Ericsson White Paper
GFTL-21:000529Uen
April 2021

ERICSSON

Explainable AI – how humans can trust AI

Content

Introduction

Artificial intelligence (AI) has achieved growing momentum in its application in many fields to deal with the increased complexity, scalability, and automation, and that also permeates into digital networks today. A rapid surge in the complexity and sophistication of AI-powered systems has evolved to such an extent that humans do not understand the complex mechanisms by which AI systems work or how they make certain decisions — something that is particularly a challenge when AI-based systems compute outputs that are unexpected or seemingly unpredictable. This especially holds true for opaque decision-making systems, such as those using deep neural networks (DNNs), which are considered complex black box models. The inability for humans to see inside black boxes can result in AI adoption (and even its further development) being hindered, which is why growing levels of autonomy, complexity, and ambiguity in AI methods continues to increase the need for interpretability, transparency, understandability, and explainability of AI products/outputs (such as predictions, decisions, actions, and recommendations). These elements are crucial to ensuring that humans can understand and — consequently — trust AI-based systems (Mujumdar, et al., 2020). Explainable artificial intelligence (XAI) refers to methods and techniques that produce accurate, explainable models of why and how an AI algorithm arrives at a specific decision so that AI solution results can be understood by humans (Barredo Arrieta, et al., 2020).

Without explanations behind an AI model's internal functionalities and the decisions it makes, there is a risk that the model would not be considered trustworthy or legitimate. XAI provides the needed understandability and transparency to enable greater trust toward AI-based solutions. Thus, XAI is acknowledged as a crucial feature for the practical deployment of AI models in systems and, more importantly, for satisfying the fundamental rights of AI users related to AI decision-making (according to European Commission ethical guidelines for trustworthy AI). Standardization bodies such as the European Telecommunications Standards Institute (ETSI) and the Institute of Electrical and Electronics Engineers Standards Association (IEEE SA) also emphasize the importance of XAI where AI models are deployed, indicating XAI's growing importance in the future (Frost, et al., 2020). AI deployers and developers must comply with these ethical guidelines and regulations to ensure that their AI solutions are explainable and trustable (Anneroth, 2019).

The aim of this white paper is to increase understanding of the emerging and vital technology that is XAI and its development as a built-in feature with AI to comply with regulations as well as examine its applicability to the telecommunications domain as an integral and fundamental part of AI-based systems to meet the demands for explainability and transparency. This paper additionally provides a framework where XAI components serve as integrated parts of AI systems (thus supporting explainability by design), in which explainability methods can be easily incorporated.

XAI — challenges and objectives

Building trust is essential for users to accept AI-based solutions and those systems incorporating decisions made by them. There are, however, significant challenges in developing explainability methods. One of them is the trade-off between attaining the simplicity of algorithm transparency and impacting the high-performing nature of complex but opaque models (when one increases the transparency aspect, privacy and the security of sensitive data come into question). Yet another challenge is to identify the right information for the user, where different levels of knowledge will come into play.
Beyond selecting the level of knowledge retained by the user, generating a concise (simple but meaningful) explanation also becomes a challenge. Most explainability methods focus on explaining the processes behind an AI decision, which is sometimes agnostic to the context of its application, providing unrealistic explanations. Researchers attempt to integrate knowledge-based systems so that the explanation becomes relevant to its application's context.

XAI helps deliver trust by supporting with the following properties:

- trustworthiness, to attain the trust of humans on the AI model by explaining the characteristic and rationale of the AI output
- transferability, where the explanation of an AI model allows better understanding of it so that it can be transferred to another problem or domain/application properly
- informativeness, relating to informing a user regarding how an AI model works in order to avoid misconception (this is also related to human agency and autonomy, which ensures humans understand AI outcomes and can take intervening actions on that basis)
- confidence, which is achieved through having a model that is robust, stable, and explainable to support human confidence in deploying an AI model
- privacy awareness, ensuring that the AI and XAI methods do not expose private data (which can be done through data anonymization)
- actionability, with XAI providing indications regarding how a user could change an action to yield a different outcome in addition to providing the rationale for an outcome
- tailored (user-focused) explanations, allowing humans — as AI system users of different knowledge backgrounds — to understand the behavior and predictions made by AI-based systems through tailored explanations based on their roles, goals, and preferences

XAI by design

AI methods are generally categorized into machine learning (ML), machine reasoning (MR), and the interplay of ML and MR. The interplay between learning and reasoning can be understood from the ETSI Zero-touch network and Service Management (ZSM) group's autonomic loops architecture (Zero touch network & Service Management (ZSM), 2017), where data collected from the environment is analyzed using ML approaches and the derived information and insights are fed to MR systems to compute decisions (which are then actuated in an automated fashion, as shown in blue color components and arrows in Figure 1).

