Framework for Trustworthy AI/ML in B5G/6G

Sokratis Barmpounakis[1], Panagiotis Demestichas[1,2]
[1]WINGS ICT Solutions, Athens, Greece; [2]University of Piraeus, Piraeus, Greece

2022 1st International Conference on 6G Networking (6GNet) | 978-1-6654-6763-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/6GNet54646.2022.9830408

***Abstract*— The world already moves towards the 6G era. AI/ML mechanisms will become structural components of the system and operate in a native manner. As the systems get more complex and intelligent, mechanisms for ensuring trust in those operations become critical. In this paper we take some first steps in the discussion on trustworthy AI towards 6G. Our discussion justifies the need for a framework and highlights that a comprehensive approach needs to be taken, for protecting the input of the AI mechanisms, for achieving an explainable operation, and for guaranteeing proper outputs. Architectural design principles, as well as resource allocation aspects are touched upon, and the important future steps are designated.**

Keywords*— *Wireless communications, Artificial Intelligence, Machine Learning, Explainability, Trust

I. INTRODUCTION

Wireless and mobile communications are entering an exciting era, characterized by their increasing on the transformation of our economy and society. Currently, the set of main drivers includes the "Fifth Generation" (5G) of wireless communications, and those to follow "Beyond the 5th Generation" (B5G), including 6G (6th Generation of wireless communications) (e.g., [1]-[4]). Some of the most prominent examples of disruptive technologies that B5G/6G networks will brace are massive Internet of Things (IoT), massive digital twinning, Augmented/Virtual Reality (AR/VR), holographic telepresence, nano-networking, and quantum computing. Towards addressing the ultra-challenging requirements that such technologies pose in terms of capacity, performance, availability and reliability, resilience, scalability, and sustainability -including energy efficiency-, B5G/6G need to reach utmost efficiency, which will be enabled by self-awareness and highly intelligent decision-making, in terms of resource allocation, management and orchestration, as well as security. All layers and planes of the protocol stack will be affected, from the physical to the services and applications, from user to control and management. This thirst, for higher orders of intelligence, reinstated and magnified the interest in Artificial Intelligence (AI) ([5]-[7]), in general, and ML (Machine Learning), in particular -as a key enabler-.

AI/ML is already considered a structural component of the B5G/6G system. 3GPP's Network Data Analytics Function (NWDAF), as well as the Management Data Analytics Function (MDAF) ([8][9]) are considered the first steps towards this direction -since Release 16-, enabling Network Functions (NFs) to access the MNO-driven analytics for different purposes, including intelligent slice selection and control. In more recent 3GPP working items [10], the aspects to be considered in relation to the potential extension of the current NWDAF functionalities towards distributed operations are described, such as whether NWDAF functional split is required, as well as which NWDAF functionality can be separated or placed in a different NF/NF Service.

Besides 3GPP, ITU also specifies an architectural framework for ML in future networks [11], which presents requirements and specific architectural components; these components include, but are not limited to, a ML pipeline as well as ML management and orchestration functionalities. The integration of such components into future networks including IMT-2020 [12] and guidelines for applying this architectural framework in a variety of technology-specific underlying networks are also described.

The above efforts and standardization trends make the forthcoming holistic adoption of AI/ML in B5G/6G a matter of time. The key challenge, nevertheless, is to move beyond *integration*, towards a true *unification* of AI/ML in future networks, in a network-native manner. In order to ensure that AI/ML components can actually serve as native components of the system, and the more intelligent the various management, orchestration, security -and decision-making as a whole- components become, the more important the issue of "trustworthy AI" [13] gets. From a general perspective trust can be defined as the "firm belief in the reliability, truth, or ability of someone or something" [14]. This is an area that gains more and more attention, as, so far, the main bulk of the "AI for wireless" work was on validating the novel AI-based mechanisms. Traditionally, trust has been associated with security. However, due to the criticality of the AI-powered wireless infrastructures and the impact that the respective decision-making can have on critical network services and use cases (often influencing even human lives, such as Connected and Automated Mobility), our work advocates that a wider approach needs to be taken; trust will derive from the evaluation of functional (i.e., related to the system behavior) and non-functional (primarily, Key Performance Indicator (KPI) accomplishment, respective costs, etc.) criteria. Besides,

regulation measures will be urgently needed, as already being introduced by the European Union's AI Act and legal framework for AI [15], as well as NIST's AI Risk Management Framework for AI in the US [16]. This paper aims at contributing to (the initiation of) a discussion that will lead to systematic and holistic approaches for encompassing trustworthy AI mechanisms in B5G/6G architectures.

