AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

ROADMAP ON AI TECHNOLOGIES & APPLICATIONS FOR THE MEDIA INDUSTRY

SECTION: "ETHICAL AND LEGAL ASPECTS OF AVAILABILITY OF QUALITY DATA FOR AI RESEARCH"

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu

Authors	**Lidia Dutkiewicz** (KU Leuven Centre for IT & IP Law) **Noémie Krack** (KU Leuven Centre for IT & IP Law) **Emine Ozge Yildirim** (KU Leuven Centre for IT & IP Law)

This report is part of the deliverable D2.3 - "*AI technologies and applications in media: State of Play, Foresight, and Research Directions*" of the AI4Media project.

You can site this report as follows:

F. Tsalakanidou et al., Deliverable 2.3 - AI technologies and applications in media: State of play, foresight, and research directions, AI4Media Project (Grant Agreement No 951911), 4 March 2022

This report was supported by European Union´s Horizon 2020 research and innovation programme under grant number 951911 - AI4Media (A European Excellence Centre for Media, Society and Democracy).

The information and views set out in this report are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf.

Ethical and legal aspects of availability of quality data for AI research

Current status

Considerable amounts of training and testing data are necessary for research and development of AI especially for Machine Learning techniques. The higher the quality of the training data, the better the system outputs will be. AI researchers are therefore looking for vast datasets in order to produce reliable and accurate outputs. For training AI models, data such as CommonCrawl[1], used for training large language models, ImageNet[2] for object recognition, or MS COCO[3] for computer vision tasks are employed. Not all data contained in these massive datasets constitute personal data and hence are not covered by the data protection legal framework; however, parts of them may contain personal data and even special categories of personal data.

The literature has been highlighting the issues of ***quality and representation*** in training data as part of publicly available datasets and databases[4]. Pornography, stereotypes, racist and other problematic contents were found to be part of datasets[5]. When used in AI research, the biased, incorrect training data will replicate or exacerbate these negative features in the AI system outputs[6]. This can be illustrated by the ***'garbage in, garbage out' principle***, opening the door to human rights violation and discrimination[7].

Data protection in research and its legal issues have been the subject of study by various scholars that we recommend reading[8,9,10,11,12,13]. Several challenges have been identified including:

- the lack of legal basis to process personal data in the publicly available datasets,
- the difficulty to comply with data subjects' rights,

[1] Common Crawl: https://commoncrawl.org/
[2] ImageNet: https://www.image-net.org/
[3] COCO dataset: https://cocodataset.org/#home
[4] Raji, Inioluwa Deborah, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. 'Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing'. In , 145–51. https://doi.org/10.1145/3375627.3375820.
[5] Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. 'Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes'. ArXiv:2110.01963 [Cs], October. http://arxiv.org/abs/2110.01963.
[6] Osoba, Osonde A, and William Welser IV. 2017. 'An Intelligence in Our Image: The Risks of Bias and Errors in Artificial Intelligence'. Rand Corporation. https://www.rand.org/content/dam/rand/pubs/research_reports/RR1700/RR1744/RAND_RR1744.pdf
[7] Richardson, Rashida, Jason Schultz, and Kate Crawford. 2019. 'Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice'. New York University Law Review 94: 42.
[8] Ducato, Rossana. 2020. 'Data Protection, Scientific Research, and the Role of Information'. Computer Law & Security Review 37: 105412.
[9] Jasserand, Catherine. 2018. 'Massive Facial Databases and the GDPR: The New Data Protection Rules Applicable to Research'. In Data Protection and Privacy: The Internet of Bodies, 169–88. Hart Publishing/Bloomsbury Publishing Plc.
[10] Weller, Katrin, and Katharina E Kinder-Kurlanda. 2016. 'A Manifesto for Data Sharing in Social Media Research'. In , 166–72.
[11] Barfield, Woodrow, and Ugo Pagallo. 2018. Research Handbook on the Law of Artificial Intelligence. Edward Elgar Publishing.
[12] Barocas, S., and A. Selbst. 2016. 'Big Data's Disparate Impact'. California Law Review 104 (671).
[13] De Bruyne, Jan, and Cedric Vanleenhove, eds. 2021. Artificial Intelligence and the Law. Intersentia.

- the difficulty to comply with technical and organisational safeguards to process data in a data protection-compliant way.

In addition to the accuracy problems linked with low quality datasets, ethical considerations arise including research quality, legitimation of the use of biased datasets. This will not only harm the quality of the research outputs, but also its further use and the more general academic reputation. Some may remember the MegaFace[14] database. Megaface was a publicly available and widely used benchmark datasets in AI research set up by the University of Washington. Thanks to researchers, journalists and activists, the alarm was triggered regarding the issues present in these datasets in respect to the right to privacy and other human rights[9]. However, repeated scandals about training data available and AI research will hinder public trust in science.

