Ericsson White Paper
GFTL-23:000757 Uen
June 2023

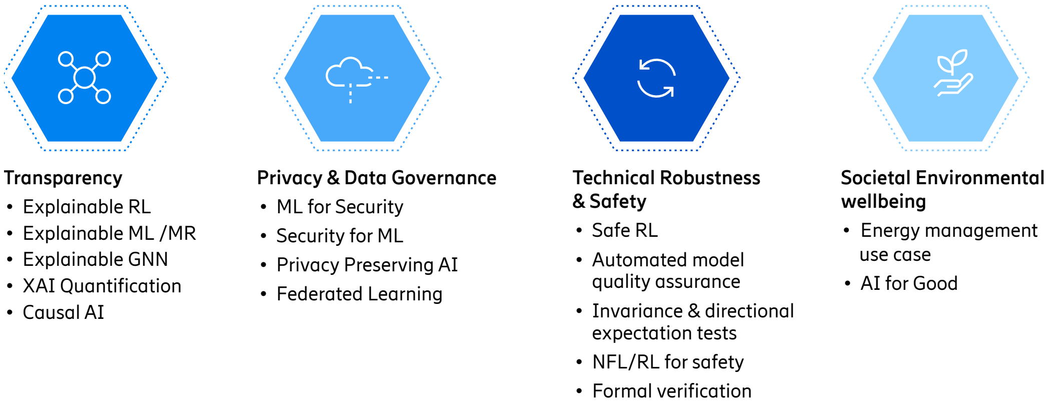
Trustworthy AI - What it means for telecom

Content

Introduction

Billions of people have come to trust and depend on modern telecom systems to support their needs and quality of life. As these systems adopt new technologies, it's important trust is maintained by understanding and addressing any new risks. Artificial intelligence (AI) differs from traditional software in its construction and operation and may introduce new and varied risks, calling for new countermeasures and guardrails. For example, the large amount of data used for AI training raises the possibility of privacy risks. Development procedures must ensure that AI models learn what is intended. The model operation should be thoroughly understood, for example, using explainability techniques. In short, to maintain the trustworthiness of the overall system, AI must itself be trustworthy: meaning, it should operate as intended and do no harm physically or ethically.

Governments, companies, and standards bodies around the world are taking notice of these facts and creating requirements regarding the trustworthiness of AI systems. The upcoming European Union AI Act [1] is one such effort. It follows principles written by the European Commission High-Level Expert Group in their "Ethics Guidelines for Trustworthy AI" [2]. Ericsson has adopted these guidelines. The framework breaks trustworthiness into seven specific areas. This paper explores how six of these areas apply to AI in telecom systems, some of which are depicted in Figure 1.



Transparency
- Explainable RL
- Explainable ML /MR
- Explainable GNN
- XAI Quantification
- Causal AI

Privacy & Data Governance
- ML for Security
- Security for ML
- Privacy Preserving AI
- Federated Learning

Technical Robustness & Safety
- Safe RL
- Automated model quality assurance
- Invariance & directional expectation tests
- NFL/RL for safety
- Formal verification

Societal Environmental wellbeing
- Energy management use case
- AI for Good

Figure 1: Ericsson's activities and technologies

Human Agency and Oversight

Human agency and oversight requirements make sure that humans can always intervene in AI-controlled systems before things like fundamental rights and safety have the potential of being affected—in other words before they could pose any harm. Implementation of such requirements means having “humans in the loop”, the difficulty of which depends on the timescale of the decision and the criticality of the system.

Sometimes AI operates at timescales, which are much too fast for human intervention, such as the optimization of radio operations in a base station. In telecom, these uses typically do not have a direct impact on the rights and safety of individuals. However, they can dramatically impact network operations, which can subsequently affect humans, so they need careful assessment.

Implementing human agency and oversight in AI-aided network operations requires user research, product function design, and user testing to make sure that network operation engineers can detect and intervene when needed. The human-machine interaction design must be aligned with existing network operation processes, rely on existing interfaces, and provide actionable information to users.

