Exploring ethics discourses in practice in AI research.

KAVLI CENTRE for Ethics, Science, and the Public

UNIVERSITY OF CAMBRIDGE Faculty of Education · Isaac Newton Trust · Cambridge Public Health

Dr Daniela Boraschi

www.kcesp.ac.uk db889@cam.ac.uk

Kavli Centre for Ethics, Science, and the Public
Faculty of Education, University of Cambridge

INTRODUCTION

Artificial Intelligence (AI) is a broad term. It encompasses the evolving scientific field focused on creating human-like intelligence in future technologies through capabilities like 'learning', 'prediction' and 'decision-making'. It also involves present-day applications using machine learning (ML) and deep learning (DL) for tasks in various fields like healthcare, finance, policing, and education. In healthcare, it is hoped that AI will help to predict and detect diseases like cancer and dementia earlier, and optimise the distribution of resources via personalised profiles. However, a growing interdisciplinary discourse is cautioning against overly optimistic views of AI, emphasising the importance of considering the negative implications of this technology, while inviting AI scientists to reflect on the socio-ethical impacts of their work. This poster presents an analysis of how these reflections are being incorporated into AI research with a focus on two important conferences in the field of AI ethics, machine learning and computational neuroscience: *ACM FAccT* & *NeurIPS*.

QUESTION

How are reflections on the socio-ethical implications of AI being incorporated into research?

METHODS

This analysis uses a mixed-method approach combining quantitative **lexicon extraction** and **sentiment analysis** with qualitative **critical discourse analysis** to explore two large datasets.

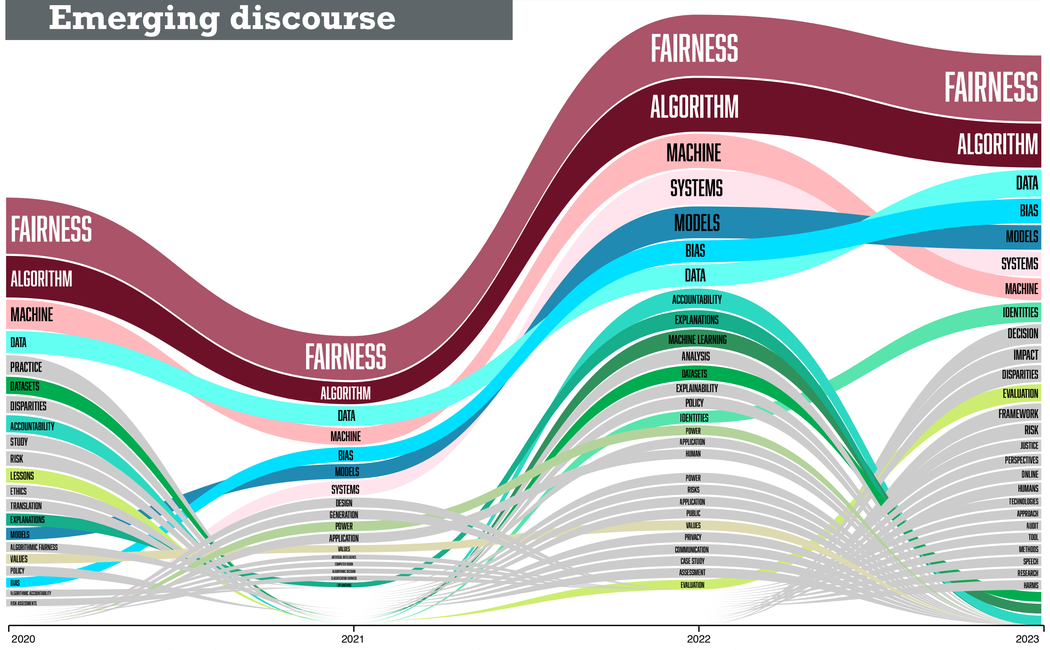
- **Dataset 1:** abstracts (n=481) of papers presented to the *ACM FAccT* from 2020 to 2023.
- **Dataset 2:** socio-ethical impact statements (n=1898) from *NeurIPS 2020* papers.
- **Dataset 3:** socio-ethical impact statements on AI and health (n=9) from *NeurIPS 2020* papers.

