Auditing AI-Systems: From Use Cases to Standardization and Regulation

3rd International Workshop – Programme

hosted by the German Federal Office for Information Security (BSI), the Association of Technical Inspection Agencies (TÜV-Verband) and Fraunhofer Heinrich-Hertz-Institute (Fraunhofer HHI)

November 24th 2022, 9:00-16:30h | Hybrid Conference
Online-Livestream | from Design Offices Humboldthafen, Alexanderufer 3-7, 10117 Berlin

9:00-9:15	**Welcome Note**	**BSI / TÜV Association / Fraunhofer HHI**
9:15-10:45	**Session 1: New Approaches from Research**	**Wojciech Samek** (Fraunhofer HHI, Berlin, Germany): Concept-Level Explainable AI **Matthias Hein** (University of Tübingen, Germany): Adversarial Robustness - the good and the ugly **Sinem Sav** (EPFL Lausanne, Switzerland): Privacy-preserving federated neural network learning
10:45-11:00	**Coffee Break**	
11:00-12:30	**Session 2: Auditing AI Systems in Practice**	**Vince Istvan Madai** (Charité - Universitätsmedizin Berlin, Germany): To explain or not to explain? XAI in healthcare **Ken Cassar** (UMNAI): Inherently Auditable Neuro-Symbolic AI **Luis Oala** (Fraunhofer HHI, Berlin, Germany): Best Practices and Toolboxes for Auditing AI Systems in Healthcare
12:30-13:30	**Lunch**	
13:30-15:30	**Session 3: General Aspects, Standardization and Regulation**	**Edson Prestes** (Federal University of Rio Grande do Sul, Brazil): IEEE 7007 - Ontological Standard for Ethically Driven Robotics and Automation Systems **Julien Chiaroni** (Directeur du Grand Défi, SGPI, France): The French approach: How to operationalize the European AI Act **Dirk Schlesinger** (TÜV AI Lab, Berlin, Germany): German Standardization Roadmap AI - Working Group Medicine **Martin R.** (National Cyber Security Center, Manchester, UK): Developing principles for the security of Machine Learning systems
15:30-15:45	**Coffee Break**	
15:45-16:30	**Panel Discussion & Closing Remarks**	Do we really need XAI for trustworthy AI applications in healthcare?

Contact: Dr. Arndt von Twickel (arndt.twickel (at) bsi.bund.de), Dr. Patrick Gilroy (patrick.gilroy (at) tuev-verband.de) and Prof. Dr. Wojciech Samek (wojciech.samek (at) hhi.fraunhofer.de)

Federal Office for Information Security

TÜV VERBAND

Fraunhofer Heinrich-Hertz-Institut

Abstracts

Wojciech Samek: Concept-Level Explainable AI

The emerging field of Explainable AI (XAI) aims to bring transparency to today's powerful but opaque deep learning models. This talk will present Concept Relevance Propagation (CRP), a next-generation XAI technique which explains individual predictions in terms of localized and human-understandable concepts. Other than the related state-of-the-art, CRP not only identifies the relevant input dimensions (e.g., pixels in an image) but also provides deep insights into the model's representation and the reasoning process. This makes CRP a perfect tool for human-machine interaction. In the talk we will demonstrate on multiple datasets, model architectures and application domains, that CRP-based analyses allow one to (1) gain insights into the representation and composition of concepts in the model as well as quantitatively investigate their role in prediction, (2) identify and counteract Clever Hans filters focusing on spurious correlations in the data, and (3) analyze whole concept subspaces and their contributions to fine-grained decision making. By lifting XAI to the concept level, CRP opens up a new way to analyze, debug and interact with ML models, which is of particular interest in safety-critical applications and the sciences.

Matthias Hein: Adversarial Robustness - the good and the ugly

The European AI act indicates that a certain level of adversarial robustness is required when using machine learning for safety-critical applications. In this talk I will discuss the difficulty of the evaluation of adverarial robustness and why sometimes even machine learning researchers fail. I briefly discuss our framework AutoAttack for a standardized evaluation of adversarial robustness as a partial solution to the problem. In the second part I will show that adversarial robustness is also useful for explaining and debugging of image classifiers and discuss our recent Diffusion Visual Counterfactual Explanations.

Sinem Sav: Privacy-preserving federated neural network learning

Training accurate and robust machine learning models requires a large amount of data that is usually scattered across data silos. Sharing or centralizing the data of different healthcare institutions is, however, unfeasible or prohibitively difficult due to privacy regulations.

We address the problem of privacy-preserving training and evaluation of neural networks in an N-party, federated learning setting. In this talk, I will present POSEIDON, the first of its kind in the regime of privacy-preserving neural network training. It employs multiparty lattice-based cryptography to preserve the confidentiality of the training data, the model, and the evaluation data, under a passive-adversary model and collusions between up to N−1 parties. To efficiently execute the secure backpropagation algorithm for training neural networks, we provide a generic packing approach that enables Single Instruction, Multiple Data (SIMD) operations on encrypted data. Our experimental results show that POSEIDON achieves accuracy similar to centralized or decentralized non-private approaches and that its computation and communication overhead scales linearly with the number of parties.