It's important to incorporate interpretability and explainability at different levels of complex AI techniques. The XAI framework presented in this white paper is tightly linked with providing explanations for both different AI techniques (ML and MR techniques) and the environment through properly defined interfaces.

The main components of this framework center on explanations, explainability for data, explainability for ML, and explainability for MR (see purple parts of Figure 1). The distinctive approach that we are taking is to apply explainability to both ML and MR as well as to the interplay between ML and MR by feeding the output of an ML model (both its predictions and explanations) into our MR techniques and applying XAI in order to generate explanations. This proactive placement provides the right AI trustworthiness early on rather than relying on reactive fixes. Furthermore, this framework allows the integration of new XAI algorithms into the respective explainability components so that in the future, newly developed XAI techniques for ML/MR can be easily deployed within the explainability for ML/MR components.

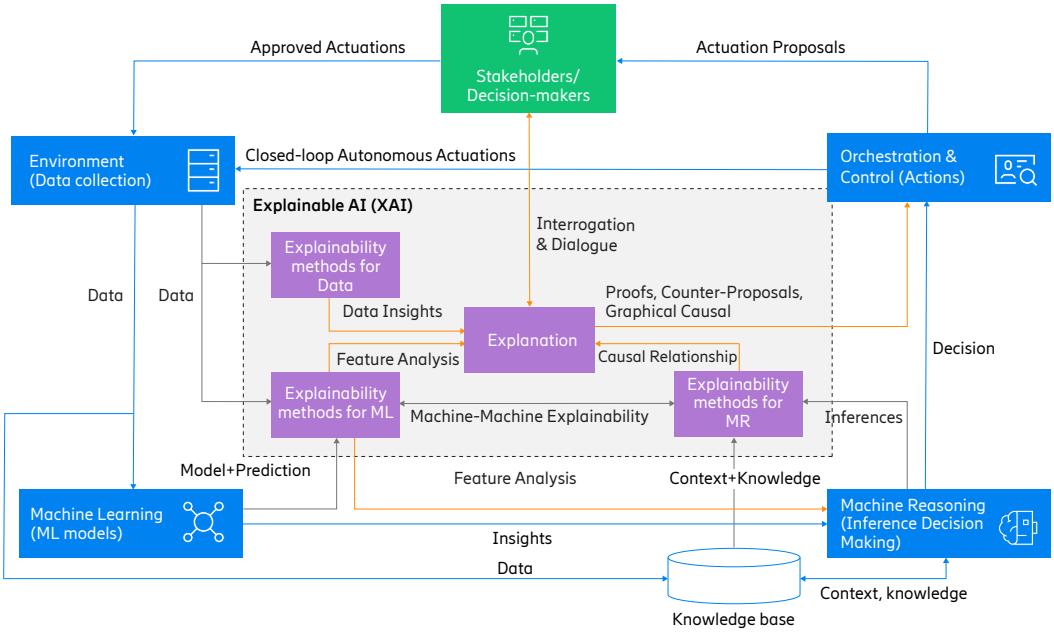

Figure 1. Overview of XAI methods and their link to data, ML, and MR

Explanation(s):
This refers to the output from XAI methods that is presented in several forms based on the used explainability technique. It contains the details on how a model generates predictions or important feature sets from the data that is causing a problem, a decision path from a decision tree model, a rule generated from a simplified model, visualization of the data, and more.

Explainability for data:
Data explanation methods are used in the initial step of building AI systems in order to identify dataset characteristics, which may include statistical information of the datasets (for example, analysis of any features correlation, choosing important instances from the datasets, and so on). This step is important for choosing a better-suited AI model that will perform predictions (for example, a decision tree can be used if the data is simple or a DNN model can be used for more complex data). Without data explanation, it would be difficult to identify the full potential of available data.