The interface for human agency and oversight may vary depending on the use case and users. It can be a GUI, CLI, Rest-based API, or even a physical interface (for example, a light). For instance, an AI system that detects and predicts network-wide congestion may use a graphical interface embedded in a dashboard used for regular network design and operations. The system may also send alerts if AI performance deteriorates (for example, by reporting too many false network congestions) to the network operations center (NOC). The alerts should include everything the NOC engineer needs, including reasons for the problem, potential root causes, and solution suggestions. The engineer then has three choices: to switch to a non-AI based function; to understand and solve the AI issue; or to escalate it. Alert volume should be considered, to not overwhelm a potentially already heavily loaded NOC. Explainable AI methods can help in generating needed, and user-tailored, reasons for problem.

The use case, user knowledge, and persistence of the AI notification influence the actions to be taken. Intermittent or ephemeral events may need to be repeated. More serious events may need an escalation path.

Transparency

Trust can arise from understanding how a system works, or from experience using it over time. The complexity and black-box nature of AI can lead to suspicion, particularly when people feel that the AI's own creators don't fully understand how it makes decisions, and what exactly it has learned. Greater transparency can help build trust by understanding and explaining AI models to humans. Explainable AI (XAI) refers to methods and techniques that produce models which show why and how an AI algorithm has made a certain decision. It helps stakeholders understand how decisions are being made in different formats: by identifying what input factors were most important in making an inference and by providing explanations and responding to "why" and/or "what-if" questions. It also helps a human operator in decision-making. If the operator is not satisfied with a response, a further investigation can be performed, using computational argumentation techniques.

Creators of AI for telecom should provide XAI methods to help build the trust of their direct customers (for example, service providers), and in turn, enable them to build trust for their subscribers. The explainability of AI should start with design and continue through implementation, as a built-in feature, to ensure transparency throughout the AI development lifecycle. In addition, different XAI techniques should be researched, and developed to explain different types of machine learning (ML) methods.

The Ericsson white paper [3] presents different XAI techniques applied to different AI/ML methods, including machine reasoning (MR), and reinforcement learning (RL). The explanations generated by these XAI techniques not only help explain the decisions to humans but also support automation, for example, in root cause analysis when combined with other AI techniques.

Explainability of ML, that is, feature analysis techniques (including SHAP and LIME) can be used in multiple telecom use cases to identify and explain the problems and root causes of specific ML model outputs in addition to ensuring the overall correctness of the ML models. These techniques can be applied to ML-based predictions to investigate the most important features that contribute to certain prediction results and validate the correctness of the ML model. The results of these explainability techniques can support MR components in identifying the root cause of the problem. 5G slice assurance is one such use case where these techniques are thoroughly investigated and tested. In this use case, certain Quality-of-Service (QoS) requirements (such as throughput, latency, and availability) are agreed upon with the customer in a service level agreement (SLA) and must be met throughout the lifecycle of the slice. ML models are used to proactively identify any potential violation of the agreed QoS requirements. Upon a violation in prediction, explainability techniques are applied to identify the most contributing features which in turn helps NOC Engineers in identifying the root cause of the problem [4]. These techniques can be applied to multiple use cases in a similar manner, like cell shaping and key performance indicator (KPI) degradation prediction, focusing on latency- and network throughput—related optimizations.

Explainability of RL: RL is suitable to solve many cellular network problems due to its dynamic nature, online training, interaction with the environment, and outstanding performance over traditional rule-based techniques for the telecom domain. An RL agent performs an action (such as applying a policy) in an environment to maximize rewards. The explainability of RL includes methods applied to different RL components, such as rewards and policy explanations.