The rest of the paper is structured as follows. Section II presents existing efforts on defining and tackling trustworthiness. Section III elaborates on the AI/ML perspective, the various AI/ML algorithms and related challenges related to their complexity and operations, while it also highlights the key requirements that place AI/ML trust, as a critical component of the overall system. Section IV presents the proposed trustworthy AI/ML framework. Section V includes elements for future study and concluding remarks.

II. Trustworthiness in Existing Works

Ziegler et al. [17] present a comprehensive set of technology enablers for security and trustworthiness, required for communication systems for the 6G era. Trustworthiness, according to the authors, must be ensured across IoT, heterogeneous cloud and networks, devices, sub-networks, and applications. To this end, they describe the importance of AI/ML but primarily as an enabling technology towards secure and privacy-preserving networks.

Li et al. [18] outline the concept of trustworthy autonomy for 6G, including elements such as how Explainable AI (XAI) can generate the qualitative and quantitative modalities of trust, while they review and present the relevant concepts of trust for 6G Radio Resource Management (RRM) automation; XAI test protocols for integration with RRM and associated KPIs for trust are presented. The authors attempt to quantify trust for 6G systems, focusing on the explainability of the ML model and decomposing the modeling of the trust into a linear combination of physical and emotional trust.

A recent whitepaper released by Ericsson [19], explores approaches to the steps needed to answer fundamental questions around trustworthy systems design. This work also tackles trustworthiness mainly from the security perspective, while AI is described as an enabler for automation-driven assurance and defense. AI@EDGE EU project [20] leverages the concept of reusable, secure, and trustworthy AI for network and service automation in industry relevant multi-stakeholder environments. 5GZORRO EU project [21] focuses on trustworthiness from several perspectives, such as resource discovery, spectrum management, edge computing, 3rd party resources and services, as well as marketplace operations.

Although, recent works already do discuss the concept of trustworthiness, AI/ML is so far mainly seen as an enabler, while trustworthy AI/ML has only been studied from the explainability perspective. This work focuses on the trustworthiness of AI/ML-driven decision-making and attempts to decompose the various requirements, as well as formulate the primary design considerations for trustworthy AI/ML operations towards B5G/6G.

III. The AI/ML Perspective

The goal of AI is to create systems that can perform tasks that would otherwise require human intervention. Those tasks rely on specific techniques, among which, machine learning and deep learning are the most popular AI subfields which are widely adopted in wireless networks. ML is a big subset -or primary enabler- of what falls in the AI world. ML develops systems that improve upon themselves, while it may apply classification, regression, clustering, data generation, etc. Also, ML may be implementing Supervised, Unsupervised, Semi-supervised, or Reinforcement Learning approaches, while the complexity of the different algorithms may vary considerably (Fig. 1). A critical challenge results from the fact that the higher the complexity, the higher the performance of the respective AI/ML mechanism usually is; this is a crucial trade-off that needs to be carefully approached. Specifically Deep Learning [22], and related techniques have proven, beyond doubt, critical enablers for the efficiency optimization and management of wireless systems, infrastructures, and services [23].

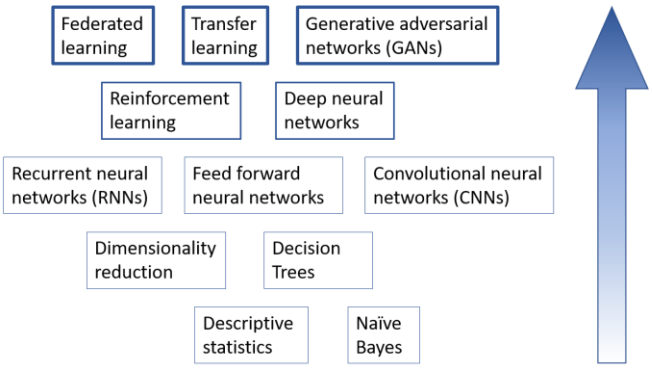

Fig. 1 Commonly used ML methods from simple ones to the cutting-edge. The interest of B5G/6G towards adoption follows a similar trend.

