Position Paper: Rethinking AI/ML for Air Interface in Wireless Networks

Georgios Kontes [1] **Diomidis S. Michalopoulos** [2] **Birendra Ghimire** [1] **Christopher Mutschler** [1]

Abstract

AI/ML research has predominantly been driven by domains such as computer vision, natural language processing, and video analysis. In contrast, the application of AI/ML to wireless networks, particularly at the air interface, remains in its early stages. Although there are emerging efforts to explore this intersection, fully realizing the potential of AI/ML in wireless communications requires a deep interdisciplinary understanding of both fields. We provide an overview of AI/ML-related discussions in 3GPP standardization, highlighting key use cases, architectural considerations, and technical requirements. We outline open research challenges and opportunities where academic and industrial communities can contribute to shaping the future of AI-enabled wireless systems.

1. Introduction

In advanced 5G and future 6G systems, AI/ML will be essential to overcome challenges related to directional transmissions, rapidly varying channels, high feedback overhead and precise localization. For instance, AI/ML beam management (Xue et al., 2024) leverages historical spatio-temporal data to swiftly predict optimal directed beams from a base station (BS) towards a user equipment (UE). This eliminates exhaustive beam search procedures and ensures that both base stations and UEs can quickly adjust to dynamic channel conditions. In addition, recurrent neural networks and transformer architectures address the problem of channel state information (CSI) prediction (Jiang et al., 2025) (i.e., the inherent delays between the estimation of channel state and its usage) by forecasting future channel states, ensuring timely and accurate link adaptation even in high-mobility scenarios. Complementing these methods, deep learning-based encoder/decoder architectures learn concise latent-space representations of channel data to compress the CSI (Lin, 2025), which is periodically reported from the UE to the BS. This leads to lower latency and improved spectral and energy efficiency. Finally, AI/ML positioning (Alawieh & Kontes, 2023), enhances the reliability of location-based services, by extracting relevant-only channel features and adapting across different deployment scenarios.

Collectively, this holistic AI/ML framework has been debated and shaped within 3GPP during Releases 18 and 19 (3GPP TR38.843, 2024). Although the 3GPP standardization task force has laid the foundations for the adoption of AI/ML in the air interface, the development and deployment of both performing and cost-efficient AI/ML models come with their own challenges. As we enter the 6G era in standardization, where AI/ML is meant to play a central role, these challenges must be discussed from day one.

2. Current Discussion & Open Challenges

Any AI/ML application at scale comes at a high maintenance cost (Ashmore et al., 2021). The availability of a clean and automated pipeline that facilitates all three phases of data governance (Sec. 2.1), model training and testing, and model deployment (including model monitoring and management) (Sec. 2.2) in both a central and distributed manner (Sec. 2.3) has proven to be equally (or more) important as the model's capabilities and performance.

For wireless networks, things are even more complicated. Here, the definition and implementation of such a pipeline for AI/ML integration are also hindered by the number of different stakeholders that need to come to an agreement. The air interface facilitates a complex interplay between UE, BS and core network (NW) vendors (constrained also by the needs and requirements of the mobile network operators) that need to agree on common and possibly synergetic solutions for all stages of the AI/ML model pipeline, while avoiding any proprietary information being exposed.

2.1. Data Governance

Data collection, management (data cleaning, privacy, standardization, enhancement, availability, etc.), and security (rejection of malicious/manipulated data) for model training, testing, and monitoring are core requirements for AI/ML-

[1]Fraunhofer Institute for Integrated Circuits IIS, Fraunhofer IIS, Nürnberg, Germany [2]NOKIA, Munich, Germany. Correspondence to: Christopher Mutschler <christopher.mutschler@iis.fraunhofer.de>.

ICML 2025 Workshop on Machine Learning for Wireless Communication and Networks (ML4Wireless), Vancouver, Canada. 2025.

based solutions (Huang & Zhao, 2024). The labeled data required for model training is usually owned by the organization that trains the model. It is similar for monitoring: once the model is deployed, usually only the model owner has access to the relevant key performance indicators (KPIs).

