WGs may of course still define additional AI/ML related terms if deemed necessary.
* Based on the above, it is recommended that the 3GPP adopts a harmonised set of AI/ML-related terminologies as much as possible, taking into consideration the definitions introduced by SA5 in TS 28.105 [x2]. Effort should be made to avoid developing different definitions for the same terms, unless necessary to address the specific needs of AI/ML domain-specific specifications being developed by different WGs.

5.2.X.3 ML model life cycle management (LCM)

Rel-18 specification [x2] addressed the AI/ML LCM management capabilities (including wide range of use cases, corresponding requirements (stage 1) and solutions (stage 2 NRMs & stage 3 OpenAPIs) for the ML model, including ML model training (which also includes validation), testing, AI/ML inference emulation, deployment and AI/ML inference steps of the lifecycle as shown below for managing the entire lifecycle of the ML model:

***Start of quoted text (from TS 28.105)***

4a.0 ML model lifecycle

AI/ML techniques are widely used in 5GS (including 5GC, NG-RAN, and management system), the generic AI/ML operational workflow in the lifecycle of an ML model, is depicted in Figure 4a.0-1.



**Figure 4a.0-1: ML model lifecycle**

The ML model lifecycle includes training, testing, emulation, deployment, and inference. These steps are briefly described below:

**- ML model training:** training, including initial training and re-training, of an ML model or a group of ML models. It also includes validation of the ML model to evaluate the performance when the ML model performs on the training data and validation data. If the validation result does not meet the expectation (e.g., the variance is not acceptable), the ML model needs to be re-trained.

**- ML model testing:** testing of a validated ML model to evaluate the performance of the trained ML model when it performs on testing data. If the testing result meets the expectations, the ML model may proceed to the next step If the testing result does not meet the expectations, the ML model needs to be re-trained.

**- AI/ML inference emulation:** running an ML model for inference in an emulation environment. The purpose is to evaluate the inference performance of the ML model in the emulation environment prior to applying it to the target network or system.

NOTE: The AI/ML inference emulation is considered optional and can be skipped in the AI/ML operational workflow.

**- ML model deployment:** ML model deployment includes the ML model loading process (a.k.a. a sequence of atomic actions) to make a trained ML model available for use at the target AI/ML inference function.

ML model deployment may not be needed in some cases, for example when the training function and inference function are co-located.

**- AI/ML inference:** performing inference using a trained ML model by the AI/ML inference function. The AI/ML inference may also trigger model re-training or update based on e.g., performance monitoring and evaluation.

***End of quoted text***

##### 5.2.x.3.1 Observations and analyses: AI/ML LCM

* The AI/ML workflow defined by SA5 [x2] represents a general framework encapsulating the various life cycle management (LCM) operations for ML model (i.e., model training, testing, emulation, deployment, and inference).
* The AI/ML LCM capabilities defined by SA5 for each of the operational steps are generic for managing of 3GPP system including the Management and orchestration, CN and RAN domains.
* It is important to recognise that “domain-specific” ML model life cycle related tasks can be developed for the specific domains by the relevant 3GPP WGs, e.g., the RAN WGs can specify data collection within the RAN domain needed to train the UE-side, network-side, or the two-sided UE/network ML models, and specific LCM operations for UE-side model over air-interface.
* While ML model and AI/ML inference function life cycle can be specified by the relevant 3GPP WG for the specific domain (i.e., RAN, CN or Management & Orchestration), the “management aspects” of life cycle (i.e., life cycle *management*) remains to be primarily a “management task” that falls within the responsibility of SA5.
* The ML models and the associated “Life Cycle” can be a use case and/or domain specific, the management of the Life Cycle (i.e., LCM) is a higher layer task which is typically a role of the OAM that encompasses the process of e.g., the governance, automation, and operational practices applied to the entire AI/ML lifecycle. It is therefore imperative to distinguish the difference between Life Cycle and Life Cycle Management.
* Where feasible, the ML model LCM workflow and associated management capabilities specified by SA5 in TS 28.105 [x2] could be considered by 3GPP for the currently ongoing and future relevant specification development. The 3GPP WG(s) should potentially provide AI/ML LCM-related requirements, if any, to SA5 to avoid duplication and contention of effort.

**NOTE:** SA5 Rel-18 specification in TS 28.105 on ML model LCM and the associated management capabilities [x2] does not address the UE-side and UE/Network-side Model LCM.

#### 5.2.x.4 ML model lifecycle management capabilities

Each step in the ML model lifecycle. i.e., the ML model training, ML model testing, AI/ML emulation, ML model deployment and AI/ML inference correspond to number of dedicated management capabilities. The specified capabilities are developed based on corresponding use cases and requirements. The management capabilities specified by SA5 [x2] are highlighted below:

***Start of quoted text (From TS 28.105)***

6.1 ML model lifecycle management capabilities

Each operational step in the ML model lifecycle (see clause 4a.0.1) is supported by one or more AI/ML management capabilities as listed below.

**Management capabilities for ML model training**

**- ML model training management**: allowing the MnS consumer to request the ML model training, consume and control the producer-initiated training, and manage the ML model training/re-training process. The training management capability may include training performance management and setting a policy for the producer-initiated ML model training.

**-** ML model training capability also includes validation to evaluate the performance of the ML model when performing on the validation data, and to identify the variance of the performance on the training and validation data. If the variance is not acceptable, the ML model would need to be re-trained before being made available for the next step in the operational workflow (e.g., ML model testing).

**Management capabilities for ML testing**

**- ML model testing management**: allowing the MnS consumer to request the ML model testing, and to receive the testing results for a trained ML model. It may also include capabilities for selecting the specific performance metrics to be used or reported by the ML testing function. MnS consumer may also be allowed to trigger ML model re-training based on the ML model testing performance results.