Novel architectures and algorithms featuring Federated Learning principles (FL) are also more and more gaining ground [24], due to certain needs and trends identified in wireless communications, namely a) the need towards online/real-time learning; this is important for maximizing the agility (reactive/proactive adaptations), powered by insights and predictions at

smaller time scales, particularly for complex wireless environments; b) the need for placing intelligence towards the (far/extreme) edge of the network (including end devices). This is crucial for various reasons: a) for minimizing the delays associated with certain delay-critical applications, b) for scalability, i.e., for minimizing the consumed resources and overhead on the network links, required for exchanging data for training/inference, and c) for preserving privacy, as data is maintained locally.

IV. Trustworthy AI Requirements, Design Principles and Proposed Enablers

A. Rationale of the framework

The integration of AI/ML in B5G/6G networks, along with the requirements for supporting online/real-time learning makes the need for trusting the AI/ML components critical. Also considering that (i) the available time to verify the network decisions is limited; and (ii) the exchanged data over wireless links might be more exposed to malicious attacks, makes this need even more important. Mechanisms are thus needed to mitigate such risks and the associated huge economic and societal impact they would entail, therefore contributing to the perspectives of economic/societal sustainability. In addition to the susceptibility to malicious attacks, the AI/ML/FL-based decision making itself is of high focus; the explainability and auditability of the decision-making operations and outputs can radically influence the impact that those decisions have on the network, -and as a result- the extent, to which human intervene towards fine-tuning/preventing the application of AI/ML/FL outputs to network operations. Computational cost -and as a result energy consumption considerations- are also crucial, as sustainability can increase trustworthiness and vice versa.

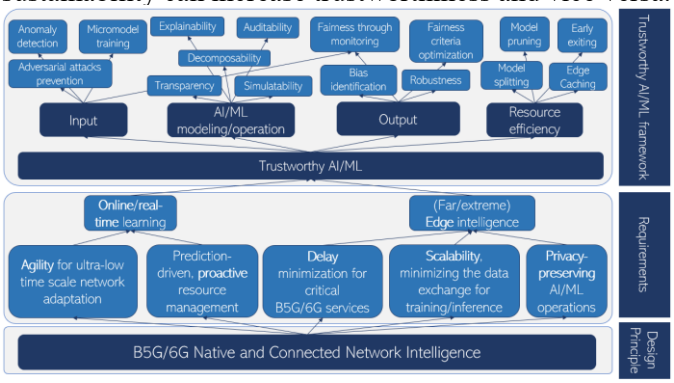

Fig. 2 Design principle, requirements and enablers of the proposed trustworthy AI/ML framework

In this paper, we interpret trust as the aggregation of answers to the following questions: What is the "quality" of the input data? What is the quality of the output? Is there bias in the respective decisions? Can the operation of the algorithm be explained/audited? How do we ensure resource efficiency of these models, operating in diverse network segments and nodes, spanning across the whole computing continuum? The above is further discussed in detail in the following sub-sections. Fig. 2 provides an overview of the trustworthy AI/ML framework, linking the AI/ML design principle with respective requirements, and ultimately to the specific enablers and methods that are proposed in this paper.

The envisioned framework identifies diverse operations, which can be applied either in a distributed or a centralized manner. In terms of the network entities, in which the trustworthy AI/ML operations can take place, different options are identified; as an example, in Fig. 3, trustworthy AI/ML operations comprising input data probing or output data processing are illustrated as part of a distributed unit (DU) operation -in line with the O-RAN specifications [25], while operations such as mechanisms for consolidating FL models, or DU entities management (and respective AI/ML components) are part of the Centralized Unit (CU) / cloud node.

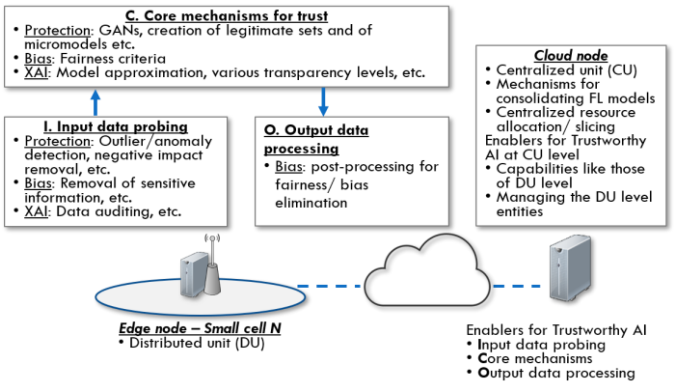

Fig. 3 Enablers for Trustworthy AI/ML/FL models in B5G/6G: (i) Enablers at DU level; (ii) Enablers at the CU level, with capabilities as those of DU level, plus the ability to manage the DU level entities

In the following, we overview the most critical aspects related to the trustworthiness of data-driven approaches for B5G/6G networks, namely the input of the AI/ML entity/component, the functionality aspects -highlighting principles such as explainability and auditability-, the quality and the impact of the output, as well as aspects related to the efficiency of the resources allocated for the execution of the respective AI/ML components/entities.

