LREC-COLING 2024

Legal and Ethical Issues in Human Language Technologies 2024 (LEGAL2024)

Workshop Proceedings

Editors
Ingo Siegert and Khalid Choukri

20 May, 2024
Torino, Italia

Proceedings of LEGAL2024: Workshop on Legal and Ethical Issues in Human Language Technologies @LREC-COLING-2024

ISBN 978-2-493814-21-0
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Preface

The year 2023 witnessed extensive discussions revolving around Artificial Intelligence (AI) and Large Language Models (LLM), marking a significant era in technological advancements. These discussions shed light on the unprecedented collection and utilization of data required by such technologies, often owned by stakeholders not directly involved in their development. Repackaging and repurposing these vast language datasets for AI and LLM endeavors become imperative, despite the intangible nature of language data, as they are subject to legal constraints.

In recent years, substantial efforts have been dedicated to adapting legal frameworks to technological advancements while considering the interests of diverse stakeholders. However, the strict consideration of legal aspects poses additional questions beyond mere recording technology and participant consent, constituting several key elements that warrant attention.

The LREC Workshop on Legal and Ethical Issues in Human Language Technologies 2024, held on Monday, June 20, 2024, as part of the LREC 2024 Conference, aims to delve into these crucial aspects. This workshop serves as a platform to explore the intricate interactions between legal and technical dimensions of data collection, processing, and distribution, particularly focusing on text crawling, speech and voice recordings, and the implications of text and speech data mining exceptions introduced by legislative bodies. Furthermore, it examines the compatibility of legal requirements for data collection and processing, as mandated by regulations like GDPR, alongside the technical feasibility of various anonymization and pseudonymization techniques.

A highlight of the workshop includes an invited talk by Jennifer Williams from the University of Southampton, offering insights into "AI Regulation Perspectives from the UK". This talk promises to enrich discussions by providing a comprehensive overview of AI regulation in the UK context. Furthermore, the workshop featured 11 accepted papers covering a range of topics. These included the use of LLMs in Finnish higher education, implications of regulations in the US super election year, intellectual property rights, annotating hate speech data, selling personal information, cultural heritage and data collection, and a comparison of voice user usage between Germany and Finland. These papers provided valuable insights into the legal and ethical challenges facing language technologies.

The workshop also delves into broader issues encompassing ethics, morality, and trust, exploring their interplay with data collection and distribution. It aims to foster dialogue between technology and legal experts, addressing current legal and ethical challenges in the Human Language Technology sector.

This volume encapsulates the proceedings of the LREC Workshop on Legal and Ethical Issues in Human Language Technologies 2024, documenting valuable insights and discussions shared during the event. We extend our gratitude to our keynote speakers and authors for their contributions, as well as to the Program Committee for their diligent review efforts.

Ingo Siegert, Khalid Choukri & Pawel Kamocki, April 2024

Organizing Committee

Ingo Siegert, Otto-von-Guericke-Universität Magdeburg, Germany
Khalid Choukri, ELRA/ELDA, France
Pawel Kamocki, IDS Mannheim, Germany
Kossay Talmoudi, ELDA, France

Table of Contents

LEGAL2024 Program

09:00 - 09:15	**Opening Session: Welcome by Workshop Chairs**
09:15 - 09:30	Participant and Organizer Introduction
09:30 - 10:30	Invited Talk: AI Regulation Perspectives from the UK (Jennifer Williams, University of Southampton)
10:30 - 11:00	**Coffee Break**
11:00 - 13:00	**Session I: Legal Frameworks and Ethical Considerations**
	Compliance by Design Methodologies in the Legal Governance Schemes of European Data Spaces *(Kossay Talmoudi, Khalid Choukri, and Isabelle Gavanon)*
	A Legal Framework for Natural Language Model Training in Portugal *(Ruben Almeida and Evelin Amorim)*
	Intellectual property rights at the training, development and generation stages of Large Language Models *(Christin Kirchhübel and Georgina Brown)*
	Ethical Issues in Language Resources and Language Technology – New Challenges, New Perspectives *(Pawel Kamocki and Andreas Witt)*
13:00 - 14:00	**Lunch Break**
14:00 - 16:00	**Session II: Considerations and Implications of AI**
	Legal and Ethical Considerations that Hinder the Use of LLMs in a Finnish Institution of Higher Education *(Mika Hämäläinen)*
	Implications of Regulations on Large Generative AI Models in the Super-Election Year and the Impact on Disinformation *(Vera Schmitt, Jakob Tesch, Eva Lopez, Tim Polzehl, Aljoscha Burchardt, Konstanze Neumann, Salar Mohtaj and Sebastian Möller)*
	Selling Personal Information: Data Brokers and the Limits of US Regulation (Denise DiPersio)
	What can I do with this data point? Towards modeling legal and ethical aspects of linguistic data collection and (re)use as a process *(Annett Jorschick, Paul T. Schrader and Hendrik Buschmeier)*
16:00 - 16:30	**Coffee Break**
16:30 - 17:30	**Session III: Applications and User Perspective**
	Data-Envelopes for Cultural Heritage: Going beyond Datasheets *(Maria Eskevich and Mrinalini Luthra)*
	Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data *(Maryam M. AlEmadi and Wajdi Zaghouani)*
	User Perspective on Anonymity in Voice Assistants – A comparison between Germany and Finland *(Ingo Siegert, Silas Rech, Matthias Haase and Tom Bäckström)*
17:30 - 18:00	**Wrap-Up of the Workshop and Closing Ceremony**

Compliance by Design Methodologies in the Legal Governance Schemes of European Data Spaces

Kossay Talmoudi, Khalid Choukri, Isabelle Gavanon
ELDA, DELCADE
9 Rue des Cordelières, 75013 Paris, 19 Rue du Colisée, 75008 Paris
{Kossay, Khalid}@elda.org, Igavanon@delcade.fr

Abstract

Creating novel ways of sharing data to boost the digital economy has been one of the growing priorities of the European Union. In order to realise a set of data-sharing modalities, the European Union funds several projects that aim to put in place Common Data Spaces. These infrastructures are set to be a catalyser for the data economy. However, many hurdles face their implementation. Legal compliance is still one of the major ambiguities of European Common Data Spaces and many initiatives intend to proactively integrate legal compliance schemes in the architecture of sectoral Data Spaces. The various initiatives must navigate a complex web of cross-cutting legal frameworks, including contract law, data protection, intellectual property, protection of trade secrets, competition law, European sovereignty, and cybersecurity obligations. As the conceptualisation of Data Spaces evolves and shows signs of differentiation from one sector to another, it is important to showcase the legal repercussions of the options of centralisation and decentralisation that can be observed in different Data Spaces. This paper will thus delve into their legal requirements and attempt to sketch out a stepping stone for understanding legal governance in data spaces.

Keywords: Data space, compliance, legal governance

1. Introduction

As the data economy is developing at an unprecedented pace, major regulatory changes have operated on the European level to maximise the value of data while regulating how many actors collect, use, and share it.

In this sense, the European Commission introduced a high-level European Data Strategy in February 2020 that introduced, among others, a landscape compiled of European Common Data Spaces that harness the data potential in various key sectors such as health, agriculture, and language.

Data Spaces are set to be one of the catalysers of data innovation. However, these Data Spaces require a set of legal standards that need to be taken into account as early as possible in the process of their conception, thus ensuring their sustainability and legal compliance in the long term.

Some of the major legal texts regulating data exchanges in the European Union are the Data Governance Act ("**DGA**") [1], the Data Act ("**DA**")[2], and the Open Data Directive[3]. These texts refer to the importance of complying with horizontal obligations stemming, among others, from the General Data Protection Regulation ("**GDPR**")[4], the E-Privacy Directive[5], the Copyright in the Digital Market Directive[6], and the proposed AI Act. Compliance with these texts requires a "by design" methodology that needs to guide the conceptualisation of Data Spaces as it is indispensable in the mission of fostering data potential responsibly.

Organisational modalities need to be put in place in order to unlock the data potential while respecting the obligations laid down in the different legal texts. There is thus an important need for a practice-facing methodology to allow dialogue between technicians building the Data Spaces and lawyers who are expected to proactively accompany the building of Data Space. Therefore, the legal obligations and frameworks governing the transfer of data need to be translated into actionable tools.

This paper tries to answer the question as to how a practical organisation of compliance be implemented in the framework of Common European Data Spaces. In fact, the data collected needs to be compliant among others, with the GDPR, trade secrets, and copyright law, among other obligations. In order to analyse this, it is important to identify the scope of the compliance with applicable law before identifying debtors of such warranties.

The legal obligations need to be laid down, not only in the documentation but also in the processes that allow for data transfers thus reflecting the operationalisation of these obligations and allowing for going beyond the "box-checking exercise" to rather dynamic compliance aligned with state-of-the-art technologies.

Legal rules would apply differently depending on the range of services offered by the Data Space and answers to operational questions are a prerequisite to the analysis. These questions include the creation of value in data spaces, their economic model, their definition as a marketplace or as an intermediation service. The

[1] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX%3A52020PC0767
[2] https://eur-lex.europa.eu/eli/reg/2023/2854
[3] https://eur-lex.europa.eu/eli/dir/2019/1024/oj
[4] https://eur-lex.europa.eu/eli/reg/2016/679/oj
[5] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=celex%3A32002L0058
[6] https://eur-lex.europa.eu/eli/dir/2019/790/oj

questions regarding the architecture are also of great relevance. In this sense, it is important to know whether the data space is set to centralise all data or merely hold a catalogue of metadata, and to understand the services it would offer.

This paper intends to focus mainly on data governance and not on data space governance, which shall be addressed in a second step, once the legal constraints applicable to the data are identified.

2. Impact of the various applicable rules

2.1. Commitments to respect the third parties' rights to data in compliance with the requirements of the Data Act and the Data Governance Act

- **Reconciling the rights of third parties with the uses envisaged**

The DGA presents chapters covering specific types of data, such as data from public sector institutions that are not covered by the Open Data directive obligations in accordance to them containing rights of third persons, including copyright, trade secret, and personal data. The Data Act in its Chapter sets out standards of data sharing in business-to-business and consumer-to-business framework. It notably presents obligations regarding data created through the use of Internet of Things (IoT) devices.