Questions answered by different explainability methods

By explainability for data	By explainability for ML	By explainability for MR
What kind of data do we have?	Why does the model predict this outcome?	What caused event X? If Y caused event X, what would change the outcome of X?
What does this data inform us of?	Which factors significantly impact the decision?	How does/did the system perform a given action?
Which parts of the data are valuable?	What can be done to obtain a different result?	Why is the system concerned with a parameter, goal, or action?
How is the data distributed?	Which conditions need to be fulfilled in order to maintain current results?	Why is a given inference or decision necessary?
	What happens if the data is modified?	When did the system consider/ reject a goal, decision, or inference?
		What are the effects of a decision or parameter setting?
		What is the relationship between two terms, values, goals, or actions?
		What factors does the system consider/ignore in deriving an inference?
		What methods does the system use or avoid in achieving a goal/inference?

Explainability for ML:
An ML model is built using a certain method based on the properties of the available data. A state-of-the-art model is used to achieve very high performance in predicting or classifying instances; however, as the complexity of the models is continuously increasing, it is becoming harder to understand how particular predictions are made and which factors contributed (and how much) to a given prediction. XAI methods are used to capture and explain the characteristics of complex ML models. When a prediction is complemented with an explanation, it is expected that the ML model will gain more trust from humans.

Explainability for ML models requires the data, an ML model, and the output/prediction of the ML model as input in order to generate the explanation (as depicted in Figure 1). The explanations are presented directly to humans or analyzed further by an MR system. Having not only prediction but also explanation from ML models enriches information for MR in making decisions (for example, in fulfilling an intent). Without ML explainability, one could not be sure whether the model complies with the rules or standards that the model should fulfill (for example, being not biased towards sensitive features).

Explainability for ML can be used at different stages in developing an AI system. Once a black box ML model is built with satisfactory performance, XAI methods (for example, SHAP (Lundberg & Lee, 2017), XGBoost (Chen & Guestrin, 2016), Causal Dataframe (Kelleher, 2017), PI (Altmann, et al., 2010), and so on) are applied to obtain the general behavior of a model (also known as "global explanation"). While the global explanation, which explains the whole data, is important, it is also necessary to obtain the explanation of every single data prediction (also known as "local explanation"). Whenever new/more data is fed into the model, the rationale of every ML prediction is presented in a local manner, where one explanation may be different from another. Some methods mostly used for local explainability include LIME (Ribeiro, Singh, & Guestrin, 2016), SHAP (Lundberg & Lee, 2017), ELI5 (Korobov & Lopuhin, 2019), CFProto (Looveren & Klaise, 2019), and more.

XAI for MR:
Machine reasoning is a field of AI that complements ML by aiming to computationally mimic abstract thinking by way of uniting known information with background knowledge, making inferences regarding unknown or uncertain information. These cover techniques such as AI planning, constraint solving, logical inference, and so on.

Various techniques for explanations in reasoning systems (Cyras, et al., 2020) have also been developed, such as contrastive explanations, argumentation techniques for explaining conflict resolution, and analysis of unsolvability or inconsistencies. Explainability techniques in MR have the advantage of being understandable by human operators who are not domain experts or data scientists since the system ontology can be mapped to user models (due to the symbolic nature of MR). Furthermore, MR techniques such as answer set programming, automated theorem proving, and minimum unsatisfiable core (among others) provide explanations both in ML and MR.

The model-based nature of MR techniques makes them well suited for explanations that are actionable, tailored, and interactive (Chakraborti, Sreedharan, & Kambhampati, 2020). Without MR explainability, it would be hard to explain cognitive processes involved in complex AI-based decision-making tasks.