The analysis is conducted with two open-source and web-based software *CORTEXT* and *Raw Graphs*, and the qualitative software *Atlas.ti*. It uses:

- **Lexicon extraction** to identify prevalent socio-ethical concerns by counting the most frequent words in the dataset.
- **Sentiment analysis** to capture the attitudes in relation to socio-ethical concerns by scoring the degree of subjectivity and positive/negative emotions in the language.
- **CDA - Thematic analysis** to reveal how scientists are articulating socio-ethical concerns in terms of strategies, meanings, values and power in relations to wider contexts.

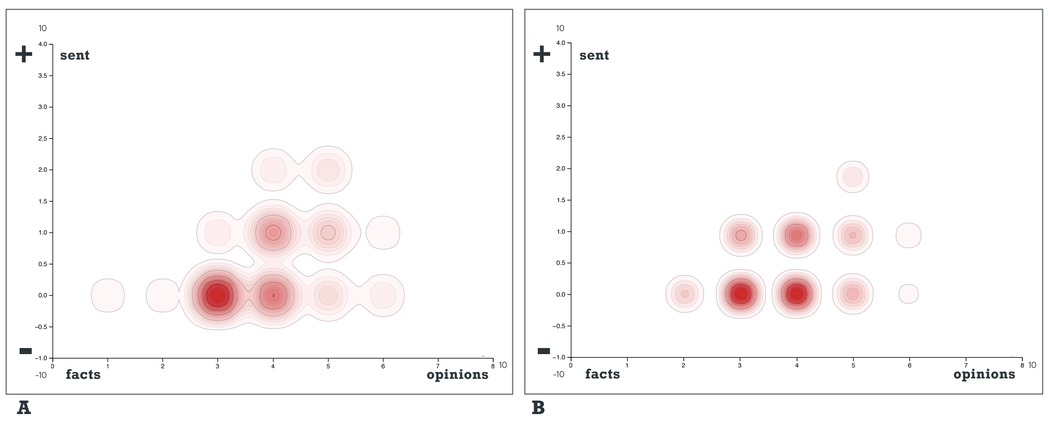
PRELIMINARY RESULTS

Emerging discourse

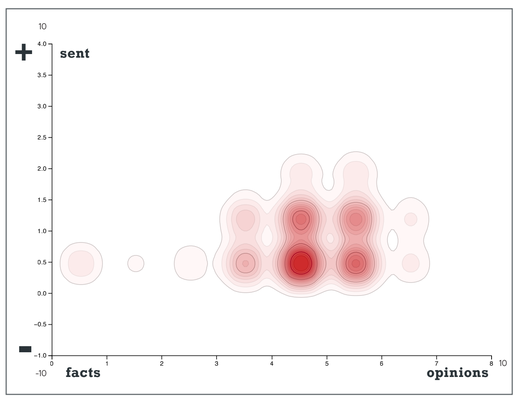


Changes in high-frequency words in *ACM FAccT* paper titles (n=481) from 2020 to 2023 show the evolution of socio-ethical themes. 'Fairness' consistently maintains a high-frequency, while other topics peak and fluctuate over time. Additionally, color-coded bands indicate less frequent but consistent topics.

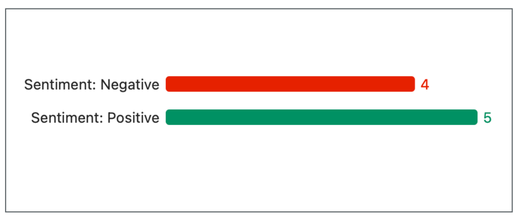
Scientists' attitudes



Sentiment extracted from abstracts (n=481) of papers presented at *ACM FAccT* in 2020 (image A) and 2023 (image B). The overall attitude is fairly neutral, with an increasing trend towards the use of subjective and positive language.