In its explanatory memorandum of the DGA, the European Commission states that the general objectives of such legislation is: "*(1) making public sector data available for re-use in situations where such data is subject to rights of others (such as privacy rights, IP rights, trade secrets, or other commercially sensitive information); (2) sharing data among businesses, against remuneration in any form; (3) allowing personal data to be used with the help of a personal data sharing intermediary that safeguards data subjects' rights under the GDPR; and (4) allowing data use on altruistic grounds*"

According to article 3 DGA, rules on re-use apply to *"data held by public sector bodies which are protected on grounds of: (a) commercial confidentiality, including business, professional and company secrets;(b) statistical confidentiality; (c) the protection of intellectual property rights of third parties; or (d)the protection of personal data"*

The obligations to respect third-party rights are horizontal in the new legislative landscape pertaining to data in the EU. These obligations are considered as a cornerstone of responsible data sharing and data services. This covers novel types of entities created by the DGA such as data intermediaries and data altruism organisations whose scope and role in the various data spaces is in the process of being defined.

- **Contractual fairness in the Data Act**

Article 8 of the Data Act states that: *"Where a data holder is obliged to make data available to a data recipient under Article 5 or under other Union law or national shall do so under fair, reasonable and non-discriminatory terms and in a transparent manner"*

The data act thus prohibits clauses that are deemed unfair in a business-to-business contract when the object of the contractual relationship is access and use of data. Clauses are deemed unfair if they limit liability for intentional acts or gross negligence that may bring prejudice to the contracting party. Some clauses can also be deemed unfair if a dominant party to the contract puts in place limitations on the other party regarding use or reuse of data during the period of the contract.

These limitations need to be integrated in the functioning of the data space through the standardisation of contract templates.

2.2. Data not covered by the DA and DGA

- **Data collected through text and data mining**

The Text and Data Mining exception incorporated in the 2019 Copyright in the Digital Market Directive aims to create a possibility of using copyrighted material that is accessed lawfully, either through subscriptions, open licences, or online availability, to extract new data while balancing the interests of rights holders.

A permissive Article 3 is limited to research organisations and cultural heritage institutions, while Article 4 is broader, allowing TDM by any entity. However, from the latter usage, rights holders can signal an opt-out that makes it impossible to mine the data.

The EU's soon-to-be AI act is noteworthy in its explicit connection between the use of copyrighted works for training AI models and the text and data mining exception outlined in Article 4 of the 2019 Copyright Directive. This entails an obligation of compliance with the opt-out option. The exceptions thus represent an important step in adapting EU copyright law to the development of AI technologies. It is therefore important that the data transactions in the data space are subject to due diligence when it comes to respecting opt-out options.

- **All other data**

Data and data sets can otherwise be protected under copyright or the database sui generis right stemming from the EU Database Directive 96/9/EC[7] when it comes to intellectual property mechanisms and GDPR when it comes to data that contains personal data. As well as data that is covered by trade secrets, in this sense, templates of confidentiality agreements can be proposed in order to align with the rights of third parties.

3. Warranties of data compliance with applicable laws

The various definitions of a data space allow a level of freedom of interpretation of its scope as they can be different depending on the sectoral needs.In this sense, the definition given by the Data Spaces Support Center (DSSC) of a Common European data space is "A sectoral/domain-specific data spaces established in the European single market with a clear EU-wide scope that adheres to European rules and values." Data spaces can therefore have different architectures and can centralise the data or not.

3.1 Warranting compliance in a decentralised data space architecture

One of the fundamental repercussions of the architecture is the different levels of responsibility that data spaces have in regard to data made available. In the framework of a decentralised data space, the data space operator does not centralise the data that transits through it. Rather, the data space acts as an enabler of transactions, generally via a central catalogue of metadata.

The data space operator does not assume this responsibility on behalf of the participants. Each participant in the data space must take full responsibility for their compliance. However, the data space operator can intervene in the framework of decentralised data spaces. This cannot be conceived as the entity that would be liable if non-compliant data transmits through it, rather it would mean that the data space operator offers help towards compliance as a service. In this sense, this service can be done through the provision of templates of compliance documentation and templates related to contractual transactions that can be adopted through data audit services that can be plugged into the data space. It is to be highlighted that in such cases where the data is not held and, by principle, not accessed by the data space operator, it is the data space participant who would define what data is exchanged, and in the case of data containing personal data, how the data is processed and for which objectives; thus coinciding with the "*means and purposes*" of GDPR.

In its guidelines 4/2019 on Article 25 Data Protection by Design and by Default (DPbDD)[8], the EDPB states that: *"Although not directly addressed in Article 25, processors and producers are also recognized as key enablers for DPbDD, they should be aware that controllers are required to only process personal data with systems and technologies that have built-in data protection."* The fact that the decentralised data space has no control over the data therefore does not exempt it from the expectations of complying with data protection by design and by default considerations as it enables eventual data processing activities.

Data governance arrangements should be in place at data spaces entry and exit points, based on the "compliance by design" principle, to optimise compliance with data (personal, non-personal, copyrighted…) protection laws, reduce the cost of compliance (barriers to entry) and gain efficiency.

For example, this is done via the verification that no data transfers allowed by the data space include personal data that has not gone through due diligence by the data participants. This is to allow the free flow of personal and non-personal data while respecting European values.

By clearly delineating the roles and responsibilities, the ecosystem ensures that all participants are aware of their obligations and can make informed decisions about their involvement in the data space, particularly about handling personal data.

The data space operator may put in place rules that close the space to non-compliant data. Using a code of conduct, it is possible to state that only data assessed by ex-ante against data protection, copyright, and trade secret considerations can transit via the data space It is notable that this does not necessarily imply a legal obligation of compliance from the data space operator's side when it is a decentralised architecture; however, it is evident that providing obligations of compliance to the exchanged data constitutes an added value for reusers who can be aware that the data was collected lawfully.

In fact, the data space operator plays a crucial role in facilitating compliance for data providers, enabling them to focus on the core aspects of data sharing and utilisation within the data space ecosystem. By offering standardised tools, the data space operator can significantly reduce the burden on data providers when it comes to implementing compliance measures. This can encourage more data providers to participate in the data space, as the administrative overhead is minimised. As the data spaces are a soft infrastructure, they can also create value via the provision of compliance services.

[7] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A31996L0009

[8] https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf

A similar analysis should be made regarding compliance with other bodies of rules such as copyright law or protection of trade secrets.

3.2 Warranting compliance in a centralised data space architecture

In this sense, although more and more data space initiatives are taking up that route, data spaces are not necessarily decentralised. This corresponds to a central platform that aggregates the data brought to the table by participants. Where the data space is an entity that centralises the data that will be the subject of transactions, may they be in the framework of selling, lending, or sharing on an altruistic basis, responsibilities in regard to compliance with relevant legislation can differ.

It is to be noted here that compliance with the rights of third parties becomes, among others, the responsibility, not only of the data space participant but also of the data space operator who centralises the data. In this sense, it becomes adequate to conceptualise the data space participants as controllers of data, where they can be individual controllers or joint-controllers on the basis of Article 26 GDPR. In this architecture, the data space operator would be generally conceived as a data processor that would therefore have specific legal responsibilities.

It is also important to assess the centrality of the data space against the risks of having unfair advantages towards data participants. As a centralised data space is naturally more closed and less penetrable by eventual data participants, this may result in unfairness that is proscribed by the Data Act.

4. Implementing mechanisms to automate compliance warranty in data spaces

4.1 Metadata interoperability

Data sharing requires incentives for data holders to develop metadata to increase certainty about rights and obligations in relation to data; indeed, efficient metadata management, with the development of standards for semantic and technical interoperability, is of utmost importance. Therefore, Article 28 of the Data Act provides that "*Operators of data spaces shall comply with, the following essential requirements to facilitate interoperability of data, data sharing mechanisms and services: a) the dataset content, use restrictions, licences, data collection methodology, data quality and uncertainty shall be sufficiently described to allow the recipient to find, access and use the data;[…]*"

Special focus on these matters shall be made for an optimal allocation of data to the benefit of society. These cross-sectoral interoperability standards, will be one of the responsibilities of the European Data Innovation Board (EDIB) as one of its missions is proposing guidelines for common European data spaces pursuant to Article 30 of the DGA.

4.2 Integration of Distributed Ledger Technology and Smart Contracts

The integration of Distributed Ledger Technology (DLT) into data spaces can have many benefits for data space participants. By integrating DLT into the data space the ecosystem can foster greater trust, transparency, and efficiency in the exchange and utilisation of valuable data assets.

DLT, mostly reflected in blockchain, empowers data providers as it significantly enhances the value proposition within the data space ecosystem. By creating a permanent, tamper-proof record of data contributions, blockchain can ensure that sellers' efforts are irrefutably acknowledged and their intellectual property rights are protected.

Similarly, smart contracts enabled by DLT can automate the process of compensating data providers upon the fulfilment of predefined conditions, ensuring timely and fair remuneration for their data assets.

DLT's transparency and immutability allow data providers to conclusively prove the origin and ownership of their data, while also guaranteeing that the information has not been altered from its original form, preserving its quality and reliability. As data transactions are recorded on the distributed ledger, providers can establish a verifiable track record of their contributions, which can lead to increased business opportunities and enhanced pricing power over time. Data users can easily verify the complete history of a data asset, including its origins, ownership changes, and any relevant licensing terms, reducing administrative overhead.

Smart contracts can streamline the process of obtaining the necessary rights to use the data, further enhancing efficiency and reducing the risk of licensing disputes. The decentralised architecture of DLT makes it significantly more challenging to engage in fraudulent activities, as altering recorded data would require consensus across the entire network.

Moreover, DLT's inherent security features can provide an additional layer of protection for the storage and transaction of sensitive language data, mitigating the risks of unauthorised access or tampering.

5 Conclusion

By aligning best practices and harmonising regulatory requirements across different data space initiatives, a cohesive and interoperable compliance framework emerges thus resulting in data spaces where data can flow freely and securely.

As data spaces emerge as infrastructure enabling the exchange and utilisation of valuable information assets, the need for a comprehensive compliance framework becomes paramount. This is translated through the integration of compliance services into the foundation of the data space infrastructure

At the heart of this compliance-driven approach lies the imperative to safeguard compliance of the data to applicable laws.