Machine-to-machine explainability

Today's complex AI deployments are often composed of multiple AI agents, and this is expected to increase in the future. Such agents may be autonomous yet coordinate among each other to meet business intents. In such a scenario, it is important for machines to generate explanations towards other machines (or agents). We term this idea "machine-to-machine explainability" (M2M explainability) or "agent-to-agent explainability." Consuming explanations from other agents also enables explanations to be tailored to humans by presenting the view from across the full system at a higher level of abstraction. MR explainability techniques leverage compositionality, and can therefore enable such M2M explanations (examples of some techniques are computational argumentation and iterative contrastive explanations). This pertains to the arrow from XML to XMR in Figure 1. M2M XAI is especially crucial for trust in multi-vendor environments where there could be agents/AI components present from various vendors interfacing with each other to meet high-level intents. Without M2M explainability, orchestration and management (O&M) of complex systems would not be efficient and would require significant human intervention. Machine reasoning is a key enabling technology for such system-level explainability, arising from the compositionality provided by MR.

Hierarchical explainability:
While various organizations of explainers can be envisioned, an interesting class of M2M explainability is hierarchical explainability. Here, a hierarchy of explainers communicates with each other to construct meaningful explanations (whether in response to user questions, or otherwise); for example, an explanation in response to a high-level user question (such as why certain specified intents are not met) can "stitch" together explanations from various levels of the hierarchy (see Figure 2). The explanation can expose how various sub-intents were weighed in order to provide the end solution, then going deeper to explain the exact constraints at the system level that invalidated certain parts of the solution space as well as the exact feature attributions that led the system to the root cause of a fault. Hierarchical explainability is particularly relevant in telecommunications systems, which are themselves often organized hierarchically. This can also shape future network architecture, with higher-level explainers being key parts of the O&M layer.

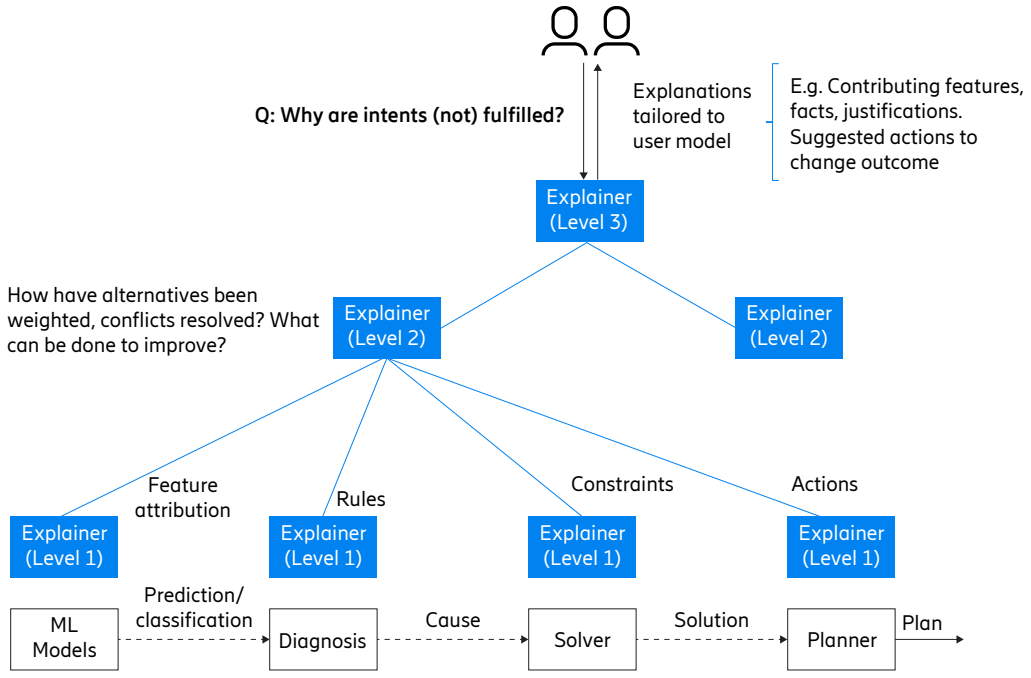

Figure 2: An illustration of hierarchical explainability

Application in telecom use cases

The importance and demand of XAI's application in telecom use cases is increasing as more complex AI models continue to be deployed to automate network operations. Specifically, future 6G networks will be heavily AI-infused (Wikström, et al., 2020). As the users of this automated process, humans may wonder how a certain decision is made. Without XAI's assurance for our solutions, AI acceptance as a whole could be limited.