Sentiment extracted from socio-ethical impact statements (n=1898) from *NeurIPS* 2020 papers. Despite the requirement to address the potential negative socio-ethical implications of AI, overall, the reflections are fairly neutral, showing a small trends towards the more frequent use of positive and subjective language.



Sentiment analysis of socio-ethical impact statements (n=9) from *NeurIPS* 2020 papers by scientists working in a UK-based lab on AI for healthcare. Even though positive sentiment is prevalent, scientists are considering the potential negative implications in their reflections.

Three key themes

Ethics and methods

Ethics of datasets creation.

Ethics of AI testing.

Generalisability, applicability and safety of outcomes derived from AI models (ML and DL) trained with synthetic and/or sculpted datasets.

Power and values

Dominant framing of AI-based mathematical solutions to complex problems.

Quantification, efficiency, and probability influence the interpretation and application of values such as fairness.

Interdisciplinary partnerships vs partnerships driven by mathematical skills, values and agendas.

Risk and morality

Risk of harm from incorrect estimations and/or intentional human harm.

Who decides the type and level of acceptable risk, and how?

TALKING POINTS

This exploratory research, in its current stage, contributes to both sociological and philosophical literature on AI ethics, as well as to more technical approaches to ethical AI R&D with the three following points of consideration.

A dominant scientific discourse is evolving around AI *fairness*. 'Fairness' is broadly defined as a value that involves treating individuals equally, in a just and reasonable manner. Within this discourse, fairness has acquired a dual role. It operates as a universally positive value guiding efforts to develop equitable AI systems. It is also a mathematical intervention at the dataset level through which distributions and predictions among different groups or individuals can be manually adjusted in order to achieve equal distributions of resources.

While acknowledging the importance of commitment and efforts, this research emphasises an inherent tension within this discourse, most notably in healthcare. Human moral values, such as fairness, are in tensions with efforts to operationalise them through mathematical or probabilistic tools and frameworks. This quantitative transformation deprives values, such as fairness, of their intrinsic human or societal essence, such a difference, altruism or kindness. This matters because finding the best (or optimal or most probable) outcome for societies, populations, and individuals risks the unintentional discrimination against those who fall outside the selected criteria for equal distributions of resources.

NEXT STEPS

- Empirically grounded research to examine further emerging themes in healthcare.
- Collaborations with scientists to support reflections of socio-ethical implications of AI R&D.

References

Ashurst, C., Hine, E., Sedille, P., & Carlier, A. (2021). AI Ethics Statements -- Analysis and lessons learnt from NeurIPS Broader Impact Statements. *arXiv.org*. https://doi.org/10.48550/arxiv.2111.01705

Bærøe, K., Gundersen, T., Henden, E., & Rommetveit, K. (n.d.). Can medical algorithms be fair? Three ethical quandaries and one dilemma. *BMJ Health & Care Informatics*. Volume 29:Issue 1 (2022).

Eustasio del Barrio, Gordaliza, P., & Jean-Michel Loubes. (2020). Review of Mathematical frameworks for Fairness in Machine Learning. *arXiv.org*. https://doi.org/10.48550/arxiv.2005.13755

Hacking, I. (2018). Risk and Dirt. In *Risk and Morality* (pp. 22–47). Toronto: University of Toronto Press. https://doi.org/10.3138/9781442679382-004

Hacking, I. (1983). *Representing and Intervening*. Cambridge University Press. https://doi.org/10.1017/CBO9780511814563

Acknowledgments

I would like to express my sincere appreciation to my whole team at the Kavli Centre for Ethics, Science, and the Public for their invaluable support throughout the research. Special thanks to Dr Richard Milne and Prof Anna Middleton for their mentorship and expertise, and to Prof Mihaela van der Schaar for contributing in the research with enthusiasm. I also acknowledge the financial support provided by the Kavli Foundation and the Isaac Newton Trust to facilitate this ongoing post-doctoral research.