As the data space ecosystem continues to evolve, compliance-driven approaches need to mutate and be flexible enough to respond to the needs of data space participants.

6 Bibliographical references

- Duisberg, A., "Legal Aspects of IDS: Data Sovereignty-What Does It Imply?." *Designing Data Spaces* 61 (2022).
- Ruohonen, J., Mickelsson, S. Reflections on the Data Governance Act. *DISO* **2**, 10 (2023).
- Margoni, T., Ducuing, C., Schirru, L., Data Property, Data Governance and Common European Data Spaces. Computerrecht: Tijdschrift voor Informatica, Telecommunicatie en Recht, (2023)

A Legal Framework for Natural Language Processing Model Training in Portugal

Rúben Almeida, Evelin Amorim
INESC TEC, FCUP-Universidade do Porto
ruben.f.almeida@inesctec.pt, evelin.f.amorim@inesctec.pt

Abstract

Recent advances in deep learning have promoted the advent of many computational systems capable of performing intelligent actions that, until then, were restricted to the human intellect. In the particular case of human languages, these advances allowed the introduction of applications like ChatGPT that are capable of generating coherent text without being explicitly programmed to do so. Instead, these models use large volumes of textual data to learn meaningful representations of human languages. Associated with these advances, concerns about copyright and data privacy infringements caused by these applications have emerged. Despite these concerns, the pace at which new natural language processing applications continued to be developed largely outperformed the introduction of new regulations. Today, communication barriers between legal experts and computer scientists motivate many unintentional legal infringements during the development of such applications. In this paper, a multidisciplinary team intends to bridge this communication gap and promote more compliant Portuguese NLP research by presenting a series of everyday NLP use cases, while highlighting the Portuguese legislation that may arise during its development.

Keywords: Portuguese NLP, Legal NLP, PLN Português

1. Introduction

In recent years, deep-learning-based methods have permitted great advances in Computer Science (CS) fields previously known for their computational complexity (Bishop and Bishop, 2023). One of those fields was natural language processing (NLP); the CS field focused on machine understanding of human languages and the generation of coherent human text (Deng and Liu, 2018).

Long before the introduction of state-of-the-art (SOTA) generative large language models (LLMs) like Chat-GPT (OpenAI, 2022), NLP relied on simpler approaches based on rules/word counts (Jurafsky and Martin, 2014) to deliver human-interpretable models capable of addressing simple NLP tasks. The evolution towards deep learning permitted NLP to quit the controlled atmosphere of university labs and embrace mainstream societal adoption at the expense of costly training processes and vast volumes of textual data.

The capabilities revealed by SOTA NLP models were accompanied by ethical and legal concerns among prominent NLP researchers[1], big-tech CEOs[2], politicians[3], and economists[4] who appealed for regulation and higher ethical standards during the development of NLP solutions as a way of restricting the usage of such capabilities to induce harm in society.

However, the pace at which new LLMs are currently being developed largely surpasses the pace at which new regulations are introduced. This phenomenon, together with the traditional communication barriers between the formality of legal expertise and the dynamic world of CS, opens the door to legal vacuums, promoting undesirable copyright[5] and privacy infringements[6].

In this paper, we intend to bridge this gap by listing the legal concerns that may arise during everyday NLP challenges. We focus on the Portuguese legal system to present a study targeting both computer scientists and legal experts, with the ultimate goal of promoting awareness about the topic in one of the most peripherally and underdeveloped[7] countries of the European Union (EU). Despite our focus on Portugal, we believe researchers in other EU countries may find the concepts introduced in this paper useful, due to the European nature of many of the legislation listed in this paper.

The paper is structured as follows: Section 2 de-

[1] https://twitter.com/ylecun/status/1733481002234679685

[2] https://www.reuters.com/technology/musk-experts-urge-pause-training-ai-systems-that-can-outperform-gpt-4-2023-03-29/

[3] https://www.ft.com/content/9339d104-7b0c-42b8-9316-72226dd4e4c0

[4] https://www.weforum.org/agenda/2017/03/taxing-robots-wont-work-says-yanis-varoufakis/

[5] https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html

[6] https://www.forbes.com/sites/emmawoollacott/2023/09/01/openai-hit-with-new-lawsuit-over-chatgpt-training-data/?sh=2f3d1b856d84

[7] https://en.wikipedia.org/wiki/List_of_countries_by_Human_Development_Index

LEGAL2024 @LREC-COLING-2024, pages 6–12

scribes the existing literature about the topic. The lack of research, focused exclusively on the Portuguese case, forced us to broaden our scope towards other EU countries. Section 3 provides a quick overview of the current SOTA for Portuguese NLP, with a special emphasis on explaining the importance of Brazilian NLP research for Portuguese NLP. In Section 4, we briefly introduce the Portuguese legal system, performing a high-level description of the legislation that may engage with the development of NLP solutions. Section 5 lists common licensing agreements used by NLP researchers to publish their works. Finally, in Section 6, we introduce some use cases representing everyday challenges faced by NLP researchers highlighting the legal concerns associated with them.

2. Related Work

The novelty of this subject translates into a lack of literature about the topic. The absence would be even worse if we focused exclusively on the Portuguese legal landscape. Therefore, in the following section, we broaden our scope outside Portugal towards other EU countries.

2.1. Portuguese

The existing literature concerning the legal implications of artificial intelligence (AI) systems in Portugal is constrained to regional conferences and journals in the Portuguese language. We highlight the pioneering works of (Guimarães et al., 2021) that compiled summarized versions of master theses developed by law students about different CS topics, including Big Data (Silva Costa, 2021).

Recently, (Barbosa, 2023) introduced the analysis concerning data privacy during LLM development. Despite focusing on the Portuguese legal framework, the writer relies heavily on EU regulations to provide a brief overview of the recent developments surrounding the topic. In particular, it is demonstrated that the General Data Protection Regulation (GDPR) alone is incapable of regulating LLMs like ChatGPT. The author mentions the importance of the AI Act to establish a legal framework founded on the concept of **risk** to deal with the inaccuracies of these models.

Focusing on copyright issues, we highlight the works of (Nobre, 2012), which provide an extensive listing of what is protected and what is excluded from copyright protection.

Lastly, it is worth mentioning the interesting considerations of (Pereira, 2020) regarding copyright eligibility for AI outputs. In Portugal, the copyrights for AI results are free, unlike in other countries like the United Kingdom, where the research team who made them inherits the rights[8].

2.2. European

At the European level, we highlight the works of (Eckart de Castilho et al., 2018) and (Kelli et al., 2020) focusing on copyright. Both works discuss whether an LLM trained in copyrighted protected data inherits the same license as its training data. Both authors concluded there is a lack of clarity on the topic, but it all depends on how much copyright data is used. If a model can reproduce data that is protected by copyright, it is called a derivation of the dataset and has the same copyright as the dataset used.

Nevertheless, most of the European literature about the topic is produced by enterprises. We highlight the efforts of ML6[9], KPMG[10] or EY[11] to promote the debate about the topic. These blog posts often have anonymous authors who are not checked by others. This makes it hard to know if they are accurate and up-to-date.

3. Portuguese NLP: Quick Overview

Portuguese is the official language of 260 million people spread across five continents. In the context of NLP, Portuguese is considered a mid-resourced language (Joshi et al., 2021). Meaning that "they have a large amount of unlabeled data...and are only challenged by lesser amount of labeled data" (Joshi et al., 2021). Despite this classification, the vast majority of these resources are Brazilian Portuguese, produced by Brazilian research teams.

Similar to other mid-resource languages, the shorter investment produces a reduced number of Portuguese LLMs. Until recently, BERTimbau (Souza et al., 2020), a Brazilian Portuguese version of BERT, was the only Portuguese LLM. Progressively, more complex architectures have emerged: Albertina PT (Rodrigues et al., 2023) or the generative models Sabiá (Pires et al., 2023), Gervásio (Santos et al., 2024), and GlórIA (Lopes et al., 2024).

[8] https://pec.ac.uk/blog_entries/copyright-protection-in-ai-generated-works/

[9] https://www.ml6.eu/blogpost/navigating-ethical-considerations-developing-and-deploying-large-language-models-llms-responsibly

[10] https://kpmg.com/ie/en/home/insights/2024/01/eu-artificial-intelligence-act-art-int.html

[11] https://www.ey.com/en_ch/forensic-integrity-services/the-eu-ai-act-what-it-means-for-your-business

3.1. Leveraging Brazilian Portuguese Resources

Despite the major phonological, morphological, lexical, syntactical and semantic differences between the numerous Portuguese varieties around the globe (Scherre and Duarte, 2016), prompt-engineering[12] and fine-tuning[13] of Brazilian LLMs help European Portuguese researchers achieve SOTA results (Almeida et al., 2024).

4. Portuguese Legal System

Portugal, as an EU member, must ensure its legal system is in harmony with EU law. In this section, we list the legislation that engages with NLP development, while describing how EU legislation impacts the Portuguese legal system.

4.1. Scientific Exceptions

Many of the legislation listed in the following subsections establish exceptions to scientific work. For that reason, we believe it is paramount to clarify what is considered **scientific work** by the Portuguese law.

The (European Parliament and Council of the European Union, 2019) directive introduces a comprehensive definition of scientific work founded on the concept of profit: “research organisation means a university, including its libraries, a research institute or any other entity, the primary goal of which is to conduct scientific research or to carry out educational activities involving also the conduct of scientific research: (a) on a not-for-profit basis or by reinvesting all the profits in its scientific research; or (b) pursuant to a public interest mission recognised by a Member State" (European Parliament and Council of the European Union, 2019).

4.2. National Legislation

Portuguese legislation is divided into two major sets of laws: the civil code and the penal code. The penal code lists those actions that constitute a crime and require the existence of **willful misconduct** to be taken into consideration. While in the civil code, the penalties restrained themselves to some kind of compensation, the penal code establishes tougher penalties like prison time. Committing a crime is a severe action that usually implies not only a penal condemnation but also a parallel civilian compensation to repair any harm provoked in society.