**Management capabilities for AI/ML inference emulation:**

* **AI/ML inference emulation:** a capability allowing an MnS consumer to request an ML inference emulation for a specific ML model or models (after the training, validation, and testing) to evaluate the inference performance in an emulation environment prior to applying it to the target network or system.

**Management capabilities for ML** model **entity deployment:**

**- ML entity loading management**: allowing the MnS consumer to trigger, control and/or monitor the ML model loading process.

**Management capabilities for AI/ML inference:**

**- AI/ML inference management:** allowing an MnS consumer to control the inference, i.e., activate/deactivate the inference function and/or ML model/models, configure the allowed ranges of the inference output parameters. The capabilities also allow the MnS consumer to monitor and evaluate the inference performance and when needed trigger an update of an ML model or an AI/ML inference function.

The use cases and corresponding requirements for AI/ML management capabilities are specified in the following clauses.

***End of quoted text***

##### 5.2.x.4.1 Observations and analyses: ML model lifecycle management capabilities

* ML model lifecycle management (LCM) capabilities are crucial for the effective deployment, operation, and optimization of AI/ML-enabled features and capabilities in both the NG-RAN and 5GC. These capabilities ensure that AI/ML models are not only developed and trained correctly but also tested, deployed, evaluated, and operated efficiently in the network environment.
* The management capabilities outlined in TS 28.105 offer a structured approach to managing the various steps of the ML model lifecycle. This structured approach is applicable to AI/ML-enabled features and capabilities in NG-RAN, 5GC, and management system, ensuring consistency and reliability in the deployment and operation of AI/ML technologies for different domains.
* The AI/ML LCM management capabilities are foundational for integrating advanced AI/ML features into 5G networks. By ensuring that ML models are effectively managed from the training step through to inference, these capabilities provide robust and reliable AI/ML-driven network enhancements.
* The AI/ML LCM workflow and associated management capabilities specified by SA5 in [x2] should be considered as the baseline for the AI/ML E2E framework for the 3GPP. These capabilities provide a comprehensive foundation for ensuring that AI/ML models and related processes are consistently managed across all steps of their lifecycle, promoting seamless integration and operation for all domain within the 5G system.

#### 5.2.x.5 AI/ML functionalities management scenarios

The Rel-18 specification [x2] also documented AI/ML functionalities management scenarios in relation with managed AI/ML features which describe the possible locations of ML training function and AI/ML inference function involving the various 3GPP system domains.

***Start of quoted text (From TS 28.105)***

4a.2 AI/ML functionalities management scenarios (relation with managed AI/ML features)

The ML training function and/or AI/ML inference function can be located in the RAN domain MnS consumer (e.g. cross-domain management system) or the domain-specific management system (i.e. a management function for RAN or CN), or Network Function.

For MDA, the ML training function can be located inside or outside the MDAF. The AI/ML inference function is in the MDAF.

For NWDAF, the ML training function can be located in the MTLF of the NWDAF or the management system, the AI/ML inference function is in the AnLF.

For RAN, the ML training function and AI/ML inference function can both be located in the gNB, or the ML training function can be located in the management system and AI/ML inference function is located in the gNB.

Therefore, there might exist several location scenarios for ML training function and AI/ML inference function.

**Scenario 1:**

The ML training function and AI/ML inference function are both located in the 3GPP management system (e.g. RAN domain management function). For instance, for RAN domain-specific MDA, the ML training function and AI/ML inference functions for MDA can be located in the RAN domain-specific MDAF. As depicted in figure 4a.2-1.

 

**Figure 4a.2-1: Management for RAN domain specific MDAF**

Similarly, for CN domain-specific MDA the ML training function and AI/ML inference function can be located in CN domain-specific MDAF.

**Scenario 2:**

For RAN AI/ML capabilities the ML training function is located in the 3GPP RAN domain-specific management function while the AI/ML inference function is located in gNB. See figure 4a.2-2.

****

**Figure 4a.2-2: Management where the ML model training is located in RAN domain management function and AI/ML inference is located in gNB**

**Scenario 3:**

The ML training function and AI/ML inference function are both located in the gNB. See figure 4a.2-3.

****

**Figure 4a.2-3: Management where the ML model training and AI/ML inference are both located in gNB**

**Scenario 4:**

For NWDAF, the ML training function and AI/ML inference function are both located in the NWDAF. See figure 4a.2-4.

**Figure 4a.2-4: Management where the ML model training and AI/ML inference are both located in CN**

***End of quoted text***

##### 5.2.x.5.1 Observations and analyses: AI/ML functionalities management scenarios

* The functional arrangement scenarios defined by SA5 specifications demonstrate that different part of the ML model life cycle can be managed depending on the use case.
* The functional arrangements represent management deployment scenarios where for example ML model training related tasks can either be a domain specific or as a cooperative multi-domain task involving for example RAN and management & orchestration or CN and management & orchestration (OAM) domains.
* The LCM workflow defined by SA5 serves as a management framework to accommodate and enable all the possible functional arrangement scenarios within or cross-domains in the 3GPP system.
* The functional arrangement scenarios, coupled with the ML model LCM as defined by SA5 in TS 28.105 [x2], can be considered in the ongoing and any future 3GPP relevant specification development.

5.2.X.6 Activities Summary

Editor's note: This clause describes high-level AI/ML activities e.g. LCM for AI/ML, data collection/storage/exposure, model training/delivery/ (de)-activation/inference emulation, inference/storage/exposure, performance evaluation and accuracy monitoring. Sub-clause(s) may be added to capture details.