Meta-analysis of academic discourse about interpretability, transparency, and fairness

Stanisław Giziński²,¹, Michał Kuźba²,¹, Przemysław Biecek¹,²
¹Faculty of Mathematics and Information Science, Warsaw University of Technology
²Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw

Introduction
Artificial intelligence models have an increasing impact on our lives. But instead of “in the service of society” we see growing concerns that such models may strengthen some existing inequalities and biases. These, sometimes spectacular, failures have a negative influence, can harm trust, stop funding and research progress and even lead to the next AI winter. In recent years, we observe a growing discourse in academia and high-tech companies on how to prevent such outcomes. Some approaches to detect and mitigate these risks include models interpretability, transparency guidelines, and fairness. In order to have a better understanding of the growing number of discussions in this area, we have created an automated pipeline for the analysis of research papers related to such topics. We automatically track trends and stakeholders in the discourse on Responsible AI. In this work we present preliminary results from our meta-analysis.

Database of regulations
We curate a comprehensive database of AI regulations across all countries, organizations and areas of applications. It can be found on https://github.com/ModelOriented/MAIR.

Data gathering and preprocessing
We downloaded over 500 papers from arXiv related to Explainable AI, Interpretable Machine Learning, fairness and transparency. We use arXiv and Semantic Scholar APIs to download LaTeX sources and some meta-data. We are particularly interested in modelling the discourse on Responsible AI and understanding actors (stakeholders) and links between them and their work. In order to do that, we construct a directed graph with nodes corresponding to papers from the dataset and edges for citations. We are particularly interested in understanding who the stakeholders are. We perform an analysis of affiliations declared by authors of the papers. We extract them by parsing LaTeX sources and performing Named Entity Recognition using [2]. Then, we label papers by the type of their authors affiliations – academic (e.g. University of Warsaw), business – (e.g. Google) or both. Our preliminary results show that the discourse is not dominated by any organization and around 80% of papers include an academic affiliation. The view on the graph is presented on Figure 1.

Additional analysis
In addition to previous analysis, we analysed S2ORC corpus [3]. After filtering computer science category papers, we lemmatized all abstract using spacy lemmatizer [2] from computer science category, and picked papers from the field of machine learning by searching for specific noun phrases. Final corpus contained around 0.5M papers related to AI. We then checked for keywords[1], to explore how some topics change in time.

References

Acknowledgements
This work was financially supported by NCN grant 2019/34/E/ST6/00052

✉ kuzba.michal@gmail.com
✉ s.gizinski84@gmail.com
✉ przemyslaw.biecek@gmail.com

Figure 1: A graph representation of arXiv dataset of papers related to interpretability, transparency and fairness. A single node represents a paper and edge (directed) a citation. The size of a node corresponds to the citation count and colours to the type of affiliations: academic (blue), business (red), academic and business (green), unknown affiliation (grey).

Figure 2: Interpretability, transparency, and fairness related topic in papers across time.