Regarding NLP, most of the legislation that applies is restricted to the civil code; however, in recent years, with the mass adoption of CS in society, many penal considerations have arisen focusing on copyright infringement, privacy violations, or cybercrime.

Finally, before enumerating the Portuguese legislation directly relevant to NLP, it is essential to clarify the distinctions between public, semi-public, and private crimes. Public crimes warrant the intervention of public prosecutors to safeguard the interests of the victim. Semi-public crimes trigger public prosecution only upon the filing of a complaint with the police. In cases of private crimes, typically of lesser severity, public prosecution does not intervene, and a complaint must also be lodged with the police. In instances where the penal code does not explicitly specify the classification of the crime, it defaults to being considered public[14].

Sensitive Data Protection: Article 35 of the Portuguese constitution (República Portuguesa, 1976) strictly prohibits the digital processing of sensitive ethnic, political, sexual, or religious data without explicit consent and for scientifically motivated purposes that may positively impact society. Consequently, NLP researchers are advised to refrain from operating on such sensitive data.

Right to Expression: Article 79 of the Portuguese civil code (República Portuguesa, 1966) affirms the right of individuals to freely express themselves without fear of their words being recorded or utilized by a third party. This legislation outlines exceptions for scientific research. Without these provisions, the use of tweets to train NLP models would be deemed illegal.

4.3. European Legislation

The great investment the EU has been making to regulate AI has had a great effect in peripheral countries like Portugal, helping reduce inequalities in access to technology between richer and poorer countries within the EU.

Before enumerating EU regulations that engage directly with the development of NLP solutions, it is important to clarify some terminology regarding EU law:

Regulation: Regulations come into force automatically upon approval by the European Parliament and European Council. Requiring a 65% majority for acceptance, regulations are always accompanied by a transition period to allow EU citizens to adapt to the new legal framework.

[12] https://en.wikipedia.org/wiki/Prompt_engineering

[13] https://en.wikipedia.org/wiki/Fine-tuning_(deep_learning)

[14] https://www.ministeriopublico.pt/perguntas-frequentes/crime

Directive: Directives represent EU laws that member states must adopt within a specified transition period, integrating them into their own legal systems. Despite penalties for countries failing to timely adopt directives, Portugal has earned a reputation for delays in transposing these directives. A pertinent example is the 2019/790 directive, where the transposition was delayed by over two years[15].

Below, we briefly introduce the most relevant EU legislation that engages with NLP development:

General Data Protection (Regulation): Enacted in 2016, this EU regulation (European Parliament and Council of the European Union, 2016) sets the foundational standards for online data protection. Emphasizing transparency and consent, it mandates that all corporate operations involving personal data must be pre-informed and explicitly consented to by users. The regulation applies specifically to private data - information that, if disclosed, can identify the individual to whom it pertains. It also regulates cross-border data transfers, stipulating that EU data protection laws accompany the data wherever it goes. Outside the EU, in countries lacking an **adequacy decision**[16], such as Brazil, additional measures may be necessary to ensure equivalent levels of protection as those mandated within the EU. This regulation was integrated into Portuguese legislation in 2021, complemented by the introduction of the **Portuguese Charter on Human Rights in the Digital Age** (Assembleia da República, 2021). Concerning NLP, explicit consent is required for data collection and usage in training NLP models, with researchers urged to minimize the use of private data whenever possible. While scientific research exceptions exist, the legislature advises their avoidance, favoring maximal consent.

Copyright in the Digital Single Market (Directive): The 2019/790 EU directive (European Parliament and Council of the European Union, 2019) on copyrights introduced a **right to [text] mine**, framing the legal context for NLP development. It broadened exceptions, permitting scientists to use copyrighted data for NLP model training if not pursued for profit. Transposed into Portuguese legislation in 2023 (Presidência do Conselho de Ministros, 2023), this directive provides a legal framework for NLP endeavors.

Database Sui Generis Protection (Directive): Directive 96/97EC (European Parliament and Council of the European Union, 1996) established copyright protection for databases, granting 15 years of copyright protection to those who compile independent works into structured units. This regulation was revisited during the approval of the 2019/790 directive, with exceptions allowing copyright infringements for scientific endeavors also applying to databases.

AI Act (Regulation): Recently approved, this regulation (European Commission, 2021) aims to be the world's first comprehensive AI legal framework, positioning the EU at the forefront of AI legislation. It sets rules for AI systems, including NLP, based on their societal risk levels. Additionally, it mandates transparency mechanisms during the training of LLMs exceeding 10^{25} FLOPS. Given that most NLP research does not involve sensitive data or yield automated decisions significantly impacting Europeans' daily lives, it is categorized as **minimal-risk**, exempting researchers from additional considerations. The AI Act is in the final stages of approval and is expected to be fully implemented within two years.

5. NLP Licensing System

In Table 1, we list the licenses of many NLP resources.

Model Name	Year	License
BERT	2018	Apache 2.0
GPT-2	2019	MIT
Bloom	2022	Custom License
Falcon	2023	Apache 2.0
Llama 2	2023	Custom License
Mistral	2023	Apache 2.0
Phi-2	2023	MIT
Gemma	2024	Custom License
BERTimbau	2020	MIT
Albertina PT	2023	MIT
Sabiá	2023	Llama 1 License
Gervásio	2024	MIT
Glória	**2024**	**ClueWeb22 License**
Wikitext	2016	CC-BY-SA 3.0
CNN-Dailymail	2017	Apache 2.0
Flores	2019	CC-BY-4.0
OSCAR	2023	CC0-1.0

Table 1: Licenses for different SOTA NLP resources. The first entries cover non-Portuguese LLMs, while the second set focuses on Portuguese architecture. The last set of resources are commonly used NLP datasets.

The results reveal that many LLMs adopt an Apache 2.0 or MIT license. The datasets identified tend to adopt Common Crawl licenses. It is worth mentioning the case of the Portuguese LLM Glória (Lopes et al., 2024), where the usage of the

[15] https://www.cuatrecasas.com/pt/portugal/art/transposicao-da-diretiva-do-mercado-unico-digital-1

[16] https://www.edpb.europa.eu/sme-data-protection-guide/international-data-transfers_en

clueweb22 dataset (Overwijk et al., 2022) as part of the training corpus required Glória's authors to adopt this license as well.

The information provided in this section is complemented by the extensive analysis made by the choose a license platform[17].

6. Use Cases

In this section, we cover three use cases that represent everyday NLP tasks. In the following subsections, we assume the perspective of a Portuguese NLP operating under EU law. The green color used in the flowcharts represents the yes/permitted case; in contrast, the red color represents the no/not permitted case.

6.1. Load Brazilian Portuguese Dataset From HuggingFace

This use case describes a standard practice in Portuguese NLP, where Brazilian datasets are used to train new NLP models. In Figure 1 we provide a flowchart summarizing the legal questions that may arise during the process.

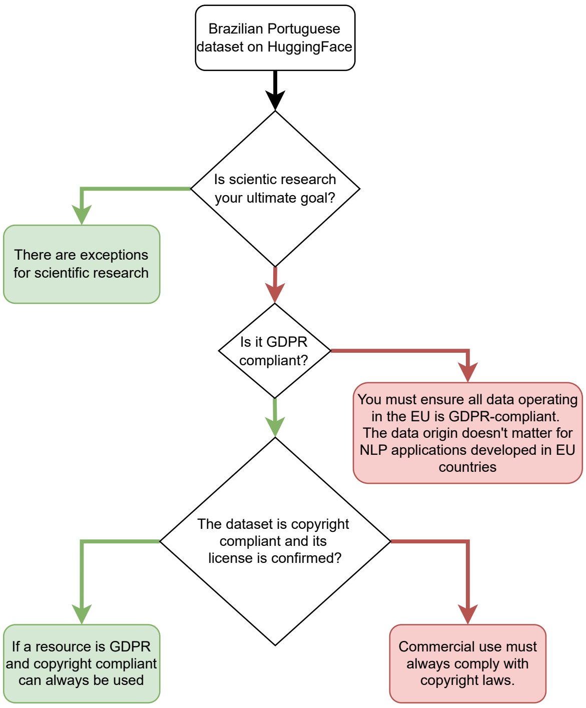
Brazilian Portuguese dataset on HuggingFace
Is scientic research your ultimate goal?
There are exceptions for scientific research
Is it GDPR compliant?
You must ensure all data operating in the EU is GDPR-compliant. The data origin doesn't matter for NLP applications developed in EU countries
The dataset is copyright compliant and its license is confirmed?
If a resource is GDPR and copyright compliant can always be used
Commercial use must always comply with copyright laws.

Figure 1: Flowchart summarizing the legal questions associated with the loading of a Non-EU dataset

It is worth mentioning that the geographical origin of the dataset has no impact on the overall legal assessment of the resource.

6.2. Crawl Portuguese Websites to Produce a Large NLP Corpus

The usage of web crawling techniques is paramount for LLM training. SOTA LLMs require a vast amount of data, whose scale is only comparable to the amount of information existing on the web. In Figure 2, we describe the legal considerations researchers should pay attention to while crawling websites.

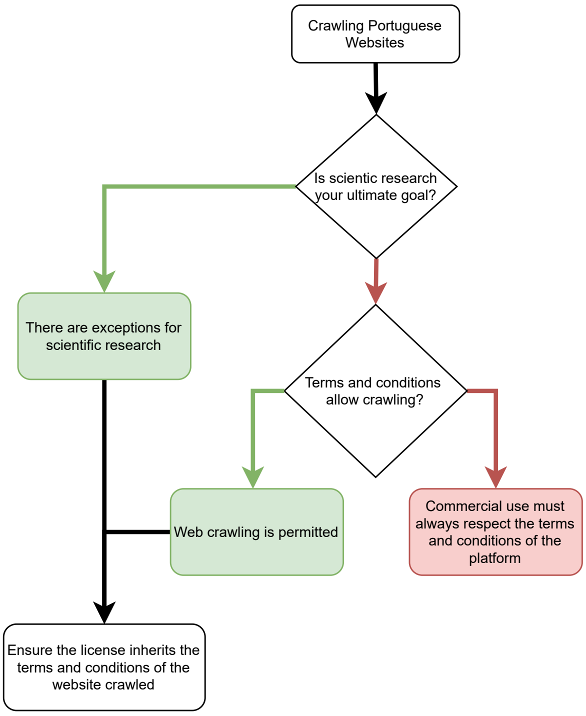
Crawling Portuguese Websites
Is scientic research your ultimate goal?
There are exceptions for scientific research
Terms and conditions allow crawling?
Web crawling is permitted
Commercial use must always respect the terms and conditions of the platform
Ensure the license inherits the terms and conditions of the website crawled

Figure 2: Flowchart highlighting the legal considerations web crawling may arise.