In a base station, the antennas are tilted up, down, or kept the same to optimize KPIs, that is, the coverage of the network, increased quality by reducing interference, and capacity/throughput of the network. Coverage refers to the area from which a UE can access the cellular network, while capacity refers to the amount of traffic the cellular network can handle simultaneously. Remote electrical tilt (RET) refers to adjusting the tilt of the antenna by an RL agent to optimize the above mentioned KPIs. Increasing the down-tilt reduces the area covered by the antenna, with the risk of leaving a certain area without coverage but increases the capacity in the covered area due to a stronger signal. In contrast, up-tilt results in a larger area covered but lower capacity due to a weaker signal. Explainability is important in the RET optimization of antennas in a cellular network. Explanations help in understanding the reasons behind a specific adjustment. The following explainable reinforcement learning (XRL) methods are applied to this use case [5]:

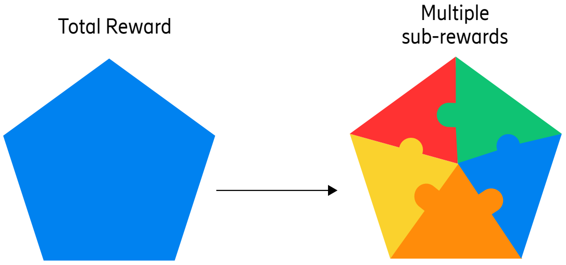


Figure 2: Total Reward is decomposed into multiple sub-reward functions for better explainability

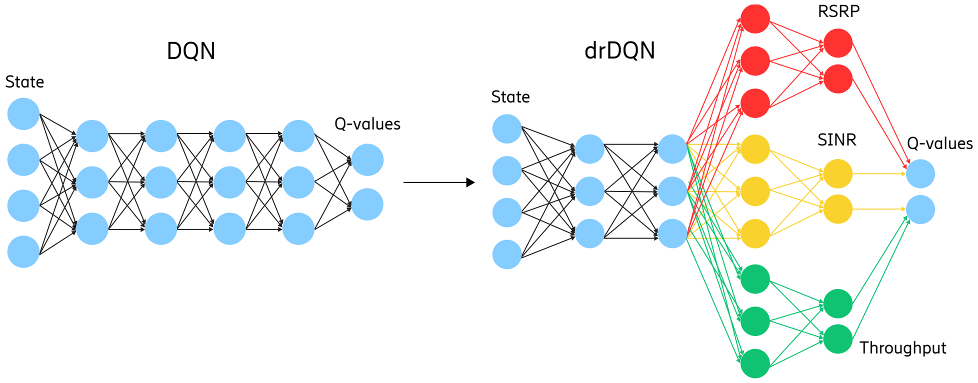


Figure 3: Reward decomposed into coverage (RSRP), quality (SINR), and capacity (throughput) sub-functions for RET use case

- Reward decomposition provides intuitive contrastive local explanations for the agent's decisions by decomposing the reward into multiple sub-functions to adjust the tilt (see Figures 2 and 3), while achieving the same performance as the original DQN algorithm. The generated contrastive explanations are very user-interpretable, as they concisely answer questions in the form of "why did you decide to down-tilt instead of up-tilting?"
- The Linear Model U-Tree (LMUT) reaches high performance while employing a fully transparent linear model capable of generating both local and global explanations (see Figure 4), however, it is less transparent than reward decomposition.
- Autonomous Policy Explanation summarizes the trained policy and explains it in natural language, thus enabling the policy to be understood by everyone including non-experts.
- Contrastive explanation through embedded self-prediction produces a local explanation about the internal representation of the RL agents (intermediate or inner layers of the deep neural network). It compares two different actions, such as why the antenna is tilted down and not up.