Some examples of XAI's potential value for telecom use cases are examined below:

Explainability methods for identifying root causes of SLA violations in 5G network slice assurance

In 5G networks, network slices must be monitored throughout their life cycles to guarantee the quality-of-service requirements that the slices have agreed to fulfill according to the given service-level agreements (SLAs). To support slice assurance, proactive measures of predictive analysis are applied to identify potential SLA violations in advance, identify their causes, and provide recommendations to solve them.

For example, the network traffic routes pass from the customer premises (from connected devices like phones, robots, or vehicles) to the transport layer, on through the black topology in the middle, and finally reach the Tier 1 data center, where it is consumed and used by a cloud-based application (see Figure 3). ML-based prediction analysis algorithms can be

used to identify probable SLA violations based on the transport layer routers' data (probable violations are predicted in this route, as shown by the red route in Figure 3).

This is important for telecommunications service providers, allowing them to quickly identify or prevent problems that may occur in operations. In a broader application, this method can be applied to a system where preventive actions can be taken to avoid unwanted events that may occur in the future.

XAI for ML:
After identifying probable SLA violations, the next step is to identify the important features from the data contributing to the violations. Multiple ML-based explainability methods (SHAP (Lundberg & Lee, 2017), XGBoost (Chen & Guestrin, 2016), PI (Altmann, et al., 2010), Causal Dataframe (Kelleher, 2017), LIME (Ribeiro, Singh, & Guestrin, 2016), ELI5 (Korobov & Lopuhin, 2019), and CFProto (Looveren & Klaise, 2019)) can be applied to analyze the probable causes of SLA violation predictions made in the slice assurance setup. The ML model, prediction of the ML model, and the data are fed to the ML explainability module (see Figure 1), which provides important features and identification of root causes contributing to the potential SLA violations. A detailed comparative analysis of the applied XAI methods could then be provided as well as a recommendation for the most suitable XAI methods (for this use case and the data set, SHAP is the best results-provider method for both global and local explanations (Terra, et al., 2020)).

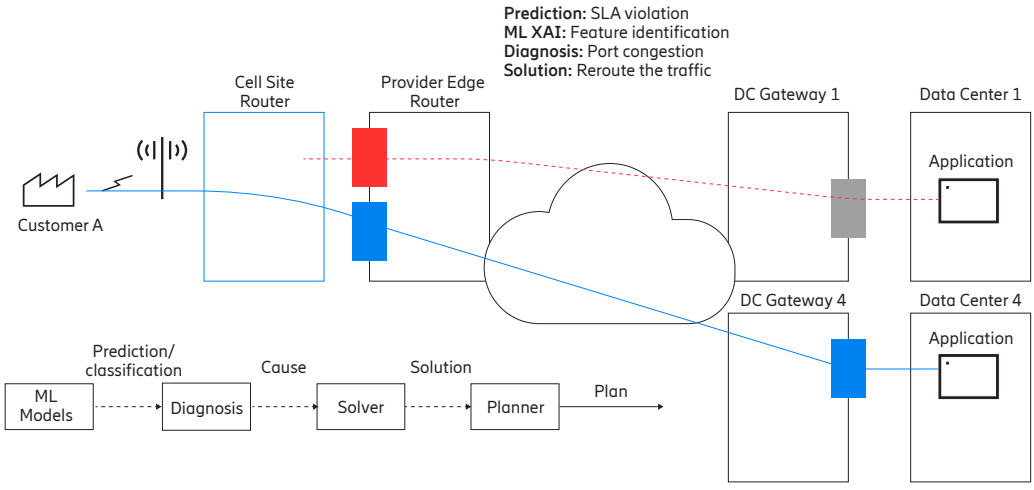

Figure 3: Explainability applied to 5G network slicing use cases

XAI for MR:
Once the main features causing an SLA violation are identified and the root cause is determined, a reconfiguration and rerouting of the slice path could be computed by MR components. The identified features would then be fed into the MR layer (as shown in Figure 1) to pinpoint the probable cause. (From an explainability point of view, we explain using both ML and MR explainability. The MR explainability is easily seen in the rules that are triggered to determine the probable cause. This is the "Diagnosis" box in Figure 3.)