6.3. Use Tweets to Produce Political Profiles: The Facebook-Cambridge Analytica case

This use case draws inspiration from the Facebook–Cambridge Analytica data scandal[18]. It aims to encapsulate the legal considerations that may emerge during the development of NLP models, particularly in scenarios involving sensitive data such as political profiling. Figure 3 provides a summarized overview of these legal aspects.

7. Conclusion & Future Work

In this study, we have presented the inaugural legal framework for NLP development in Portugal. The limited awareness of this subject in one of the EU's less affluent nations exposes vulnerabilities to

[17] https://choosealicense.com/appendix/

[18] https://shorturl.at/uxK47

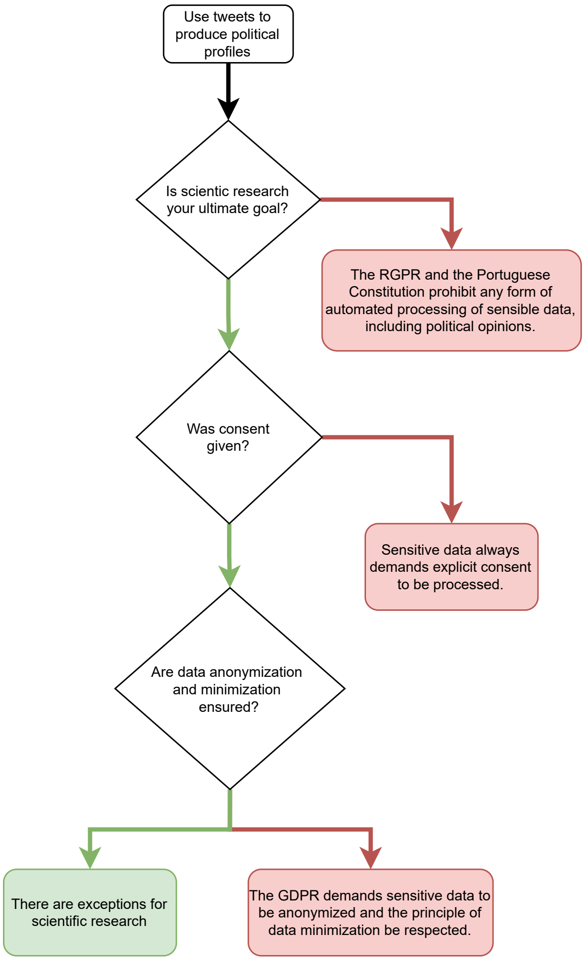
Use tweets to produce political profiles
Is scientic research your ultimate goal?
The RGPR and the Portuguese Constitution prohibit any form of automated processing of sensible data, including political opinions.
Was consent given?
Sensitive data always demands explicit consent to be processed.
Are data anonymization and minimization ensured?
There are exceptions for scientific research
The GDPR demands sensitive data to be anonymized and the principle of data minimization be respected.

Figure 3: Flowchart concerning the legal issues of processing sensitive data.

legal vacuums and potential infringements of copyright and data privacy. We have tried to introduce key NLP concepts in a way that is accessible to both computer scientists and legal experts. The three use cases outlined serve to summarize the principal insights gained from this research. Utilizing flowcharts to illustrate these scenarios aims to accelerate the procedure of obtaining data and enhance overall comprehension of the research.

Our future work will focus on expanding the scope of use cases, matching to the same structured approach, while exploring additional topics. Specifically, we aim to incorporate insights from the recently introduced regulation on AI, the European AI Act, to further enrich the legal framework for NLP development.

8. Acknowledgements

This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia, within project LA/P/0063/2020. DOI 10.54499/LA/P/0063/2020 | https://doi.org/10.54499/LA/P/0063/2020 The author also would like to aknowledge the project StorySense, with reference 2022.09312.PTDC (DOI 10.54499/2022.09312.PTDC).

9. Bibliographical References

Rúben Almeida, Ricardo Campos, Alípio Jorge, and Sérgio Nunes. 2024. Indexing portuguese nlp resources with pt-pump-up. *arXiv preprint arXiv:2401.15400*.

Assembleia da República. 2021. Carta Portuguesa de Direitos Humanos na Era Digital. Diário da República Eletrónico. Lei n.º 27/2021 de 17 de maio.

Mafalda Barbosa. 2023. Proteção de dados e inteligência artificial – perigos e soluções (também a propósito do chatgpt. *Revista Do Direito Da Responsabilidade*, 5(1):796–833.

Christopher M Bishop and Hugh Bishop. 2023. *Deep Learning*. Springer Nature.

Li Deng and Yang Liu. 2018. *Deep Learning in Natural Language Processing*. Springer, Singapore.

Richard Eckart de Castilho, Giulia Dore, Thomas Margoni, Penny Labropoulou, and Iryna Gurevych. 2018. A legal perspective on training models for natural language processing.

European Commission. 2021. COM(2021) 206 final. European Commission Document. Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL ESTABLISHING HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE REGULATION) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS.

European Parliament and Council of the European Union. 1996. DIRECTIVE 96/9/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Official Journal of the European Union. Directive on the legal protection of databases.

European Parliament and Council of the European Union. 2016. REGULATION (EU) 2016/679 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Official Journal of the European Union. Regulation on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

European Parliament and Council of the European Union. 2019. DIRECTIVE (EU) 2019/790 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL. Official Journal of the European Union. Directive on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC.

Maria Raquel Guimarães, Rute Teixeira Pedro, Maria Regina Gomes Redinha, Fernanda de Araujo Meirelles Magalhães, et al. 2021. Direito digital.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2021. The state and fate of linguistic diversity and inclusion in the nlp world.

Dan Jurafsky and James H Martin. 2014. *Speech and Language Processing : an Introduction to Natural Language processing, Computational linguistics, and Speech Recognition*. Dorling Kindersley Pvt, Ltd, India.

Aleksei Kelli, Arvi Tavast, Krister Lindén, Kadri Vider, Ramunas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Stranák, et al. 2020. The impact of copyright and personal data laws on the creation and use of models for language technologies. In *Selected Papers from the CLARIN Annual Conference 2019*, pages 53–65. Linköping University Electronic Press.

Ricardo Lopes and João Magalhães and David Semedo. 2024. *GlórIA – A Generative and Open Large Language Model for Portuguese*.

Teresa Nobre. 2012. *Direito de Autor e Direitos Conexos*. University of lisbon.

OpenAI. 2022. Introducing chatgpt.

Arnold Overwijk, Chenyan Xiong, and Jamie Callan. 2022. Clueweb22: 10 billion web documents with rich information. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 3360–3362.

Alexandre Pereira. 2020. *A PROTEÇÃO JURÍDICA DO SOFTWARE EXECUTADO POR ROBOTS (E DAS OBRAS GERADAS POR IA)**, page 239–254. Universidade De Coimbra, Coimbra.

Pires, Ramon and Abonizio, Hugo and Almeida, Thales Sales and Nogueira, Rodrigo. 2023. *Sabiá: Portuguese large language models*. Springer.

Presidência do Conselho de Ministros. 2023. Decreto-Lei n.º 47/2023. Diário da República. Decreto-Lei transpondo a Diretiva (UE) 2019/790 sobre direitos de autor e direitos conexos no mercado único digital.

Erasmo Marcos Ramos. 1998. A influência do bürgerliches gesetzbuch alemão na parte geral do novo código civil português. *Revista da Faculdade de Direito da UFRGS*, (15).

República Portuguesa. 1966. Código Civil. Diário do Governo. Article 79.º: Direito à imagem.

República Portuguesa. 1976. Constituição da República Portuguesa. Official Journal of the Portuguese Republic. Article 35: Utilização da informática.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing neural encoding of portuguese with transformer albertina pt-*.

Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024. Advancing generative ai for portuguese with open decoder gervásio pt*.

Maria Marta Pereira Scherre and Maria Eugênia Lammoglia Duarte. 2016. Main current processes of morphosyntactic variation. *The Handbook of Portuguese Linguistics*, pages 526–544.

Inês Silva Costa. 2021. A proteção da pessoa na era dos big data: a opacidade do algoritmo e as decisões automatizadas. *Revista Electrónica De Direito*, 24(1):33–82.

Souza, Fábio and Nogueira, Rodrigo and Lotufo, Roberto. 2020. *BERTimbau: pretrained BERT models for Brazilian Portuguese*. Springer.

Intellectual Property Rights at the Training, Development and Generation Stages of Large Language Models

Christin Kirchhübel[1], Georgina Brown[2]
Atlantic Chambers, Liverpool, UK[1]
Department of Linguistics and English Language, Lancaster University, UK[2]
ck@atlanticchambers.co.uk[1], g.brown5@lancaster.ac.uk[2]

Abstract

Large Language Models (LLMs) prompt new questions around Intellectual Property (IP): what is the IP status of the datasets used to train LLMs, the resulting LLMs themselves, and their outputs? The training needs of LLMs may be at odds with current copyright law, and there are active conversations around the ownership of their outputs. A report published by the House of Lords Committee following its inquiry into LLMs and generative AI criticises, among other things, the lack of government guidance, and stresses the need for clarity (through legislation, where appropriate) in this sphere. This paper considers the little guidance and caselaw there is involving AI more broadly to allow us to anticipate legal cases and arguments involving LLMs. Given the pre-emptive nature of this paper, it is not possible to provide comprehensive answers to these questions, but we hope to equip language technology communities with a more informed understanding of the current position with respect to UK copyright and patent law.