B. Input

As mentioned previously, the approach for AI in B5G/6G networks is heading towards a decentralized learning process. Methods for the collaborative training of DNN models have been proposed. [26] suggests a first phase of local model training and then, a second phase for the finalization of the training through the global aggregation of the updated parameters. Google's TensorFlow Federated [27] is an open-source framework for experimenting with machine learning and other computations on decentralized data. Nevertheless, it

largely remains an open problem that certain generative ML models can craft systematic (evasion and poisoning) attacks against machine learning classifiers.

A process, by which an adversary entity injects malicious data in order to influence the training process and degrade/revert the algorithm's performance/outputs is data poisoning. A sample adversary that can create a maliciously trained network is presented in [28]. Moreover, as described in [29], there can be deep generative models that learn via the maximum likelihood principle, by constructing an explicit probability density distribution or by providing some way of "indirect interaction" with this probability distribution. In the FL case, the effect of poisoning can spread to various areas of the network, in which the model is distributed. The mitigation of this effect is largely an open problem, nevertheless, there are specific directions for future work.

To prevent such adversarial attacks, the use of Generative Adversarial Nets (GANs) is considered as one of the cutting-edge approaches [30]. According to the GAN paradigm, two major approaches have been applied in these cybersecurity studies: (i) the GAN is used to improve generalization to unforeseen adversarial attacks, by generating novel samples that resembles adversarial data and can then serve as training data; (ii) the GAN is trained on data with the goal of generating realistic adversarial data that can thus fool a security system.

In parallel, it must be noted that blockchain has been gaining support in the last years with many applications. Lately, some approaches that combine FL with blockchain have been gaining ground. This way privacy-aware ML models at the core of an FL ecosystem can enable the entire network to learn from data in a decentralized manner [31]. Lastly, a permissioned blockchain-based federated learning method is described, where incremental updates to an anomaly detection machine learning model are chained together on the distributed ledger for intrusion detection [32]. By integrating FL with blockchain technology, the solution supports the auditing of ML models without the necessity to centralize the training data.

B5G/6G systems, due to the native integration of AI/ML-driven intelligence, will need advanced mechanisms that are able to spot system vulnerabilities, especially in a distributed and real-time operation context. In our envisioned approach, data sanitization could rely on techniques such as anomaly detection, as well as on more complex methods, such as elimination of training data, which are proven to have substantial negative impact on the ML task accuracy/performance [33]. Training with micromodels in order to reduce the risk of an attack is also one of the potential solutions. Leveraging also on the GAN paradigm, realistic samples that resemble adversarial data can be generated and be used in the training, for enabling FL models to obtain a sense of "immunity" to the poisonous attack.

C. Functionality and operation

Another issue that ensures trust to the AI/ML powered algorithms conducting various network management, orchestration, resource allocation, or security operations, is to have the ability to explain or interpret their behavior and outputs. B5G/6G will be providing a plethora of services, ranging from those with stringent requirements (self-driving cars, robots, health-related) to less critical ones, which nevertheless, will still need to be served at a certain level of quality. Apparently, as the consequences of resource allocation algorithms will be most significant, the need for explainability (interpretability) becomes of primary importance. Manual inspection of representative samples, outliers, and misclassifications of ML models will radically increase the models' explainability/interpretability, which is a vital feature for their application and evolution.

Enforcing explainability in the constructed models and decisions taken is gaining more and more attention [34][35]. Departing from traditional black-box models, where no insight is given to the decisions that models make, explainable AI/ML approaches allow to know the reasons behind (e.g., certain users were refused to get access to the network due to a partial malfunctioning of a base station, along with a sudden traffic surge). Knowing the grounds for such decisions can be crucial when, e.g., those decisions have legal/contractual implications (e.g., violation of service level agreements). A method to balance the model interpretability and data privacy in FL is defined in [36]; the model uses Shapley values to reveal detailed feature importance for host features and a unified importance value for federated guest features. On the other hand, [37] suggests using differentially private generative models (including GANs), trained with federated learning, in place of data inspection and demonstrates that they can be used effectively to debug many commonly occurring data issues even when the data cannot be directly inspected.

