Trustworthiness and explainability

Barbara Hammer

Bielefeld University

10 septembre 2026

Explainable AI (XAI) aims at the integration, extension, or substitution of otherwise black-box machine learning models by components, which are comprehensible for humans. Albeit many XAI methods exist, it is unclear, what information exactly they reveal, why different methods lead to different results, and in which sense humans can trust the results. Within the talk, I will give examples how XAI technologies can be used in applications relating to critical infrastructures, and which challenges such as uniqueness, plausibility, or vulnerability exist. Further, I will present first results how to target some of these challenges : Specifically, I will have a closer look at feature-based explanations and how to extend those to incorporate the effect of interactions, leading to more reliable results, for explaining multimodal foundation models, for example.