Given the above elements, improving access to relevant and sufficient data for AI research is needed more than ever. Complementary initiatives contributing to the fight against problematic data proliferation and use are equally needed. A recent study[15] put forward measures which would improve the situation. This includes making ethically salient information about datasets clear and accessible, active stewardship of the data and its uses, employing ethics review procedures to promote responsible data uses, and the advance review of datasets and publications.

The scholar community is also willing to address these issues and many journals announced extra ethics checks on AI research to be published. We can name the Conference and Workshop on Neural Information Processing Systems (abbreviated as NeurIPS and formerly NIPS), where papers could be rejected due to ethical and legal doubts associated with the data used[16].

The situation is not easy for AI researchers as doing AI research means coping with the fundamental tension between data protection (including data minimisation) and the vast amount of data needed for meeting accuracy and quality requirements. Furthermore, the diverse and vague legal frameworks are not helping to provide clear guidance to AI research[17].

All of these challenges can constitute a barrier to AI research harming the research impact, the publishing opportunities and society's trust in academia and research in general. Trust is essential in scientific research.

Possible ways forward

Availability of quality data is a core part of European technology policy discussions. Already in 2018, the European Council acknowledged that '*high-quality data are essential for the*

[14] Megaface database: http://megaface.cs.washington.edu/

[15] Peng, Kenny, Arunesh Mathur, and Arvind Narayanan. 2021. 'Mitigating Dataset Harms Requires Stewardship: Lessons from 1000 Papers'. ArXiv Preprint ArXiv:2108.02922.

[16] Beygelzimer, Alina, Yann Dauphin, Percy Liang, and Jennifer Wortman Vaughan. 2021. 'Introducing the NeurIPS 2021 Paper Checklist'. Medium (blog). 26 March 2021. https://neuripsconf.medium.com/introducing-the-neurips-2021-paper-checklist-3220d6df500

[17] Rogers, Anna, Tim Baldwin, and Kobi Leins. 2021. 'Just What Do You Think You're Doing, Dave?'A Checklist for Responsible Data Use in NLP'. ArXiv Preprint ArXiv:2109.06598.

development of AI'[18]. A year later, the HLEG Ethics Guidelines for Trustworthy AI[19] referred in particular to ***data governance*** as one of the requirements for AI systems, highlighting the impact of biases in training datasets. In 2020, the European Commission released its White Paper on AI[20], which underlined how ***data availability*** was essential for training AI systems and how data quantity and quality were key components for building trustworthy and unbiased AI systems.

To meet the quality requirement, the EU policy instruments provide several measures aiming to solve the challenge relating to the lack of available data for the EU digital transformation:

- The ***European data strategy***[21] puts forward firstly EU-wide common, interoperable data spaces in strategic sectors as a solution to this lack of available data. Set up by the Commission, they would provide trustful, accountable and non-discriminatory access to high-quality data for the training, validation and testing of AI systems. Secondly, a horizontal governance framework for data access and use is said to '*facilitate decisions on which data can be used for scientific research purposes in a manner compliant with the GDPR*'.
- The ***Data Governance Act***[22] proposal puts forward a new interesting concept called *data altruism,* according to which data can be made available for purposes of 'general interest'.
- The ***EU AI Act***[23] proposal provides that for the development of high-risk AI systems, certain entities, such as digital innovation hubs, testing experimentation facilities and researchers, '*should be able to access and use high-quality datasets within their respective fields of activities*' (Recital 45).

Recent policy and legal initiatives show how importance on the availability of quality data has been acknowledged by policymakers. It remains to be seen how this will materialise tangibly for AI researchers. The research community is in need of effective tools and clear guidance to operate legally. Despite these initiatives, a wide binding data access framework for AI research would be more than welcome to enable researchers to access data in a harmonised and legally-compliant way.

---

[18] European Council. 2018. 'European Council Meeting of 28 June 2018 – Conclusions (EUCO 9/18)'. https://www.consilium.europa.eu/media/35936/28-euco-final-conclusions-en.pdf.

[19] High-Level Expert Group on Artificial Intelligence. 2019. 'Ethics Guidelines for Trustworthy AI'. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[20] European Commission. 2020. 'White Paper on Artificial Intelligence: A European Approach to Excellence and Trust (COM(2020) 65, Final)'. Brussels. https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificialintelligence-feb2020_en.pdf

[21] European Commission. 2020. Communication on a European Strategy for Data. https://digital-strategy.ec.europa.eu/en/policies/strategy-data

[22] European Commission. 2020. Data Governance Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020PC0767

[23] European Commission. 2020. Artificial Intelligence Act. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

AI4media

ARTIFICIAL INTELLIGENCE FOR THE MEDIA AND SOCIETY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951911

info@ai4media.eu www.ai4media.eu