We employ three MR explainability techniques:

1. contrastive explanations, which enable a human operator to ask "why" questions, and approve or reject a network actuation sequence computed by the reasoning system (if the human operator is not satisfied with a response, a further interrogation of the system can be performed, using computational argumentation techniques)
2. argumentation, showing the facts that support and conflict with intent or goal (this can reveal the reasoning behind a proposed solution)
3. persona-aware explanations, generated by proactively considering the user model (in terms of preferences, intents, and user privileges) while generating explanations

Such explanations reveal the justification behind the reasoning solutions. We also use MR techniques to make sense of the interplay between multiple explainers to arrive at meaningful high-level explanations (as in Figure 2).

XAI applied to other telecom use cases

One use case where the explainability methods for ML are applied is key performance indicator (KPI) degradation prediction focusing on latency- and network throughput–related KPIs, where an ML model predicts KPI degradations that may happen in the future. The SHAP explainability method for ML has been identified as the best suitable method to find the potential cause of a degradation, which can be later complemented with a rule-based system to provide recommendations for solving the problem.

Another use case is sleeping cell prediction, where an ML model predicts accessibility degradation four hours in advance and recommends the possible actions for humans to take to prevent it. Again, SHAP is used here for identifying the important features contributing to the potential problem.

Some other use cases where explainability for ML is used include workload prediction, anomaly detection, customer complaints reduction, and security threat analysis.

Conclusion

Explainable AI becomes a key factor in AI-based systems when it comes to meeting the demands of understandable, transparent, interpretable, and (consequently) trustworthy AI-based solutions. XAI plays a fundamental role in gaining human operator trust on one hand while supporting established guidelines, standards, and regulations on the other. This white paper has outlined the importance and potential uses of XAI applications, especially for the telecom domain. Our presented explainability framework incorporates XAI by design, works on several AI techniques (both ML and MR), and supports the easy incorporation of newly developed XAI methods. It is envisioned that cognitive networks will be a central feature to 6G (Wikström, et al., 2020), which includes AI models being deployed at scale to raise the intelligence of networks. Due to the expected scale, complexity, and criticality of such future networks, understanding and trusting these AI models and their behavior through seamlessly integrated XAI technologies becomes even more crucial.

Definition and References

AI model refers to a formal or mathematical representation generated using statistical data, facts, and expert input that captures input–output patterns or represents knowledge for decision-making.

Black box model refers to an AI model taking its input to generate its output (which can be in the form of predictions or recommendations) without human operators having any knowledge of its internal workings.

Transparency means that an AI model is understandable by itself.

Explainability means that the model is able to explain of why and how an AI algorithm arrives at a specific decision while maintaining its accuracy.

Understandability means that AI model is able to make humans understand its functionalities.

An **explaine**r is a component/agent that generates explanations.

An **explainee** is the consumer of these explanations.

1. Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. Bioinformatics, 10, 1340-1347.
2. Anneroth, M. (2019). Responsible AI – a human right? Available from https://www.ericsson.com/en/blog/2019/2/responsible-ai--human-right
3. Barredo Arrieta, A., Diaz Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado González, A., Garcia, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion, 58, 82-115.
4. Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2020). The Emerging Landscape of Explainable Automated Planning & Decision Making. Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, (IJCAI-20), 4803-4811.
5. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA. Association for Computing Machinery. pp. 785-794.
6. Cyras, K., Badrinath, R., Mohalik, S. K., Mujumdar, A., Nikou, A., Previti, A., Sundararajan, V., & Vulgarakis, A. (2020) Machine Reasoning Explainability. Arxiv. [Preprint] Available from: https://arxiv.org/abs/2009.00418
Related tutorial: https://www.ericsson.com/en/reports-and-papers/research-papers/aamas-2021-tutorial
7. European Comission. (2020). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. Available from https://ec.europa. eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence- altai-self-assessment
8. Frost, L., Meriem, T. B., Bonifacio, J. M., Cadzow, S., Silva, F. d., Essa, M., Forbes, R., Marchese, P., Odini, M., Sprecher, N., Toche, C., & Wood, S. (2020). Artificial Intelligence and future directions for ETSI. ETSI White Paper No. #34, 1(ISBN 979-10-92620-30-1).
9. Kelleher, A.(2017). Causality. Available from https://github.com/akelleh/causality
10. Korobov, M., & Lopuhin, K. (2019). ELI5 Documentation, Release 0.9.0., Available from https://readthedocs.org/projects/eli5/downloads/pdf/latest/
11. Looveren, A. V., & Klaise, J. (2019). Interpretable Counterfactual Explanations Guided by Prototypes. Arxiv. [Preprint] Available from: https://arxiv.org/abs/1907.02584
12. Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (NIPS 2017). Curran Associates, Inc. pp. 4765-4774
13. Mujumdar, A., Cyras K., Singh S., & Vulgarakis A. (2020). Trustworthy AI: explainability, safety and verifiability. Available from: https://www.ericsson.com/en/blog/2020/12/trustworthy-ai
14. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, California, USA.