Challenges. In current 3GPP discussions, the fundamental issue of which data can be collected in a way that respects user's security and privacy aspects plays central role. In parallel, the key questions of data ownership and access mechanisms still remain unanswered. For instance, in positioning, a UE-sided model requires area-specific labels (its true position, provided by the network) to be used for training, fine-tuning or monitoring. The situation is similar for NW-sided models. If we consider NW-sided beam management, the BS has full control on the transmitted beams but requires UE-sided beam measurements for a proper (training, testing, or monitoring) dataset.

2.2. Life Cycle Management (LCM)

2.2.1. Model Complexity vs Generalization

The energy needed for the training and inference operations of high-performing models increases with model size, as larger models exhibit better performance and generalization capabilities (Brutzkus & Globerson, 2019). Although UE and NW energy savings strategies are central within 3GPP discussions, the implications of using AI/ML models on the energy footprint remain largely underexplored. In fact, fulfilling several requirements set forth in Release 19 AI/ML for air interface would result in larger and more complex model architectures and thus increased energy consumption for both the UEs and the NW. Examples include: (i) training with larger/mixed datasets, which results in larger AI/ML models; (ii) switching between smaller (even site-specific) models, which necessitates additional signaling between the UE and the NW, and switching decision mechanism, and (iii) fine-tuning, which depends on a pre-trained (and thus potentially large) model.

Challenges. Robust generalization of UE-sided models requires tackling diverse radio environments (e.g., channel conditions and interference) that quickly degrading performance if models are not adapted. Training or fine-tuning one large or multiple specialized models places heavy energy and resource demands on UEs and requires detection of distribution shifts for unseen or rare scenarios. For smaller, specialized models, model switching or individual model update introduces additional complexity and latency as UEs and NW nodes must streamline when to switch or retrain.

2.2.2. Model Testing

One of the most important steps prior to model deployment is model testing. Different types of test can be implemented, ranging from the evaluation of prediction accuracy on a holdout data set to more elaborate testing methods that involve different versions of the model being evaluated by real users. Again, the organization that owns the model usually has all test data available and designs test protocols.

Challenges. Defining testing procedures for all use cases is not straightforward (e.g., there are no agreed testing scenarios for positioning methods that track moving UEs). There are two main constraints: (i) designing tests that are universal for the AI/ML models developed by all vendors at every UE type, and (ii) defining tests for fine-tuned or area-specific models. The former constraint requires widely applicable and simpler testing mechanisms, while the latter opens the discussion on on-device post-deployment testing protocols.

2.2.3. Model monitoring and management

Monitoring analytics are diverse: From simple (hardware) system performance metrics (e.g., memory utilization) and model input/output monitoring for detecting out-of-distribution inputs/outputs, to time-consuming and cost-inducing label-based monitoring approaches. Usually, the detection of model performance drop triggers a new data collection and model retraining (or fine-tuning) procedure.

Challenges. As also mentioned above, monitoring label availability (for label-based monitoring analytics) requires alignment between UE and NW sides, which blurs the ownership status of the monitoring dataset. Second, even though for UE-(BS-)sided models model monitoring and management are handled by the UE (BS), the NW is still responsible for the performance in an area. This implies that the NW can request the UE (or BS) to start a monitoring session on their respective models to determine if performance is adequate or a fallback to a different approach is required. What further complicates things is that reporting exact monitoring KPIs might result in security issues: If a malicious/adversarial UE reports wrong KPIs to the NW, the NW cannot easily detect this, while if the UE reports its model predictions to the NW, there is a danger of model extraction (Tramèr et al., 2016; Oliynyk et al., 2023).

2.2.4. Performance KPIs

Performance KPIs for AI/ML model training, testing, and monitoring consider the perfect predictive performance of the model in all use cases: beam management strives to identify the beam with the highest RSRP among top predicted beams, or predicting the correct RSRP for all beams; CSI prediction is tasked to forecast future channel conditions perfectly; CSI compression aims at reconstructing CSI flawlessly at the BS; and positioning is expected to estimate UE positions at millimeter accuracy.