Enablers of this area need to offer visibility and transparency in the AI/ML/FL-powered resource-allocation algorithms' operation. Such enablers can be based on model approximation approaches (construction of simpler models that can yield explanations, typically, with a lower accuracy), data auditability (scrutiny of the data that was used for basing the model decisions). Emphasis should be placed to the striking of a balance between oversimplification (of the algorithms applied in the system) and complexity replication, as well as on supporting different levels of transparency. In this respect, different levels of transparency should be considered, indicatively algorithmic transparency, decomposability, and simulatability [34]. The outcome of

investigating the different levels will be an optimal solution to the trade-off between model interpretability and performance. Inherently interpretable models (e.g., General additive and Bayesian models) can serve as a baseline, in contrast to post hoc external XAI techniques like model approximation or data auditing. The foreseen outcome would be a set of explainability techniques for deep learning models, such as model simplification, feature relevance and local explanations.

D. Output

Even under proper (non-poisoned) datasets, AI/ML decisions can be biased, for example, be unfairly favoring a certain user, service, or traffic class, through the respectvie resource allocation decision-making; another scenario is using certain resources in a static manner, therefore, potentially "suffocating" neighboring areas and/or making things hard for reaching a global optimum. Bias may exist due to the input data, or because of the way algorithms are trained and structured. Federated learning is susceptible to newly introduced sources of bias, due to the lack of fairness across devices, as well as system and statistical heterogeneity in the network. When FL is deployed on the network edge, the number of training samples on each device might be disproportionate, when deciding which clients to sample based on connection type and quality, device type, location, activity patterns, and local dataset size criteria.

The detection of such imbalances in the training data via data-driven (i.e., ML approaches) is of course very meaningful. General directions for addressing "unfair" bias are outlined in [38], while [39][40] focus on more specific methods regarding the processing of algorithm input and output. In order to ensure bias elimination and fairness, the authors of [41] propose an optimization objective inspired by fair resource allocation in wireless networks; the algorithm encourages a more uniform accuracy distribution across devices. From a more general point of view, a FL framework -where the centralized model is optimized for an arbitrary target distribution formed by a mixture of the client distributions- is presented in [42]. Besides, in [43] the authors investigate the robustness of FL approaches to bias in the training data resulting from the heterogeneity of end devices (resulting in e.g., different sampling rates or sampling quality) in edge networks.

The mitigation approach should be comprehensive, in line with directions appearing in the literature [39][40], ranging from output to input data processing. For instance, the outputs of AI/ML/FL models can be closely monitored and post-processed, prior to being applied in the network, to satisfy fairness constraints. Moreover, gradually, there can be means that focus on the pre-processing of the inputs, to remove (change the representation / importance of) sensitive data that lead to unfair decisions. B5G/6G should employ enablers that will be capable of identifying and mitigating bias in resource allocation decisions, by considering various fairness criteria.

Two layers of bias mitigation techniques are foreseen, namely one on the input and one on the output of the resource allocation layer. Firstly, balancing of different input streams could eliminate impact discrepancies that result from heterogeneity of the sources and data points. The second layer could focus on ensuring that fairness criteria are optimized, prior to the application of decisions on the system. The result would be the capability of identifying bias in the resource allocation algorithms' output and mitigating it, by modifying both the input and output, via data sanitization and extensive verification.

E. Ensuring resource efficiency

Another critical aspect is the efficiency -in terms of required and consumed resources-, with which, the AI/ML components are deployed/operating within a specific network segment. Regarding the deployment of distributed or federated learning on edge devices of 5G and beyond networks, main building blocks, as well as different neural network architectural splits and their inherent tradeoffs are investigated in [44]. Aiming to minimize the computation, communication, and storage costs, an approach to resource management via integrating model pruning both at the server and in a distributed manner is presented in [45]. Moreover, an optimization of the mobile edge computing, caching and communication is proposed by the authors of [46], by bringing more intelligence to edge systems.