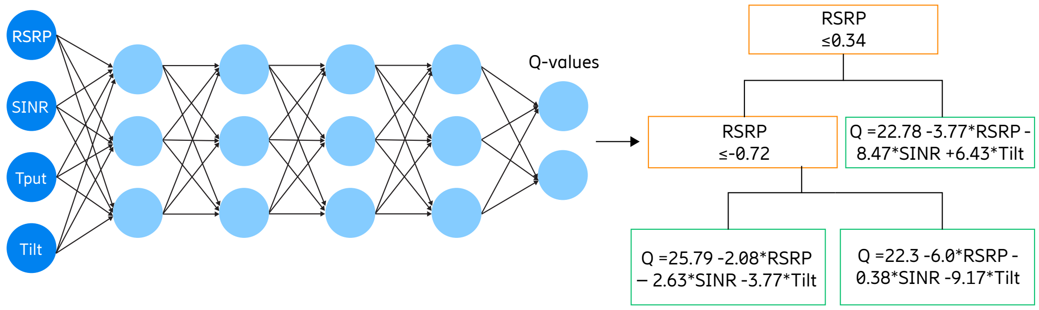

RSRP
SINR
Tput
Tilt
Q-values
RSRP
≤0.34
RSRP
≤-0.72
Q =22.78 -3.77*RSRP - 8.47*SINR +6.43*Tilt
Q =25.79 -2.08*RSRP – 2.63*SINR -3.77*Tilt
Q =22.3 -6.0*RSRP - 0.38*SINR -9.17*Tilt

Figure 4: LMUT distils a transparent model from a trained agent by transferring the NN to a decision tree with linear regression in the leaf node, where the tree is inherently transparent

In addition to enabling transparency and AI automation, Ericsson has seen the potential of reducing the input feature set by using XRL. A novel method was developed for connecting explanations from both the input (feature analysis) and output (reward) ends of a black-box RL model, resulting in fine-grained explanations [6]. Reward prioritization, performed by the user, generates two different levels of explanation, and allows RL agent reconfigurations when unwanted behaviors are observed.

Privacy and Data Governance

It is usually necessary or desirable to prevent AI systems' data from being disclosed. If the data includes the personal data of individuals, it may be subject to stringent legal requirements. Business data might contain intellectual property or be subject to contractual constraints. Sometimes, the data or other information related to it can be inferred from an AI model, especially when combined with publicly available data sources. Laws and regulations requiring privacy for individuals, such as the EU General Data Protection Regulation (GDPR) [13], predate the mass adoption of AI across industries. Despite being written in a manner to make them future-proof, such regulations don't necessarily anticipate the extent of potential AI risks. There are also often contractual requirements for privacy. And even in the absence of such requirements, it is generally understood that maintaining privacy is ethically the right thing to do.

Ensuring privacy in telecom AI impacts the entire AI lifecycle and requires the application of Privacy by Design and Privacy by Default (as defined by GDPR [7] and other global privacy laws), starting with controls where the training data is collected and continuing through model use (inference). Many controls relevant to AI are the same as or similar to those used in other types of data processing. The nature of the data, the purpose of the collection, who will use it, and how and when it will be used, should be clearly communicated. Only the minimum amount of data required for the intended purpose should be collected. Unnecessary fields should be redacted or masked. Controls like pseudonymization, encryption, authentication, and authorization should be used to ensure appropriate access. Uses of the data should be logged and auditable. Once no longer needed, data should be securely erased.

Since AI models learn from the training data, in some sense the data is encompassed within the model. Therefore, it's unsurprising that attacks exist on models to extract the training data itself or make inferences about it. Privacy enhancing technologies (PETs), including differential privacy, exist to help make AI models less susceptible to such attacks.

Developers can also analyze model sensitivity to data extraction, for example, how many queries are required. Such metrics can inform decisions about the privacy risk involved in deploying the model.

Data can be exposed during training if it is improperly secured. When federated learning is used, multiple participants with independent datasets can contribute to building a single global model. This helps them keep their datasets private, but the protocols used must be carefully designed to ensure no information is leaked.

Diversity, Non-discrimination and Fairness

Bias in models can come from bias in the training data, which itself may stem from historical prejudices and inequities. It may also be caused by disproportional representation. One example is natural language processing systems used for interaction with subscribers, such as chatbots, or support ticketing systems. Even within a single language, training such systems should consider different speaking styles, idioms, and education levels. Not everyone speaks perfectly, but everyone deserves the same level of service.