Challenges. Such stringent prediction-accuracy-oriented

Table 1. Promising research directions to unlock the adoption of AI/ML in air interface within 6G.

Research Direction	Challenges						
	Data Governance	Model Life Cycle Management					Interoperability
		Monitoring & Management	Model Complexity	Generalization	KPI alignment	Simplified Testing	
Multi-task learning	✓	✓		✓			✓
Conditional architectures		✓	✓				✓
Root Cause analysis	✓	✓				✓	✓
Opportunistic data collection	✓	✓	✓			✓	
Self-supervised/ Meta/Active learning	✓	✓		✓			
Reinforcement Learning & AI-based Optimization	✓		✓	✓	✓		✓

KPIs are not well aligned (see Russell & Norvig (2020) for a definition of alignment) to the real world applications. For example, video streaming requires different beam quality than text communication, many positioning tasks do not require millimeter or centimeter accuracy, and for some tasks incomplete CSI information is enough (Jiang et al., 2025). The challenge here, in addition to identifying the right KPI for each task, is to train and deploy models that strike the right balance between complexity and performance requirements, as posed by the underlying application.

2.3. Model Interoperability

Many scenarios require "two-sided" models that are deployed in a distributed fashion between UE and BS, which involve a joint operation of the models at both ends. In addition to low complexity, AI/ML models should strive for a robust exchange of information between the entities, enabling a coordinated operation of the model as a whole.

Challenges. The AI/ML model at UE- and NW-side may come from different providers, facing the challenge of interoperability. Vendors at each side need to support the model yet without disclosing their respective proprietary models. This implies that UE and NW vendors must architect sophisticated interoperability frameworks that facilitate selective sharing of their models. This would ensure the preservation of complete model confidentiality while fostering the harmonious and successful operation of two-sided models.

3. Future Research Directions

We discuss promising research directions that support true integration of AI/ML in wireless air interfaces (Tab. 1).

3.1. Multi-Task Learning

AI/ML use cases are typically developed isolated to one another, which results in high development cost (Jiang et al., 2025). Possibly overlapping data is collected separately and thus trained models encode redundant information. To address this, multi-task learning (Zhang & Yang, 2021) deploys a single backbone model as a common representation of the channel (e.g., see Ott et al. (2024)), which in sequence provides targeted "heads" tailored to the involved use cases.

This also facilitates model monitoring and management, as maintenance of a single multi-task model is more straightforward than tracking multiple specialized models, while retraining or fine-tuning can be done within a unified pipeline. As task representations share latent features, fine-tuning (or even adding a new use case / model head) can require fewer data, since such shared knowledge can be effectively utilized. Furthermore, if complementary tasks on the UE side (e.g., beam management and resource scheduling, respectively) partially share the backbone model, the integration of two-sided training or inference routines in multi-task models is facilitated as well, thus enabling model interoperability.

3.2. Conditional Neural Architectures

Even with multi-task approaches, universal "one-size-fits-all" models may still be suboptimal for a number of UEs. Some UEs frequently encounter a narrow set of channel conditions (e.g., repeated commuting routes), therefore, on-device personalized sub-models can yield substantial gains, like increased performance and reduced signaling overhead.

A way to tackle this could be the adoption of two-sided models, enhanced with early exit architectures (Teerapittayanon et al., 2016). This would allow earlier termination when prediction confidence is high, exploiting the fact that many data samples can be classified using only the earlier layers of the model. For challenging conditions (e.g., a shift from LOS to multipath NLOS), the inference can continue through additional layers, ensuring uninterrupted performance.

Some of the neural network layers can be shared between the UE and NW vendors, allowing flexible execution that

eases the UE-side computational burden and ensures a high level of interoperability between the distributed sub-models, while protecting the proprietary properties of the UE- and NW-side parts of the model. This way, (distributed) model monitoring and management is also more principled, as performance degradation can be isolated to specific exits, instead of re-verifying the entire network.