Keywords: Intellectual Property, copyright, Large Language Models (LLMs)

1. Introduction

Intellectual Property (IP) can be protected by patents, trademarks, copyright, and design rights, amongst others. As relatively uncharted territory in law, we consider copyright and patent law specifically in relation to the training and development of Large Language Models (LLMs), as well as the IP status of the outputs that LLMs generate. Because little to no IP caselaw yet exists specifically in relation to LLMs, this paper turns to discussions and caselaw concerning neighbouring Artificial Intelligence (AI) technologies as these will likely extend to LLMs as legal cases arise in the future.

2. Copyright

Copyright is an unregistered right meaning it arises automatically. It is available for literary, dramatic, musical or artistic works, sound recordings, films, broadcasts, and the typographical arrangement of published works provided that these works are 'original'. Owners of copyright have the exclusive right to do the 'acts restricted by the copyright' specified in Section 16(1) of the Copyright, Designs and Patents Act 1988 (the "1988 Act"), which includes, among others, making a copy of the work. In the absence of any defences or exceptions, copyright infringement would occur when the whole or a substantial part of copyright protected work is copied without permission, for example. When thinking about copyright in the context of LLMs, it is logical to differentiate between 'input', i.e., the data used to train a LLM, and 'output', i.e., the data generated by a LLM.

2.1 Input

A pertinent issue is the extent to which training a LLM poses copyright infringement risks. Training a LLM relies on text and data mining (TDM) of large amounts of data. While some LLM developers are more transparent than others about the data that they rely on, there is strong evidence to suggest that, in many instances, the data used will be covered by copyright protection. For example, in written evidence to the House of Lords Communications and Digital Committee (the "House of Lords Committee") who conducted an inquiry into 'Large Language Models and Generative AI' (report dated 2 February 2024), Open AI admitted that it was "impossible to train today's leading AI models without using copyrighted materials" and attempting to do so "would not provide AI systems that meet the needs of today's citizens" (Open AI—written evidence (LLM0113)). Another example may be seen in the case of Getty Images (US), Inc. v. Stability AI, Inc., 1:23-cv-00135. Getty has issued copyright infringement proceedings (among others) against Stability AI for 'scraping' millions of images from the Getty Images Websites without Getty's consent and then using those images as input to train and develop its AI model. Getty claims that, in many cases, the output delivered by Stability AI includes a modified version of a Getty Images watermark, from which it can be inferred that Stability AI has been trained on Getty's data.

TDM is not a uniform process; rather, it varies between LLM developers. While some TDM methods may involve the copying of whole works, other TDM approaches may 'only' collect links to websites. Some may require a copy of the work to be retained, others may only necessitate temporary copies which are discarded once the relevant information has been extracted. As such, whether TDM involves copying of the whole or substantial part of the work is a moot point and likely to be case-specific.

On the assumption that TDM is considered to involve copying, it follows that developers run the risk of copyright infringement. One way for developers to avoid this risk of copyright infringement would be to rely on the exception afforded by Section 29A of the 1988 Act which permits TDM for non-commercial purposes. In the EU context, Article 3(1) of the Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market appears to provide a similar exception, whereby TDM is permitted for purposes of scientific research. Needless to say, while provisions such as these may be a solution for developers operating in the research environment, they would be inadequate for commercial purposes. Another way of avoiding copyright infringement would be to obtain permission from the copyright holders by way of a licence. As already highlighted, during their inquiry into LLMs in Autumn 2023, the House of Lords Committee received oral and written evidence around the issue of copyright, and it became abundantly clear that LLM developers were using copyrighted data to train models without permission, i.e., without a licence and for commercial purposes.

In view of the evidence presented, there is a clear tension between the interests of developers on the one hand, and copyright holders on the other. There appears to be agreement between developers that access to copyright protected works is essential to ensure that AI systems perform the best they can, and the need for licence agreements could be prohibitive for this quest. However, using copyright protected works without permission goes against the whole purpose of copyright which is to reward original creations and incentivise innovation.

The UK government attempted to resolve this tension through the introduction of an AI copyright code of practice. In summer 2023, the UK Intellectual Property Office (UKIPO) set up a working group involving stakeholders from the technology, creative and research sectors. Members included the BBC, the British Library, Financial Times, Google Deepmind, IBM, Microsoft, Stability AI, UK Research and Innovation. The UKIPO said that: *'The code of practice aims to make licences for data mining more available. It will help to overcome barriers that AI firms and users currently face, and ensure there are protections for rights holders. This ensures that the UK copyright framework promotes and rewards investment in creativity. It also supports the ambition for the UK to be a world leader in research and AI innovation.'* (UKIPO, 2023). In February 2024, the UK government announced that it had shelved plans to put in place this code as it had become clear that the working group would not be able to reach agreement. (Department for Science, Innovation & Technology, 2024: 19).

While we wait for the government to clarify the relationship between IP and AI, all we are left with is the existing law. Considering this issue in the context of the 1988 Act, there is legal uncertainty as to whether commercial AI developers are infringing copyright when training AI systems on copyright protected material without a licence. As alluded to above, it may be argued that TDM does not actually involve 'making a copy' for the purposes of the 1988 Act, and even if it did, this copy may only be temporary and therefore fall within the exceptions allowing for transient or incidental copies provided for by Section 28A of the 1988 Act. Further, even if TDM were to involve copying, given that this is only part of the training stage, and given that the actual AI model does not directly reproduce the copyright protected work, but rather reflects the data / information contained within that work, does copyright even apply in those circumstances? Understandably, litigation around issues such as these has already started and is likely to grow in the near future.

2.2 Output

Questions around copyrightability also arise for LLM output – can AI-generated material attract copyright protection in the first place, and if so, who is the author? In relation to the former, the 1988 Act expressly provides for the copyright protection of literary, dramatic, musical or artistic work which is computer-generated in circumstances such that there is no human author of the work. This is in contrast to the position in the US, for example, where only works with human authors can receive copyright protection. Even though current UK legislation seems to explicitly cater for copyright in AI-generated work, in order for this work to be a true candidate for copyright protection, it needs to be 'original'.

Traditionally, the test for originality in the UK had a low threshold, requiring the work to be produced with 'sufficient skill, labour, and judgment'. This changed following the judgment of the European Court of Justice in the case of Infopaq International A/S v Danske Dagblades Forening [2009], with UK courts adopting the EU requirement of work having to exhibit the 'author's own intellectual creation' in order to be deemed ‘original’. There is some uncertainty around how UK courts are going to interpret the originality requirement going forward. However, in view of provision 9(3) of the 1988 Act, which states, *‘In the case of a literary, dramatic, musical or artistic work which is computer-generated, the author shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken’,* it would appear that works without a human author could meet the originality requirement. It follows that AI-generated work which has had some element of human involvement may very well pass the originality test and therefore could be copyrightable.

Turning now to the question of authorship, there are a number of possible authors – the developer of the AI system, the user of the AI system, i.e., the prompt engineer, or the AI system itself. Section 9(3) of the 1988 Act (reproduced above) provides that the author of computer-generated work *‘shall be taken to be the person by whom the arrangements necessary for the creation of the work are undertaken.’* Given that statutory drafting refers to a ‘*person*’, it may be the case that the AI system cannot be the author for the purposes of the 1988 Act. In the absence of a contract, whether it is the person who built the AI system, or the prompt engineer ‘who made the necessary arrangements’ is likely to be case-specific.

As with the input stage, the topic of copyright infringement is also relevant to the output stage. Guadamuz (2024) presents a detailed discussion of the relevant issues including possible defences to arguments that AI generated output infringes copyright. Briefly here, the points that are likely going to be debated in this sphere include the copyright infringement potential of memorisation, i.e., LLM models “memorising” specific fragments of their training data, and then reproducing these fragments in their output (Emanuilov and Margoni, 2024). Rather than reproducing existing work ‘verbatim’, perhaps a more likely scenario involves AI output resembling input data so the legal analysis will revolve around similarity of input vs. output.

We do not have existing legal authority on these issues, but this is likely to change in the near future in view of Getty Images v Stability AI [2023] EWHC 3090, a claim which is currently in the process of being litigated in the High Court. The Getty case will be of particular interest as it raises IP right infringement issues around input as well as output.

3. Patents

There is a question of whether the outputs of AI can be patented. Patents fall within the so-called registered rights as they are granted on application to the UKIPO. Patents provide the patent holder with an exclusive right over the invention, e.g., an exclusive right to a product or a process, for a period of time. In order to qualify for a patent, the product or process must be new, involve an inventive step, be capable of industrial application, and not specifically excluded from protection. In exchange for the patent grant, the applicant must disclose technical information about the invention to the public.

While issues around AI and copyright remain to be tested in the courts, we do have some legal authority in the area of patent law.

3.1 Can AI be an inventor?

The question of whether AI can be an inventor under current UK patent law has already been considered by the Supreme Court in the case of Thaler v Comptroller- General of Patents, Designs and Trade Marks [2023] UKSC 49. Dr Thaler filed two patent applications under the Patents Act 1977 (the "1977 Act") for inventions solely created by an AI system called DABUS of which Dr Thaler was the owner. The Hearing Officer for the Comptroller-General of Patents, Designs and Trade Marks (the "Comptroller") issued a decision that (i) DABUS could not be an inventor for the purposes of Sections 7 and 13 of the 1977 Act because it was not a person, and (ii) Dr Thaler was not entitled to a patent based on his ownership of DABUS in circumstances where DABUS was listed as the inventor. Dr Thaler appealed to the High Court and the Court of Appeal but both appeals were unsuccessful.

The Supreme Court decided that: i) an inventor within the meaning of the 1977 Act must be a natural person, and ii) that the doctrine of accession does not apply as this is not a case where new tangible property is produced by an existing item of tangible property. It follows that DABUS is not an inventor for the purposes of 1977 Act and the Act did not confer on Dr Thaler the property in or the right to apply for and obtain a patent for any technical development made by DABUS. Accordingly, the Comptroller was right to find Dr Thaler's applications as withdrawn under Section 13(2) of the Patents Act. The Supreme Court acknowledged that had it been Dr Thaler’s case that he was the inventor (rather than DABUS), and that he had used DABUS as a '*highly sophisticated tool*', the outcome of the proceedings '*might well have been different*'.