Steps to avoid bias include:

- understanding what categories exist in the input data that need to be treated equally
- making sure each category is adequately represented in the training set, regardless of size
- being aware of historical inequities that might be relevant to the problem at hand and adjusting the data accordingly
- ensuring model robustness, including thorough testing
- considering categories separately throughout the model development process, so that the model performs well for each

Even a seemingly solely technical problem can have a bias that impacts people, sometimes in subtle ways. Consider the use of AI to structure and operate mobile networks. This can be impacted (or biased) based on how and where that data is collected, or by cognitive bias already present in the system's creators. If data collection is skewed toward a particular group (for example, people with a certain economic status, which might correlate to other factors such as race), the resulting system might be inadvertently biased. Telecom systems have to work everywhere in the world and should provide equally good service to all people.

If more investment (that is, equipment, optimization, effort) goes into certain areas, some of these areas might get disproportionately better service. Such areas could have different usage patterns. When ML that is used to plan, deploy, optimize, and operate networks is trained on data collected in some regions, networks (or products) might be created that work well only in those regions.

Technical Robustness and Safety

Human safety is not typically directly impacted by telecom operations but can be affected by situations like loss of service. Telecom systems can be important components in emergency communications and disaster handling. Another example is an application like autonomous vehicles, where the loss of communication might impact the ability to proceed safely.

AI can assist in these situations when it makes the network itself more robust. But this means that the AI itself must be robust. Careful attention to AI quality must be taken during training and deployment. Fallback mechanisms should be in place for cases where the AI cannot decide or makes an out-of-bounds decision to, for example, transfer the control to a human operator.

AI is also subject to new types of attacks. Data extraction attacks, which have already been mentioned, compromise the training data and potential privacy. Researchers have also demonstrated attacks on inference using adversarial examples. Carefully crafted inputs are fed to the model, causing its inference to be biased in a direction chosen by the attacker. Where the attacker has access, poisoning the training data can be used to influence model operation. A motivated attacker might use these mechanisms to affect network operations or steal a service. The training pipeline, resulting models, and surrounding application context should be analyzed for susceptibility to such attacks.

Automated model quality assurance is crucial for technical robustness and safety. Models must be thoroughly tested against performance metrics that reflect potential real-world scenarios. These metrics are use case and model specific. For instance, in the case of a classification model, accuracy, false positive or false negative rates might be measured, while in the case of clustering, Silhouette Coefficient [8] and Dunn Index [9] might be chosen.

It is also important to communicate this information in an understandable and useful manner to model recipients, keeping in mind that they may not have expertise in AI. For example, when communicating a performance metric about a model, information should be included about its meaning and what values are considered good or bad.

Since a model's performance depends on the data set used for training and evaluation, data quality is essential for model quality. A training data set should accurately represent reality and cover the events or objects of interest: the training data set must have the same statistical properties the real objects have, and if there are relationships between the attributes of the real object or among the real objects, those relationships must be preserved in the training data set.

The RET use case mentioned earlier shows that poor AI decisions could lead to interference and compromised network performance. Anywhere an AI algorithm is used in the operation of a network, a failure of that algorithm can lead to inefficiency, instability, or, at worst, downtime.

Another important concern is that AI models could, inadvertently or maliciously, take actions that are unsafe for humans. As already mentioned, telecom systems by nature are not safety-critical, and the danger could be reduced network performance. One scenario could be when the model explores the space of all possible states and actions. This becomes significantly important for RL, where space exploration is seen as an effective way to train an RL agent to capture a near-optimal policy. However, unchecked exploration can lead the system to visit a dangerous state, for example, when the system tries to tilt the antenna at an overly high angle. Safe RL methods provide a shield to block unsafe actions that might result from free exploration of state-action spaces. The intention is to allow the agent some state exploration of the environment while having boundaries using safety specifications defined by a human developer. The specifications, or boundaries, can be user dependent. One such use case where these techniques are successfully tested is RET Optimization [10].