Association for Computing Machinery.

15. Terra, A., Inam, R., Baskaran, S., Batista, P., Burdick, I., & Fersman, E. (2020). Explainablity Methods for Identifying Root-Cause of SLA Violation Prediction in 5G Network. GLOBECOM 2020 - 2020 IEEE Global Communications Conference. Taipei, Taiwan.
16. Wikström, G., Persson, P., Parkvall, S., Mildh, G., Dahlman, E., Balakrishnan, B., Öhlén, P., Trojer, E., Rune, G., Arkko, J., Turányi, Z., Roeland, D., Sahlin, B., John, W., Halén, J., Björkegre, H. (2020). Ever-Present Intelligent Communication - A research outlook towards 6G. Ericsson White Paper (GFTL-20:001402).
17. Zero touch network & Service Management (ZSM) (2017). Available from https://www.etsi.org/technologies/zero-touch-network-service-management

Authors

Rafia Inam is a Senior Project Manager at Ericsson Research in the area of AI. She has conducted research for Ericsson for the past six years on 5G for industries, network slices, and network management; AI for automation, and service modeling for intelligent transport systems. She specializes in automation and safety for cyber-physical systems and collaborative robots, trustworthy AI, XAI, risk assessment and mitigations using AI methods, and reusability of real-time software. Rafia received her PhD in predictable real-time embedded software from Mälardalen University in 2014. She has co-authored 40+ refereed scientific publications and 50+ patent families and is a program committee member, referee, and guest editor for several international conferences and journals.

Ahmad Terra is a WASP industrial PhD student at KTH Royal Institute of Technology with a focus on explainability for telecommunication. He holds an MSc in machine learning from KTH. Prior to joining Ericsson, Terra worked at ABB as a graduate robotics engineer. At Ericsson, Terra has previously worked on his master thesis by applying reinforcement learning for risk mitigation in a collaborative robot. His current focus is on exploring the actionability of XAI applied to 5G network slicing.

Anusha Mujumdar is a Senior Researcher at Ericsson Research focusing on trustworthy AI technologies, including explainability and safety and verification in AI systems. She joined Ericsson in 2017 after earning her PhD in control theory. Anusha's area of interest lies predominately in autonomous systems, including reinforcement learning and automated planning and optimization as applied to telecom networks and cyber-physical systems. She has coauthored more than 15 peer-reviewed journals and conference papers and has filed several patents.

Elena Fersman is a Research Director in AI at Ericsson and leads a team of researchers located in Sweden, the US, India, Hungary, and Brazil. She is a docent and adjunct professor in cyber-physical systems at the Royal Institute of Technology in Stockholm, specializing in automation. She holds a PhD in computer science from Uppsala University, an MSc in economics and management from St. Petersburg Polytechnic University, and she completed her postdoctoral research at Paris-Saclay University. At Ericsson, she has occupied various positions ranging from product management to research leadership. Elena is a member of the Board of Directors of RISE Research Institutes of Sweden and has coauthored over 50 patent families.

Aneta Vulgarakis Feljan is a Sector Manager in Machine Reasoning at Ericsson Research. Her main interests are in AI-based cyber-physical systems and the combination of model-driven and data-driven AI. She specializes in knowledge representation and reasoning, automated planning, behavioral modeling and analysis, trustworthy AI, and the application of these in telecom and non-telecom use cases. Aneta is the coauthor of more than 40 refereed publications on software engineering and AI topics and is a coinventor of over 50 patent families. Before joining Ericsson Research, Aneta was a scientist at ABB Corporate Research. Her PhD in computer science from Mälardalen University focused on component-based modeling and formal analysis of real-time embedded systems.