The best allocation of the diverse enablers based on performance and resource consumption factors needs to be pursued in B5G/6G networks. The cost efficiency of deploying all pertinent enablers needs to be examined. Such study should include a joint evaluation of the speed of convergence to decisions, accuracy, resource efficiency in terms of energy consumption and more. Another important aspect is the size of the network of the federated learners. All these aspects, deriving from the real-time/online application/learning of the models need to be addressed proposing an optimal deployment methodology.

V. Future Work and Concluding Remarks

AI is considered a driving enabler for the evolution of future communication systems towards an intelligent, self-aware ecosystem with decentralized capabilities exploiting resources across the computing continuum, from cloud down to the extreme edge, including end devices. In order to for this enabler to flourish, the decision-making and impact that AI/ML has on critical

aspects of the network management operations must be characterized by robustness, transparency, reliability and ultimately trustworthiness. Managing trust is thus a critical issue for ubiquitous AI. Current ongoing activities approaching AI from an ethical and a normative perspective are also expected to help bridge the gap between AI and network management.

This paper discussed the facets of trust in AI/ML mechanisms for B5G/6G networks. Arising issues need to be addressed, to ensure that the incorporation of such enablers in the network will not or undermine its capacities or impede its global acceptability. Essential next steps for developing the framework for trustworthy AI for B5G/6G comprise elaboration on functional problem definitions, fine-tuning and realization of proposed solutions, evaluation of alternatives and validation with respect to their effectiveness with diverse datasets and through different pilots. The issue of the system-wide design and of tentative standardization needs to be addressed as well. In this context, schemes for the flexible allocation of functions to the physical elements will need to be adopted.

ACKNOWLEDGMENTS

This work has been partially supported by EC H2020 DAEMON project (Grant agreement No. 101017109).

