Contribution ID: **58** Type: **not specified**

Benchmarking Trust: A Metric for Trustworthy Machine

Thursday, 16 May 2024 15:05 (25 minutes)

In the evolving landscape of machine learning research, theconcept of trustworthiness receives critical consideration, both concern-ing data and models. However, the lack of a universally agreed upondefinition of the very concept of trustworthiness presents a considerablechallenge. The lack of such a definition impedes meaningful exchange andcomparison of results when it comes to assessing trust. To make mattersworse, coming up with a quantifiable metric is currently hardly possible.In consequence, the machine learning community cannot operationalizethe term, beyond its current state as a hardly graspable concept.
In this talk, a first step towards such an operationalization of the notion of is presented –The FRIES Trust Score, a novel metric designed to evaluate the trustworthiness of machine learning models and datasets. Grounded in five foundational pillars –fairness, robustness, integrity, explainability, and safety –this approach provides a holistic framework for trust assessment based on quality assurance methods. This talk further aims to shed light on the critical importance of trustworthiness in machine learning and showcases the potential of the implementation of a human-in-the-loop trust score to facilitate objective evaluations in the dynamic and interdisciplinary field of trustworthy AI.

Type of presentation

Invited Talk

Primary author: RUTINOWSKI, Jérôme (TU Dortmund University)

Presenter: RUTINOWSKI, Jérôme (TU Dortmund University)

Session Classification: Invited session

Track Classification: Spring Meeting