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

Stanisław Giziński[2, 1], Michał Kuźba[2, 1], Przemysław Biecek[1, 2]

[1]Faculty of Mathematics and Information Science, Warsaw University of Technology

[2]Faculty of Mathematics, Informatics, and Mechanics, University of Warsaw

## Introduction

Artificial intelligence models have an increasing impact on our lives. However, 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 observed 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. 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 application. 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 metadata. We are particularly interested in modelling the discourse on Responsible AI and understanding the actors (stakeholders) and the links between them and their work. 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 the affiliations declared by authors of the papers. We extract them by parsing LaTex sources and performing Named Entity Recognition using spaCy[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 organisation and around 80% of papers include an academic affiliation. The view of the graph is presented in Figure 1.

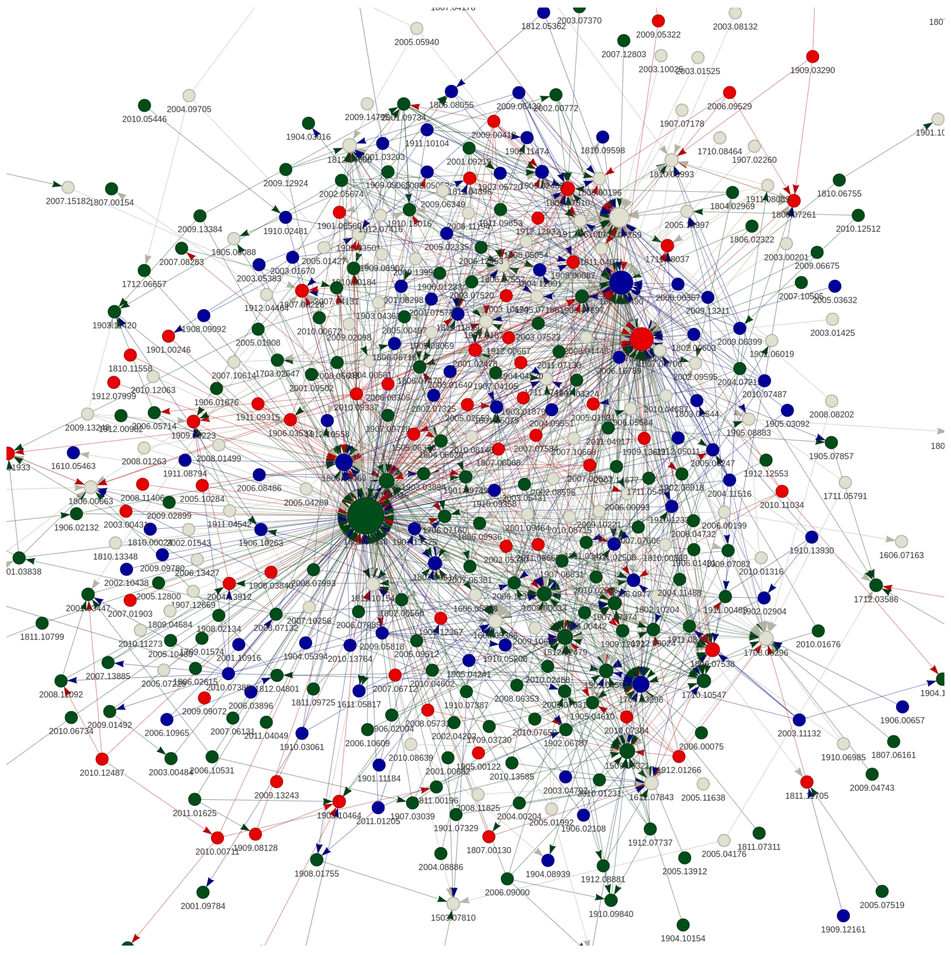

Figure 1: A graph representation of arXiv dataset of papers related to interpretability, transparency and fairness. A single node represents a paper and edge (directed) a citation. The size of a node corresponds to the citation count and colours to the type of affiliations: academic (blue), business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

## Additional analysis

In addition to the previous analysis, we analysed S2ORC corpus [3]. After filtering computer science category papers, we lemmatized all abstracts 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.