REFERENCES

[1] Hexa-X, Horizon 2020, https://hexa-x.eu/ (accessed 06/2022).
[2] Daemon, Horizon 2020, https://h2020daemon.eu/ (accessed 06/2022).
[3] Ian F. Akyildiz, Ahan Kak, Shuai Nie, "6G and Beyond: The Future of Wireless Communications Systems", IEEE Access, Vol. 8, pp. 133995-134030, July 2020, DOI: 10.1109/ACCESS.2020.3010896.
[4] Chris Kelly, "6G: A $1 trillion opportunity for telcos", ITP corporation, February 2021.
[5] Internet Society, "Artificial Intelligence and Machine Learning: Policy Paper", 2017.
[6] "AI and Machine Learning in 5G", ITU News Magazine, No. 5, 2020,
[7] P. Demestichas, et al., "Beyond 5G and Artificial Intelligence", panel discussion, EUCNC 2019, Valencia, Spain.
[8] 3GPP, TS 23.791, V16.2.0, "Study of Enablers for Network Automation for 5G (Release 16).
[9] 3GPP, TS 28.533, V15.2.0, "Management and orchestration; Architecture framework (Release 15).
[10] 3GPP, TS 23.700, V17.0.0, "Study on enablers for network automation for the 5G System (Release 17).
[11] ITU, Y.3172, Architectural framework for machine learning in future networks including IMT-2020.
[12] ITU, ITU-R M.2083-0, IMT Vision – Framework and overall objectives of the future development of IMT for 2020 and beyond.
[13] High-Level Expert Group on Artificial Intelligence setup by the EC, "Ethics recommendations for trustworthy AI", April 2019.
[14] Merrian-webster dictionary Web site, https://www.merriam-webster.com/dictionary/trust (accessed 06/2022).
[15] European Union, The Artificial Intelligence Act, https://artificialintelligenceact.eu/ (accessed 06/2022).
[16] NIST, AI Risk Management Framework, https://www.nist.gov/itl/ai-risk-management-framework (accessed 06/2022).
[17] V. Ziegler, et al., "Security and Trust in the 6G Era," in IEEE Access, vol. 9, pp. 142314-142327, 2021, doi: 10.1109/ACCESS.2021.3120143.
[18] C. Li, et al., "Trustworthy Deep Learning in 6G-Enabled Mass Autonomy: From Concept to Quality-of-Trust Key Performance Indicators," in IEEE Vehicular Technology Magazine, vol. 15, no. 4, pp. 112-121, Dec. 2020, doi: 10.1109/MVT.2020.3017181.
[19] Ericsson, "Building trustworthiness into future mobile networks", url: https://www.ericsson.com/en/reports-and-papers/white-papers/building-trustworthiness-into-future-mobile-networks (accessed 06/2022).
[20] AI@EDGE H2020 project, Deliverable D2.1, "Use cases, requirements, and preliminary system architecture".
[21] 5GZORRO H2020 project, Deliverable D2.2, "Design of the 5GZORRO Platform for Security and Trust".
[22] I.H. Sarker, "Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions". SN COMPUT. SCI. 2, 420 (2021).
[23] N. C. Luong et al., "Applications of Deep Reinforcement Learning in Communications and Networking: A Survey," in IEEE Communications Surveys & Tutorials, vol. 21, no. 4, pp. 3133-3174, Fourthquarter 2019, doi: 10.1109/COMST.2019.2916583.
[24] S. Niknam, et al., "Federated Learning for Wireless Communications: Motivation, Opportunities, and Challenges," in IEEE Communications Magazine, vol. 58, no. 6, pp. 46-51, June 2020, doi: 10.1109/MCOM.001.1900461.
[25] O-RAN Alliance, O-RAN specifications, url: https://www.o-ran.org/specifications (accessed June 2022).
[26] B. Lim, et al., "Federated Learning in Mobile Edge Networks: A Comprehensive Survey", doi:https://doi.org/10.48550/arXiv.1909.11875
[27] TensorFlow Federated, https://www.tensorflow.org/federated (accessed 06/2022).
[28] T. Gu, et al., "BadNets: Evaluating Backdooring Attacks on Deep Neural Networks," IEEE Access, vol. 7, pp. 47230-47244, 2019.
[29] Ian Goodfellow (2017) NIPS 2016 tutorial: generative adversarial networks. arXiv:1701.00160.
[30] C. Yinka-Banjo, O. Ugot. A review of generative adversarial networks and its application in cybersecurity. Artif Intell Rev 53, 1721–1736 (2020). https://doi.org/10.1007/s10462-019-09717-4.
[31] A. Nagar (2019) Privacy-Preserving Blockchain Based Federated Learning with Differential Data Sharing, arXiv:1912.04859.
[32] D. Preuveneers, et al., "Chained Anomaly Detection Models for Federated Learning: An Intrusion Detection Case Study", Appl. Sci, 8(12), 2663 .
[33] M. Barreno, et al., "The security of machine learning", Machine Learning, Vol. 81, 2010, pp. 121–148.
[34] Z. Lipton, "The mythos of model interpretability", Commun. of the ACM, Vol. 61, No. 10, pp. 36-43, 2018.
[35] Stephan Raijmakers, "Artificial Intelligence for law enforcement: challenges and opportunities", IEEE Security and Privacy, Vol. 17, No. 5, pp. 74-77, September/October 2019.
[36] G. Wang, "Interpret Federated Learning with Shapley Values", doi: https://doi.org/10.48550/arXiv.1905.04519.
[37] S. Augenstein, et al., "Generative Models for Effective ML on Private, Decentralized Datasets", doi: https://doi.org/10.48550/arXiv.1911.06679
[38] J. Silberg and J. Manyika, "Notes from the AI frontier: Tackling bias in AI (and in humans)", McKinsey, 1999.
[39] M. Hardt, et al., "Equality of opportunity in supervised learning," 30th Conference on Neural Information Processing Systems, Barcelona, Spain, December 2016.
[40] B. H. Zhang, et al., "Mitigating unwanted biases with adversarial learning," AAAI/ACM Artificial Intelligence Ethics and Society 2018 (AIES '18), New Orleans, USA, February 2018.
[41] T. Li, et al., "Fair resource allocation in federated learning", doi: https://doi.org/10.48550/arXiv.1905.10497.
[42] M. Mohri, et al., "Agnostic Federated Learning", doi: https://doi.org/10.48550/arXiv.1902.00146.
[43] J. Qian, et al., "Towards Federated Learning: Robustness Analytics to Data Heterogeneity", doi: https://doi.org/10.48550/arXiv.2002.05038.
[44] J. Park, et al., "Wireless Network Intelligence at the Edge," in Proceedings of the IEEE, vol. 107, no. 11, pp. 2204-2239, Nov. 2019.
[45] Y. Jiang, et al., "Model Pruning Enables Efficient Federated Learning on Edge Devices", doi: https://doi.org/10.48550/arXiv.1909.12326.
[46] X. Wang, et al., "In-Edge AI: Intelligentizing Mobile Edge Computing, Caching and Communication by Federated Learning", IEEE Network. PP. 1-10. 10.1109/MNET.2019.1800286.