3.3. Root Cause Analysis

When an AI / ML model performs poorly, the monitoring entity should not only report a failure but also pinpoint the source of the problem. For instance, as studied within 3GPP Release 19, if the performance of a two-sided (encoder/decoder) CSI compression model degrades, an investigation on whether the encoder, the decoder, or a change in the radio environment is responsible enhances the interoperability properties of such distributed models.

Such information also allows making informative decisions on how to mitigate the effect of a detected issue, e.g., retrain only part of the model, switch to a smaller fallback model, or temporarily deactivate the AI/ML functionality. This enables more transparent model testing procedures and can even support high-level root-cause explanations, e.g., using natural language (Roy et al., 2024; Manjunath et al., 2025).

3.4. Opportunistic Data Collection

Instead of continuously collecting positioning labels or logging CSI and beam measurements, UEs and base stations can report data only when certain triggers appear. For instance, the UE or the NW might request monitoring data only if the UE's performance abruptly changes – a sign that the current environment differs from what the model has seen before. Likewise, for obtaining the true labels when legacy positioning methods are unreliable, the NW can opportunistically activate external sensors or measurements to acquire secondary information to obtain labels in challenging scenarios (e.g., using camera or other UE sensors).

Such targeted, trigger-based data logging (for training, testing and monitoring) not only enables a more agile model monitoring and management framework, but also unlocks the ability for on-demand testing in highly variable scenarios, thus allowing effective use of test and label resources instead of continuous, exhaustive logging.

3.5. Reinforcement Learning / Optimization-Based AI

AI/ML studies within 3GPP standardization prioritize improving a single metric (usually prediction accuracy), often without considering the broader objectives tied to system-wide KPIs, i.e., without involving Quality of Service (QoS) and Quality of Experience (QoE) requirements. For example, simply increasing channel prediction accuracy does not necessarily translate into the best end-to-end performance for tasks such as positioning, beam management, or channel state feedback (Jiang et al., 2025).

To address this gap, future AI/ML designs should include a family of online learning or optimization-based approaches, such as reinforcement learning (Sutton & Barto, 2018), recommender systems (Zhang et al., 2019) or mixed-integer optimization (Wolsey & Nemhauser, 1999). These adaptive algorithms, unlike fixed models solely optimized for higher prediction accuracy, can be combined with task-oriented KPIs and are able to continuously adjust to real-world conditions, such as network load fluctuations, device mobility, or localized interference patterns.

Unsurprisingly, such methods also ease the burden on data collection for model (called policy within RL) training, testing, and monitoring, since stringent requirements on high-quality labeled datasets no longer exist. Instead, task-related measurements used to calculate reward (and possibly a set of constraints) estimates must be collected. This also enhances UE- and NW-side interoperability, since each side can maintain its own policy function and still coordinate via reward exchanges or partial state observations.

3.6. Efficient use of labeling resources

Most of the widely-adopted AI/ML approaches and model architectures can benefit from algorithms that enable efficient use of labeling resources, such as self-supervised (Liu et al., 2021), meta-learning (Hospedales et al., 2021) and active learning (Ren et al., 2021) techniques. For example, in several scenarios, a meta-learned model can quickly adapt when deployed to new cell sites or applied under changed channel conditions. In CSI prediction/compression use cases, self-supervised methods can process channel measurements to predict future channel behavior or compress CSI data, without any dependency on manual annotation. Or in positioning, methods such as channel charting (Studer et al., 2018) can alleviate the dependency on positioning reference units for the availability of accurate position labels.

Complementary, active learning enables on-demand model monitoring, testing, and fine-tuning, ensuring that only the most informative data samples (e.g., when model prediction uncertainty is high or it starts to exhibit degraded performance) are collected and are used for model adjustments.