Conducting invariance and directional expectation tests is also essential to assess and assure the model's robustness. In an invariance test, label-preserving perturbations are applied to inputs and the model prediction is expected to remain the same. In a directional expectation test, a set of perturbations are made to the input which should have a predictable effect on model output.

Social and Environmental well being

AI can be used to create positive benefits for society, such as by helping protect the environment. Ericsson considers communication to be a fundamental human right, so the availability of the network is core to societal well-being. The findings of this white paper support this conclusion. AI, by helping service providers create telecom networks that are more reliable, ubiquitous, and inexpensive, contributes to the social goal of universal communication. But there are possible negatives, such as the privacy aspects discussed above. ML training can be energy intensive, so careful cost-benefit analysis is needed before it is employed. ML might be used where there are potential safety or societal risks (for example, control of critical infrastructure). Minimizing those risks is important, and this can be done using the techniques described in this white paper.

AI can help improve network operations and energy consumption. Large volumes of data can be used to optimize important goals like performance, reliability, capacity, and energy usage. Traditional optimizations, written by programmers, typically used only a few parameters and simple algorithms with limited results. AI allows large numbers of parameters to be used, better optimizations, and therefore better performance. Ericsson teams looking at specific use cases have shown that AI can lead to significant energy savings compared to traditional algorithms. [11].

AI in telecom can be useful beyond the simply better operation of the network. During the COVID-19 pandemic, Ericsson engaged in a joint project with a service provider, government officials, and two hospitals [12]. The service provider provided anonymized and aggregated data about people's movements, taken from their network. This was combined with vaccination, antibody test, and hospital COVID patient admission data. A series of

ML models used the data to predict admissions two to three weeks into the future. In eleven of sixteen weeks, the predictions had an error rate of less than thirty percent. Better resource planning for hospitals, especially during a crisis, leads to better patient care. This demonstrates that AI and telecom networks and data can be used to benefit society in novel and perhaps unexpected ways.

Conclusion

The benefits of AI in telecom networks are only just beginning to be leveraged but will clearly be an important and integral element in future networks. Trusting those networks requires trusting the AI, which can be achieved by following the presented guidelines. However, these present a number of challenges:

- Ensuring that humans retain oversight and control over AI systems, even when highly automated and operating at high speeds.
- Providing information about AI operations, using techniques such as XAI, while maintaining a fine balance between the privacy of models and data and transparency
- Protecting the data of users and businesses, while still using it to deliver AI that benefits them
- Understanding how AI might affect different communities and preventing adverse impacts
- Making AI systems robust and safe, yet practical to train and deploy.
- Considering how AI affects society broadly, both to prevent adverse impacts and to promote beneficial uses.

Trusting the AI, that is having confidence that it operates as intended and does no harm, requires diligence in addressing each of these challenges. Ericsson is working to continuously improve its AI systems, making them, and consequently their products, more trustworthy.

Glossary

AI	Artificial Intelligence
CLI	Command Line Interface
DQN	Deep Q Network
GDPR	General Data Protection Regulation, https://gdpr-info.eu/
GUI	Graphical User Interface
KPI	Key Performance Indicator
LIME	Loacl Interpretable Model-agnostic Explanations
LMUT	Linear Model U-Tree
ML	Machine Learning
MR	Machine Reasoning
NN	Neural Network
NOC	Network Operations Center
PET	Privacy Enhancing Technologies
RET	Remote Electrical Tilt
RL	Reinforcement Learning
RSRP	Reference Signal Received Power
SHAP	SHapley Additive exPlanations
SINR	Signal to Interface plus Noise Ratio
SLA	Service Level Agreement
UE	User Equipment
XAI	Explainable Artificial Intelligence
XRL	Explainable Reinforcement Learning

