SAIL
SUSTAINABLE LIFE-CYCLE OF INTELLIGENT SOCIO-TECHNICAL SYSTEMS

Spring School 2025

Topic

Innovating AI Evaluation: Beyond Accuracy and Precision.

Description

As the landscape of artificial intelligence continues to evolve, the need for comprehensive and nuanced evaluation methods increases as well. Traditional metrics such as accuracy and precision, while important, are insu cient for fully capturing the complexities and impacts of AI systems.
The SAIL Spring School aims to address this gap by introducing participants to a diverse array of evaluation strategies, such as user evaluations, ethical and societal impacts, evaluating outcomes that are co-constructed between user and AI, mathematical guarantees, interpretability and transparency assessments, context-specific metrics, etc.

When & Where

When: March 26-28, 2025 | **Where**: CITEC lecture hall, Bielefeld University, Germany

Demystifying Explainable LLMs

Peggy Lindner
University of Houston

Amaury Lendasse
Missouri University of Science and Technology

As Large Language Models (LLMs) become integral to AI-driven systems, ensuring their outputs are interpretable and trustworthy is crucial. This session explores key strategies for improving LLM explainability, including interpretable outputs, attention visualization, feature attribution techniques, and counterfactual reasoning. We will also discuss human-in-the-loop approaches, rule-based integration, and structured retrieval methods that enhance transparency and user trust. Special focus will be given to the role of explainability in agentic AI systems, where models operate with greater autonomy and influence. Attendees will gain practical insights into implementing explainability techniques for more reliable and accountable AI applications.

Funded by

Ministry of Culture and Science of the State of North Rhine-Westphalia

UNIVERSITÄT BIELEFELD
UNIVERSITÄT PADERBORN
HSBI Hochschule Bielefeld University of Applied Sciences and Arts
TH OWL

sail.nrw