4. Conclusion

There is no single silver bullet to address the aforementioned challenges. Future research should focus on hybrid solutions that combine modular, adaptive algorithms with robust data governance and continuous model monitoring, aiming to reduce the reliance on hard-to-acquire and cost-deficient label-based approaches.

References

3GPP TR38.843. Technical Report Group Radio Access Network; Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface (V18.0.0). Technical report, The 3rd Generation Partnership Project, Sophia Antipolis, France, 2024.

Alawieh, M. and Kontes, G. 5G positioning advancements with AI/ML. *arXiv preprint arXiv:2401.02427*, 2023.

Ashmore, R., Calinescu, R., and Paterson, C. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. *ACM Computing Surveys (CSUR)*, 54(5): 1–39, 2021.

Brutzkus, A. and Globerson, A. Why do larger models generalize better? a theoretical perspective via the xor problem. In *International conference on machine learning*, pp. 822–830. PMLR, 2019.

Hospedales, T., Antoniou, A., Micaelli, P., and Storkey, A. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169, 2021.

Huang, Q. and Zhao, T. Data collection and labeling techniques for machine learning. *arXiv preprint arXiv:2407.12793*, 2024.

Jiang, C., Guo, J., Li, X., Jin, S., and Zhang, J. AI for CSI prediction in 5G-advanced and beyond. *arXiv preprint arXiv:2504.12571*, 2025.

Lin, X. The bridge toward 6G: 5G-advanced evolution in 3GPP release 19. *IEEE Communications Standards Magazine*, 9(1):28–35, 2025.

Liu, X., Zhang, F., Hou, Z., Mian, L., Wang, Z., Zhang, J., and Tang, J. Self-supervised learning: Generative or contrastive. *IEEE transactions on knowledge and data engineering*, 35(1):857–876, 2021.

Manjunath, H., Heublein, L., Feigl, T., and Ott, F. Multimodal-to-text prompt engineering in large language models using feature embeddings for GNSS interference characterization. *arXiv preprint arXiv:2501.05079*, 2025.

Oliynyk, D., Mayer, R., and Rauber, A. I know what you trained last summer: A survey on stealing machine learning models and defences. *ACM Computing Surveys*, 55 (14s):1–41, 2023.

Ott, J., Pirkl, J., Stahlke, M., Feigl, T., and Mutschler, C. Radio foundation models: Pre-training transformers for 5g-based indoor localization. In *2024 14th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, pp. 1–6. IEEE, 2024.

Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

Roy, D., Zhang, X., Bhave, R., Bansal, C., Las-Casas, P., Fonseca, R., and Rajmohan, S. Exploring LLM-based agents for root cause analysis. In *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, pp. 208–219, 2024.

Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020. ISBN 9780134610993.

Studer, C., Medjkouh, S., Gonultaş, E., Goldstein, T., and Tirkkonen, O. Channel charting: Locating users within the radio environment using channel state information. *IEEE Access*, 6:47682–47698, 2018.

Sutton, R. and Barto, A. *Reinforcement Learning: An Introduction*. Cambridge, MA, MIT Press, 2018.

Teerapittayanon, S., McDanel, B., and Kung, H.-T. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd international conference on pattern recognition (ICPR)*, pp. 2464–2469. IEEE, 2016.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction APIs. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Wolsey, L. A. and Nemhauser, G. L. *Integer and combinatorial optimization*. John Wiley & Sons, 1999.

Xue, Q., Guo, J., Zhou, B., Xu, Y., Li, Z., and Ma, S. AI/ML for beam management in 5G-advanced: a standardization perspective. *IEEE Vehicular Technology Magazine*, 2024.

Zhang, S., Yao, L., Sun, A., and Tay, Y. Deep learning based recommender system: A survey and new perspectives. *ACM computing surveys (CSUR)*, 52(1):1–38, 2019.

Zhang, Y. and Yang, Q. A survey on multi-task learning. *IEEE transactions on knowledge and data engineering*, 34(12):5586–5609, 2021.