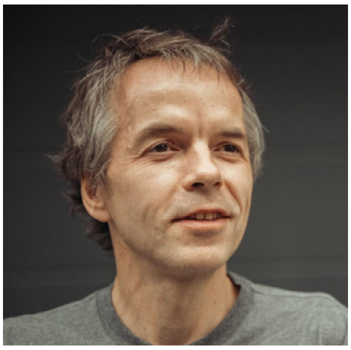
References

1. https://artificialintelligenceact.eu/
2. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
3. Rafia Inam, Ahmad Terra, Anusha Mujumdar, Elena Fersman, Aneta Vulgarakis, Explainable AI - How humans can trust AI. EricssonWhite Paper, April 2021
4. A. Terra, R. Inam, S. Baskaran, P. Batista, I. Burdick and E. Fersman, "Explainablity Methods for Identifying Root-Cause of SLA Violation Prediction in 5G Network," in IEEE Global Communications Conference, 2020
5. Franco Ruggeri, Ahmad Terra, Rafia Inam, Karl-Henerik Johansson, "Evaluation of Intrinsic Explainable Reinforcement Learning in Remote Electrical Tilt Optimization", in 8th International Congress on Information and Communication Technology, 2023
6. Ahmad Terra, Rafia Inam, Elena Fersman. BEERL: Both Ends Explanations for Reinforcement Learning. Journal of Applied Sciences, Special issue "Explainable Artificial Intelligence", Vol 12, No. 21, November 2022
7. EU GDPR Privacy by Design, https://gdpr-info.eu/issues/privacy-by-design/
8. Peter J. Rousseeuw (1987). "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis". Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.
9. Dunn, J. C. (1973-09-17). "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters". Journal of Cybernetics. 3 (3): 32–57. doi:10.1080/01969727308546046. S2CID 120919314.
10. A. Nikou, A. Mujumdar, V. Sundararajan, M. Orlic and A. V. Feljan, "Safe RAN control: A Symbolic Reinforcement Learning Approach," 2022 IEEE 17th International Conference on Control & Automation (ICCA), 2022
11. https://www.ericsson.com/en/news/2023/4/ericssons-service-continuity-ai-app-delivers-25-percent-energy-savings-for-far-eastone
12. Taghia, J., Kulyk, V., Ickin, S. et al. Development of forecast models for COVID-19 hospital admissions using anonymized and aggregated mobile network data. Sci Rep 12, 17726 (2022). https://doi.org/10.1038/s41598-022-22350-6
13. EU General Data Protection Regulation, https://gdpr-info.eu/t

Authors

Jim Reno is a Distinguished Engineer at Ericsson, where he works on security aspects of Artificial Intelligence as applied to telecommunication systems. He has more than 40 years of industry experience in fields including system software (operating systems, networking, system management, and cloud native systems), payment system security, authentication, authorization and identity management.

Rafia Inam is a senior research manager at Ericsson Research and Adjunct Professor at KTH in research area Trustworthy Artificial Intelligence. She has conducted research for Ericsson for the past nine years on 5G for industries, network slices and network management; AI for automation and intelligent transport systems. She specializes in automation and safety for cyber-physical systems and collaborative robots, trustworthy AI, and explainable AI. She won Ericsson Top Performance Competition 2021 on her work on AI for 5G network slice assurance, and was awarded Ericsson Key Impact Award 2020, and Key contributor award 2020. Rafia received her PhD in predictable real-time embedded software from Mälardalen University in 2014. She has co-authored 55+ refereed scientific publications and 55+ patent families and is a program committee member, referee, and guest editor for several international conferences and journals.



Attila Ulbert joined Ericsson in 2015 and he is currently Artificial Intelligence System Manager. In his enthusiastic journey with Ericsson, he lead the development of Ericsson's AI platform, and worked on fundamental AI studies on security, trustworthiness, and industrialization. Attila has a PhD in Informatics from Eötvös Loránd University. He is a marathoner.