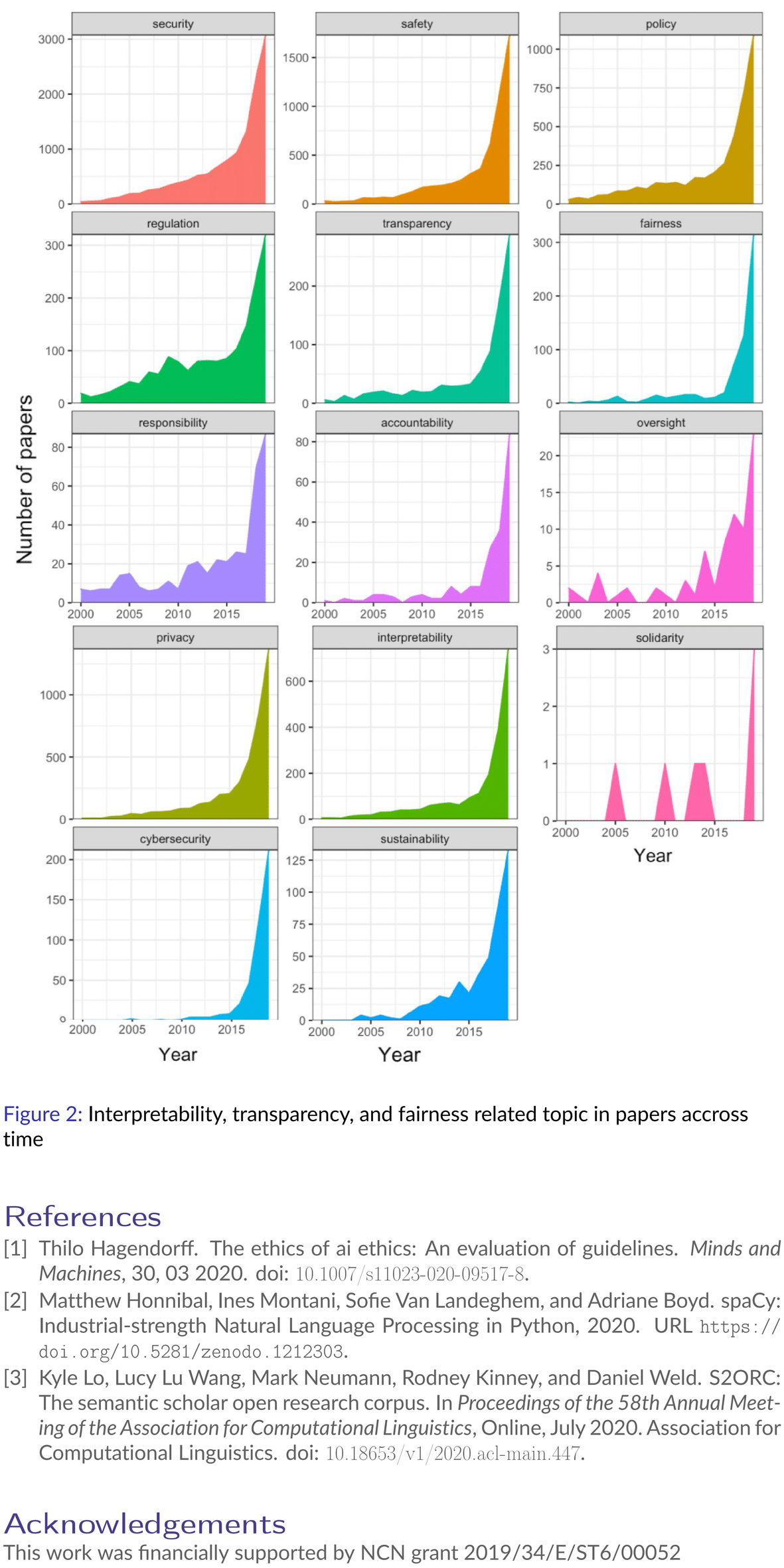


Figure 2: Interpretability, transparency, and fairness related topic in papers accross time

## References

[1] Thilo Hagendorff. The ethics of ai ethics: An evaluation of guidelines. *Minds and Machines*, 30, 03 2020. doi: 10.1007/s11023-020-09517-8.

[2] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020. URL `https://doi.org/10.5281/zenodo.1212303`.

[3] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.447.

## 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`