Title: Discussion of the possibilities of implementing selected guidelines for trustworthy artificial intelligence

[polska wersja tytułu : Możliwości wdrożenia wybranych wytycznych dla godnej zaufania sztucznej inteligencji.]

Jerzy Stefanowski
Institute of Computing Science
Poznan University of Technology
60-965 Poznan
email: jerzy.stefanowski@cs.put.poznan.pl

Abstract:
In recent years, Artificial Intelligence (AI) has made a tremendous progress in terms of new methods, tools and the applications, that have a wide impact on society and humans. Despite the positive benefits from it, there is also a growing awareness of AI limitations and risks, which restricts the people's trust in the current systems. The ideas of Responsible AI or Trustworthy AI are introduced to mitigate them. Following them several criteria for AI systems are formulated, see e.g. the list of ethical guidelines prepared by EU Commission High Level Expert Group on Artificial Intelligence. Although some of them inspire the current works on the regulation proposals, many of them are formulated on a too general level or associated with too high expectations when compared to the current possibilities of their practical implementation.
This leads to many open research questions. Based on own research, we will consider here three selected aspects of Trustworthy AI: explanations of predictions from machine learning (ML) models, fairness of proposed decisions and limited reproducibility of the current ML.
In general explanation methods attempt to make interpretable the internal logic and the outcome of the ML algorithms, in particular popular "black boxes" - neural network models, in understable terms to humans. Although many technical methods have already been introduced, several important problems are still open. In this presentation we will discuss different representation forms of explanations and their adjustment to the human recipient, difficulties of their evaluation, needs for multi-criteria analysis and ways of involving human experts in the interactive evaluation scenarios. It will be illustrated by two cases of explaining text classification with prototype deep network and different approaches exploiting counterfactuals. The other discussed aspect concerns fairness machine learning, i.e. we want to prevent situations, where decisions of intelligent systems take unexpected social implications, such as discriminations with respect to sensitive person characteristics (e.g. gender, ethnicity, people disabilities, …). We will focus on inability to satisfy simultaneously main fairness measures and illustrate it by the case of an online learning prediction system.
Finally, we will discuss the challenge of the difficulty of reproducibility of the most popular machine learning algorithms using deep neural networks. We briefly summarize a few survey analyses of irreproducibility of many recent ML research papers. Unfortunately, this raises many consequences and even doubts as to the methodology of conducting research in this field.

Keywords: Explainable Artificial Intelligence; Prototypes and counterfactual explanations, Fairness in Machine Learning, Reproducibility

Wybrane piśmienictwo:
1. Maciej Falbogowski, Jerzy Stefanowski, Zuzanna Trafas, Adam Wojciechowski: The Impact of Using Constraints on Counterfactual Explanations. W Proceedings of the 3rd Polish Conference on Artificial Intelligence, April 25-27, 2022, Gdynia, Poland, 81 - 84 (2022).
2. Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, Dino Pedreschi: A survey of methods for explaining black box models. ACM Computing Surveys (CSUR) 51(5), 1–42 (2018).
3. Odd Erik Gundersen, Yolanda Gil, and David W. Aha: On Reproducible AI: Towards Reproducible Research, Open Science, and Digital Scholarship in AI Publications. AI Magazine 3,56–68, (Sep.2018).
4. Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. ACM Comput. Surv. 54, 6, Article 115 (July 2022),
5. Kamil Pluciński, Mateusz Lango, Jerzy Stefanowski: Prototypical Convolutional Neural Network for a Phrase-Based Explanation of Sentiment Classification. W Proceedings of the Machine Learning and Principles and Practice of Knowledge Discovery in Databases ECML PKDD 2021 Conferences. Communications in Computer and Information Science, vol 1524. Springer, 457-472 (2021).

---------- uwagi dla organizatorów

Prezentacja – referat, w przypadku przyjęcia, wygłoszę w formie stacjonarnej (w sali konferencji); Język prezentacji – polski lub angielski – dowolny, dostosuję się do specyfiki organizacji konkretnej sesji.

Informacja o autorze:
Jerzy Stefanowski is a Professor at the Institute of Computing Science, Poznan University of Technology. He received the Ph.D. and Habilitation degrees in computer science from this university. He is also a corresponding member of Polish Academy of Sciences and a vice-president of Polish AI Society. His research interests include machine learning, data mining and intelligent decision support, in particular ensemble classifiers, class imbalance, rule induction, and explainable Artificial Intelligence. More information could be found at https://www.cs.put.poznan.pl/jstefanowski/