Introduction to the Trustworthy Artificial Intelligence and Machine Learning Minitrack

Line Pouchard
Brookhaven National Laboratory
pouchard@bnl.gov

Peter Salholer
FH JOANNEUM
peter.salhofer@fh-joanneum.at

With the advancement of AI technology, AI algorithms start to match human performance for certain tasks (e.g. ChatGPT) and discover loopholes in systems that were not previously found. AI in general and ML methods specifically are increasingly used with scientific data and applied with great promise to solve a broad variety of scientific problems. With the increased use of AI comes an increase in inherent complexity. Deep Learning (DL) models with billions of parameters, operating with very large data volumes on heterogeneous architectures, obscure their inner workings to human understanding. Unlike traditional ML algorithms, such as rule-based decision trees or linear-regression models where the decision boundary is clear, interpreting a learned model is difficult.

The increased need for transparency is compounded by that of avoiding bias in predictions. Numerous examples of bias have been discovered in image recognition, classification, and text generation. Formal explanations of how models achieve results, explicit representations of data, comprehensiveness and diversity of datasets used for training are crucial to foster trust in AI. Additionally, the accuracy of results obtained with AI is often the product of customization; experiments show that many are not reproducible at scale, even within expected error bounds. While reproducibility may not be needed for some uses of AI (e.g. when AI is used for the purpose of preliminary triage in drug discovery) in other uses, reproducible AI is paramount.

Researchers need to understand how Artificial Intelligence (AI) and Machine Learning (ML) results are obtained in order to gain new insights and to establish confidence in the validity of these results. The promises of AI/ML will not be realized if scientists cannot trust the results, understand how they were obtained, gain transparency into what datasets, models and model parameters have been used or what features in the data lead to results. Like any good scientific results, AI/ML pipelines should be reproducible to the most possible extent.

This minitrack explores a themes related to trustworthy AI. The only paper accepted to this minitrack is titled "Towards a Quantitative Evaluation Framework for Trustworthy AI in Facial Analysis" authored by Annika Schreiner and Nils Kemmerzell. A framework for quantifying, integrating, and evaluating several aspects of trustworthy analysis in the context of facial analysis systems (ResNet), including fairness, robustness, privacy, and explainability was presented.

URI: https://hdl.handle.net/10125/107325
978-0-9981331-7-1