It is important to note that, right at the outset of the judgment, Lord Kitchin (with whom the other Lords agreed) made clear that *‘the appeal is not concerned with the broader question whether technical advances generated by machines acting autonomously and powered by AI should be patentable. Nor is it concerned with the question whether the meaning of the term “inventor” ought to be expanded, so far as necessary, to include* *machines powered by AI which generate new and non-obvious products and processes [48] … This appeal is concerned instead with the much more focused question of the correct interpretation and application of the relevant provisions of the 1977 Act to the applications made by Dr Thaler.'* [50] The court recognised that, in view of rapid advances in AI technology, these broader questions are increasingly important and alluded to a potential shift in the legal landscape as a result. However, in citing the judgment of Laing LJ in the Court of Appeal, the Supreme Court was clear that *'If patents are to be granted in respect of inventions made by machines, the 1977 Act will have to be amended'* [79].

Whether AI can be an inventor for the purposes of patent law has received global attention with Dr Thaler filing test patent applications in different jurisdictions around the world, including the European Patent Office (EPO). The current legal position on an international level appears to be that an inventor for patentable inventions must be a human or a person with legal capacity.

3.2 Can AI be patented?

The inventions in the Thaler case concerned a food container and a light beacon. It was undisputed that these were patentable, i.e., there was no issue around novelty, for example, and the inventions did not fall within categories that are excluded from patentability. What about the patentability of the AI system itself? Under Section 1(2)(c) of the Patents Act 1977 'a programme for a computer … as such' is excluded from patent protection. Essentially, the position is that one cannot obtain a patent for a computer programme in itself; however, if the computer programme provides a 'technical contribution' to the real world, then it is patentable. A recent decision of the High Court considered this statutory provision and associated caselaw in the context of AI in the case of Emotional Perception AI Ltd v Comptroller-General of Patents, Designs and Trade Marks [2023] EWHC 2948.

Emotional Perception had applied to patent an Artificial Neural Network (ANN). The ANN was said to be capable of providing improved media file recommendations. Taking the context of music websites for example, where a user wants to receive music similar to music they already have, traditional tools would recommend music tracks based on similar categories of music (e.g., rock), those categories having been tagged as such by humans. Instead of taking the 'category' of music as the criterion for recommending similar music tracks, the ANN-based system is said to identify similar music tracks based on human perception and emotion. In brief, the system works as follows: it takes a pair of music files, which have been given a semantic label, e.g., 'happy', 'relaxing', and so on. The files are plotted in a 'semantic space', with the distance between the files indicating their semantic similarity. In addition to their semantic properties, the files are also analysed according to physical properties such as tone, timbre, speed etc., and again plotted in a 'property space'. Using back-propagation, the property space is refined in order to reflect the semantic space, so that semantically similar tracks are close in property space, whereas semantically dissimilar tracks are farther apart in the property space. The operational ANN is then able to take a music track, determine its physical attributes, plot these against the physical attributes of other music tracks in a music library or database, and by looking for those tracks which are most proximate in terms of physical characteristics, it can recommend semantically similar tracks.

An officer for the UKIPO refused grant of the patent on the basis that the ANN system was considered to be 'a program for a computer' and that the patent application related to that computer programme 'as such'.

Emotional Perception appealed to the High Court challenging the decision by the UKIPO to refuse grant of the patent. The matter came before Sir Anthony Mann, J, who considered whether i) the ANN was 'a program for a computer' therefore falling within the statutory exclusions to patentability, and ii) if it was, whether there was a technical contribution which meant it fell outside the exclusionary regime.

On the first point, Mann J, differentiated between hardware ANNs and software emulated ANNs, and concluded that neither qualifies as 'a programme for a computer' and therefore neither was excluded from patentability. In the case of hardware ANNs, it was accepted by the parties that there is no 'programme' and therefore this would not fall within the exclusions. *'The hardware is not implementing a series of instructions pre-ordained by a human. It is operating according to something that it has learned itself.' [54]* In the case of software emulated ANNs, there were two aspects in which computer programming plays a role, one being the training stage, and the other being the software platform which enabled the computer to carry out the emulation. With regards to the latter, Mann, J considered that this can be de-coupled from the ANN: *'It seems to me that it is appropriate to look at the emulated ANN as, in substance, operating at a different level (albeit metaphorically) from the underlying software on the computer, and it is operating in the same way as the hardware ANN. If the latter is not operating a program then neither is the emulation.' [56]*

The court found that the ANN, in itself, was not a computer programme because it was not operating a set of programme instructions given to it by a human. The ANN had trained itself, applying its own weights and biases. It was emulating a piece of hardware which had physical nodes and layers, and was no more operating or applying a program than a hardware system was.

With respect to the computer programme involved at the training stage, which sets the training objectives and parameters in which the ANN is to operate, the court concluded that this fell outside the actual invention that is claimed. The invention was not a claim to the computer programme at the training stage; the invention related to the idea of using pairs of files for training, and setting the training objective and parameters accordingly. The claim therefore went beyond the actual computer programme.

As explained above, even if an invention were to be a claim to a computer program, it may still be patentable if it provides a "technical contribution" outside the computer program itself. Given his conclusion that the ANN was not a computer programme, Mann J, did not need to consider the question of technical contribution, but he nevertheless did. Following a review of caselaw on what constitutes a 'technical contribution', Mann J found that the sending of a file recommendation to an end user is a matter external to the computer and amounts to a technical contribution, i.e., the ANN has a real world effect outside of the computer.

So far, established practice at the UKIPO was to treat inventions involving AI as computer implemented and therefore applications would have had to be considered under the computer program exclusion exemption, i.e., whether the invention produced a technical contribution. The position in Europe has been similar with the EPO considering inventions involving AI as computer-implemented inventions which would only become patentable if they are applied to solve a technical problem in a field of technology.

The Emotional Perception AI judgment has potentially opened up a new avenue to obtain patent protection in the UK for inventions involving ANNs and AI more generally. We say 'potentially' as UKIPO is currently appealing the decision of the High Court. We will await to see whether the Court of Appeal, like the High Court, reaches a decision favourable to patentees of AI inventions. In case the High Court decision is upheld, it will be interesting to see what influence the Emotion Perception AI judgment will have on the approach taken by the EPO.

4. Conclusion

The training of AI technology has led to copyright disputes, and there are question marks over the IP of the outputs that result from generative AI (and the IP status of the AI itself). It is quite easy to see how the cases and discussions drawn on in this paper extend to LLMs. While we await further development and resolution in these cases and discussions, this paper has aimed to put a spotlight on the issues that could feasibly arise for LLM stakeholders going forward.

5. References

5.1 Academic

Emanuilov, I., & Margoni, T. (2024). Forget me not: memorisation in generative sequence models trained on open source licensed code. DOI: 10.5281/zenodo.10635479

Guadamuz, A. (2024). A Scanner Darkly: Copyright liability and exceptions in Artificial Intelligence inputs and outputs. GRUR International. 73(2) pp. 111-127 DOI: 10.1093/grurint/ikad140.

5.2 Legislation and Caselaw

Copyright, Designs and Patents Act 1988

Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market

Emotional Perception AI Ltd v Comptroller-General of Patents, Designs and Trade Marks [2023] EWHC 2948

Getty Images v Stability AI [2023] EWHC 3090

Getty Images (US), Inc. v. Stability AI, Inc., 1:23-cv-00135

Infopaq International A/S v Danske Dagblades Forening [2009] ECJ Case C-5/08, ECLI:EU:C:2009:465

Patents Act 1977

Thaler v Comptroller- General of Patents, Designs and Trade Marks [2023] UKSC 49

5.3 Guidance

Communications and Digital Committee Large language models and generative AI. 1st Report of Session 2023-24. HL Paper 54. 2nd February 2024. URL: https://publications.parliament.uk/pa/ld5804/ldselect/ldcomm/54/5402.htm

Department for Science, Innovation & Technology. Consultation Outcome: A pro-innovation approach to AI regulation: Government response. Command paper CP1019. 6th February 2024. URL: https://www.gov.uk/government/consultations/ai-regulation-a-pro-innovation-approach-policy-proposals/outcome/a-pro-innovation-approach-to-ai-regulation-government-response

Open AI Written Evidence to HL Paper 54. URL: https://committees.parliament.uk/writtenevidence/126981/html/

UK Intellectual Property Office. (2023). The governments code of practice on copyright and AI. URL: https://www.gov.uk/guidance/the-governments-code-of-practice-on-copyright-and-ai

Ethical Issues in Language Resources and Language Technology – New Challenges, New Perspectives

Paweł Kamocki, Andreas Witt
Leibniz-Institut für Deutsche Sprache
R5 6-13, 68161 Mannheim, Germany
kamocki | witt@ids-mannheim.de

Abstract

This article elaborates on the author's contribution to the previous edition of the LREC conference, in which they proposed a tentative taxonomy of ethical issues that affect Language Resources (LRs) and Language Technology (LT) at the various stages of their lifecycle (conception, creation, use and evaluation). The proposed taxonomy was built around the following ethical principles: Privacy, Property, Equality, Transparency and Freedom.

In this article, the authors would like to: 1) examine whether and how this taxonomy stood the test of time, in light of the recent developments in the legal framework and popularisation of Large Language Models (LLMs); 2) provide some details and a tentative checklist on how the taxonomy can be applied in practice; and 3) develop the taxonomy by adding new principles (Accountability; Risk Anticipation and Limitation; Reliability and Limited Confidence), to address the technological developments in LLMs and the upcoming Artificial Intelligence Act.

Keywords: ethics, generative AI, privacy

1. Introduction

In our contribution to the previous edition of the LREC Conference (Kamocki, Witt, 2022), we proposed a tentative taxonomy of ethical issues affecting Language Resources (LRs) and Language Technology (LT) tools throughout their entire lifecycle, built around the principles of Privacy, Property, Equality, Transparency and Freedom. In this article, we would like to elaborate on this idea.