We improve POSEIDON and demonstrate its applicability on biomedical analysis for disease-associated cell classification with single-cell analysis. For this, we design a system, PriCell, for training a published state-of-the-art convolutional neural network in a decentralized and privacy-preserving manner. We compare the accuracy achieved by PriCell with the centralized and non-secure solutions and show that PriCell guarantees privacy without reducing the utility of the data.

Vince Istvan Madai: To explain or not to explain? XAI in healthcare

Explainability for artificial intelligence (AI) in medicine is a hotly debated topic. In my talk, I will present the main arguments for and against explainability for AI-based clinical decision support systems (CDSS). I will explore how technical, ethical and stakeholder considerations influence the impact and design of explainability implementations, and will illustrate the different approaches with a concrete use case.

Ken Cassar: Inherently Auditable Neuro-Symbolic AI

Recent advances in Neuro-Symbolic AI bring transformative and fundamental changes to how AI systems are audited. This talk will introduce UMNAI's Hybrid intelligence and look at inherent global and local auditability and explore some of the impacts and opportunities that lie beyond the current frontier of opaque AI algorithms.

Luis Oala: Best Practices and Toolboxes for Auditing AI Systems in Healthcare

Developers proposing new machine learning for health (ML4H) tools often pledge to match or even surpass the performance of existing tools, yet the reality is usually more complicated. Reliable deployment of ML4H to the real world is challenging as examples from diabetic retinopathy or Covid-19 screening show. Together with my collaborators, I envision an integrated framework of algorithm auditing and quality control that provides a path towards the effective and reliable application of ML systems in healthcare. In this presentation, I give a summary of ongoing work towards that vision and share examples from the practice of ML4H auditing.

Edson Prestes: IEEE 7007 - Ontological Standard for Ethically Driven Robotics and Automation Systems

In this talk, I will present the IEEE 7007 Ontological Standard for Ethically Driven Robotics and Automation System. This standard was published in the end of 2021 and is the very first ontological representation for the ethics and AI domain. It contains a formal representation for several concepts and relationships that are necessary to express information related to, e.g., norms and ethical principles, data privacy and protection, transparency and accountability.

IEEE 7007 Standard can contribute to the advance of the use of Ethics in the design and deployment of Autonomous and Intelligent Systems in multiple ways. It can be used, e.g., as a guide to teach ethical design for both human and institutional capacity building; as a framework for creating computational ethically aligned systems; as a mechanism to strengthen cooperation across States, and so on.

Due to this achievement, the Working Group that developed the IEEE 7007 Standard was a recipient of the International IEEE SA Emerging Technology Award in 2021.

Julien Chiaroni: The French approach: How to operationalize the European AI Act

The "Grand Défi on trustworthy AI for industry" aims to ensure user trust and promote the wide deployment of AI in strategic industries in Europe. To achieve these goals, it supports the development of standards and technological solutions that support design, test, validation, verification, and maintainability of AI-based systems. Three pillars are strategic in order to achieve these goals. First of all, we must revisit "classic" engineering (datas engineering, algorithmic engineering, software engineering and system engineering) to ensure the system's compliance with requirements and constraints, especially the future European regulation on AI (AI Act). Then, we have to develop new

conformity assessment schemes and new technical infrastructures (especially with simulation). Finally, we have to promote new standards that support European value and regulation. The presentation will summarize the achievement of the “Grand Défi” and the challenges that we still need to tackle.

Dirk Schlesinger: German Standardization Roadmap AI - Working Group Medicine

The 2nd version of the ‘Normungsroadmap KI’ will be officially released in early December 2022. The short presentation will provide a glimpse of the recommendations put forward by the ‘working group medicine’, especially where the group sees additional need for research and harmonization with existing norms and standards so as to provide a robust framework for conformity assessments for AI in medical devices.

Martin R.: Developing principles for the security of Machine Learning systems

In this talk I will present the NCSC's principles for the security of machine learning systems, which were published in September 2022. I will discuss our reasons for producing the principles and summarise the approach we took to developing them. I will also briefly discuss how we have tested the principles, in particular as part of a small project to retrospectively apply them to a machine learning implementation in the healthcare sector.

Machine learning has become widespread in almost every aspect of life, and it became apparent that there was a lack of practical guidance on securing systems against the specific threats inherent to the ML lifecycle. Much of the work on the security of machine learning is within an academic or research context and pragmatic advice for non-specialists has become increasingly necessary. The NCSC has therefore worked to develop a set of principles which are designed to be applicable to most machine learning implementations, being agnostic to function, model and data type. The principles are not a comprehensive assurance framework or a checklist. Instead they provide context and structure to help decision makers and risk owners make educated decisions about system design and development, considering their own specific threat environment.