1.1 Ethical Norms over Time

It is a tempting perspective to think of ethical norms as something universal and perfectly static, i.e. not changing over time. The proponents of this view on ethics would use the Decalogue as an example: formulated millenia ago (probably around 7th century BC), the Ten Commandments are still the cornerstone of ethics and a foundation of (not only Western) civilisation. The argument, however, is inherently flawed, as the biblical version of the 10th commandment ('*Thou shalt not covet thy neighbour's house, thou shalt not covet thy neighbour's wife, nor his manservant, nor his maidservant, nor his ox, nor his ass, nor any thing that is thy neighbor's*'. – Exodus 20:17) is nowadays more generally known in its simplified version: You shall not covet. The reasons behind this are certainly not purely mnemonic; rather, in today's world wives are not considered property, slavery had been abolished, and oxen and donkeys are not generally seen as particularly desirable items (compared to, for example, high-end laptops, luxury watches or electric cars). The world has changed, and ethical norms, even as fundamental as the Ten Commandments, had to be adapted to the new reality.

It is therefore not surprising that ethical guidelines concerning something as dynamic as LT and LRs also need to change, and quite often. The taxonomy proposed in our previous contribution should be revised and, if necessary, completed.

1.2 Changes since 2022

The two years that passed since our original proposal seem to be a very short period of time, even in the evolving field of LT and LRs. However, some important developments have taken place during that time.

Most importantly, since the launch of Chat GPT in late 2022, LLMs have attracted a lot of public attention. Before that date, LLMs were mostly regarded as a useful tool in applications such as Machine Translation or Speech Recognition, but few were able to predict that LLMs will become independent tools, and almost ontologically independent beings. The debate on ethical implications of LLMs is now present in mainstream media (e.g. Metz, Weise, 2023), with reports on individuals having romantic relations with language models, or even committing suicide under their influence (Xiang, 2023).

In this unprecedented context, the EU Artificial Intelligence (AI) Act is taking shape (European Commission, 2021); it is expected to be soon, and become applicable in 2024. The AI Act already existed as a draft in 2022, but seemed, we admit it, rather far removed from LT and LRs, focused instead on such applications of AI systems as biometric identification, law enforcement and administration of justice. However, ChatGPT has revolutionised the perception of LLMs, which now may seem qualifiable as high-impact AI systems. Such systems are heavily regulated by the draft AI Act; before adoption, the draft is expected to be substantially modified in such a way as to regulate LLMs (and other foundation models) even more (Volpicelli, 2023; Zenner, 2023).

As explained in our previous contribution, while we agree that ethical norms are distinct in nature from legal norms, we also believe that the two systems – law and ethics – affect each other. This mutual influence is particularly visible in the field of new technologies, as laypersons, usually unable to comprehend the functioning of these technologies and their underlying principles, are more inclined to perceive them as ‘evil’ or ‘immoral’ when they are prohibited or restricted by law. This is why, in our opinion, the AI Act, even *in statu nascendi*, has an impact on LR and LT ethics.

1.3 Continous Relevance of the Proposed Taxonomy

All the above does not mean that the taxonomy we proposed in 2022 is now outdated. Quite the contrary, we stand by all the principles we formulated in our previous contribution, i.e. Privacy, Property, Equality, Transparency and Freedom. At the same time, we admit that some of these principles, and especially the principle of Equality, is increasingly unclear and difficult to apply in the context of LLMs and generative AI. Nevertheless, and perhaps all the more so, it should remain in sight throughout the entire lifecycle of an LR or an LT tool.

2. Overview of Ethical Issues throughout the LR or LT tool Lifecycle

In providing an overview of the ethical issues affecting LRs and LT, we will follow the same structure as in our previous contribution, dividing their lifecycle into four stages: conception, construction, use and evaluation. In this contribution, we would include a tentative checklist with questions that need to be addressed at every stage.

2.1 Conception Phase

Already at the earliest stage, i.e. the conception phase (before any data collection), certain ethical considerations need to be addressed. The questions that should be asked include:

1. **Under whose responsibility is the tool or resource developed?**

This fundamental question may be overlooked in joint research projects or public-private partnerships, as it is not always well-integrated in data-intensive technology research (Wagner, 2020). Although it may seem as a legal (more than ethical) question (cf. Article 5(2) of the GDPR), it should, in our view, precede any legal analysis. The legal situation (regarding responsibility, intellectual property, warranties…) should be modelled by a contract, or a series thereof, based on the answer to this question, which can also be formulated as: if things go wrong, who is to take the blame? This person (or, more often, this entity), would then be morally obliged to minimise the associated risks, and should be given the organisational (and legal) tools to do so. With great power comes great responsibility, and, ideally, *vice versa*.

This question is also essential to address the principle of accountability, which is a basic principle of the GDPR, but also of the AI Act, and one of the OECD Principles for responsible stewardship of trustworthy AI (OECD, 2019).

In practice, responsibility can be limited to specific tasks: for example, one entity can be responsible for training a model, another one for developing an application based on that model, and yet another one for commercialising it. However, situations where responsibility is thinly spread should generally be avoided, and whenever possible, responsibility should be concentrated in as small a number of entities as possible.

2. **What is the intended purpose of the tool or resource? What are its potential uses and foreseeable misuses?**

Although sometimes the purpose might be difficult to grasp (some resources or tools can initially be general-purpose, and then shift to a more specific application, or vice versa), this question still helps to anticipate the associated risks. A resource intended for researchers (e.g. a corpus of 17th century theatrical plays) is held to a lower standard of risk anticipation than, e.g., a chatbot intended to assist air traffic controllers, as the potential harm caused by malfunctioning or misuse is much higher in case of the latter.

Defining the intended use is also instrumental in assessing the reliability of the tool or resource – it is reliable if it performs its main task in a way that is both reasonably accurate and proportionate. Accuracy in this context means that the output does not contain false information; proportionality: that the output does not contain more or less information than reasonably needed or expected. For example, a chatbot intended to provide passengers with information about train schedules is accurate if it provides information about existing trains only (not information that is out-of-date, or, worse, that is ‘hallucinated’, like an imaginary direct train connection between Prague and Oxford). It is proportionate if it provides relevant information such as time of departure, expected perturbations and a crowd forecast; it is disproportionate if it omits to provide essential information (e.g., departure time) or if it provides irrelevant information (e.g., the number of the seat next to an unaccompanied teenager, or the name of the conductor).

As per the AI Act, it is also necessary to consider ‘foreseeable misuses’, as they play a prominent role in risk assessment. For example, the abovementioned corpus of 17th century theatrical plays can hardly be misused, whereas a speech synthesis tool may potentially be misused by minors to circumvent age restrictions (by making them sound as adults on the phone). This brings us to the next question, i.e.,:

3. **What are the intended user groups?**

The intended users are, of course, closely linked to the intended use, at least *prima facie*. In particular, it should be considered here whether the tool or resource can be made available to minors, or other groups that deserve special protection (people with disabilities, elderly people, refugees), in which case a higher standard of risk anticipation and management should be applied.

Attention needs to be paid to tools and resources that, although primarily intended for a narrow group of users (e.g. university researchers), are going to be made openly available, to satisfy the requirements of the Open Science movement. Open availability means that the tool may also be used by groups like minors or convicted criminals. This should be taken into account on the one hand in risk anticipation and management (see above about foreseeable misuses), and on the other hand in the decision whether the resource or tool should actually be made openly available or not. We believe that in many cases ethical considerations related to risk mitigation should prevail over Open Science ideals. After all, FAIR data should be as open as possible, but also as closed as necessary (Landi et al., 2020).

4. **What is the potential impact of the tool or resource on the users?**

This question seems closely related to the two previous ones. However, one should consider the impact not only of the intended use by the target user groups, but also more broadly: what is the worst possible scenario involving the tool? Can it harm the user in any way, by its normal functioning or malfunctioning? How likely is this worst scenario to happen? How can the risk of it happening be further reduced?

Of course, we are aware of the fact that few human activities are completely risk-free – one can be killed or seriously injured while turning the light on, if several factors coincide (e.g. wet hands, bare feet and faulty electric installation). This does not mean that light switches should be banned, or only made available to trained professionals, or that an average user should be constantly reminded about the risk of being electrocuted while operating a switch. However, while in the presence of a relatively new technology, such as a very large language model (as opposed to a known and tested piece technology, as a light switch), any non-negligible risks should be carefully considered, anticipated and mitigated, and the users warned about their existence.

The 'impact on users', as discussed here, includes impact on their privacy, understood both as *'freedom from unauthorised intrusion'* and as a *'right to keep one's personal matters and relationships secret'.* This is related to the GDPR principle of privacy by design (Kamocki, Witt 2020), and discussed at length in our previous contribution (Kamocki, Witt 2022).

Moreover, a tool or resource that is ill-balanced since the conception phase would disregard the principle of Equality (also described in our previous contribution), and have a negative impact on some users.

The answers to all the above questions should be thoroughly documented and made available to users in an appropriate form, in the spirit of the fundamental principle of Transparency (OECD, 2019; Kamocki, Witt 2022).

2.2 Construction Phase

The construction phase contains for the most part of data collection and preparation (annotation, etc.). At this stage, the following questions should be taken into account:

1. **Are the data (and other material) subject to (intellectual) property rights?**

This question is directly related to the principle of Property, which we elaborated upon in our previous contribution (Kamocki, Witt, 2022). Even though today many researchers decide to openly share their data and code, in the spirit of Open Science, this does not change the fact that language data and software code are (almost always) protected by copyright, which means they can only be copied and shared with the right holder's permission or under a statutory exception. In the EU, such exceptions for Text and Data Mining (whether carried out for research purposes or for any other purpose) were introduced in the 2019 Directive on copyright in the Digital Single Market. In the US, to the best of our knowledge, the fair use doctrine allows for a wide range of uses related to research and new technologies in general.

In any case, it is important to decide whether the data can be used (e.g. scraped from the Internet) on the basis of an exception, or whether a permission (licence) should be obtained. Needless to say, the decision should be grounded in a thorough legal analysis, and properly documented.

2. **Are privacy-sensitive data used? If so, is the concerned individual allowed to opt-out?**

We use the term 'privacy-sensitive data' instead of 'personal data' for a good reason: as explained in our previous contribution, we want to examine the issue of privacy not from the legal perspective, but from a broader, ethical one.

In general, it seems to us that in most cases providing the person whose interests (privacy or others) are affected by the data the right to opt-out by withdrawing 'their data' from the processing – even if such an opt-out is not required by law – is the best way to address 'data sensitivity'. However, in certain situations an opt-out may require a delicate balance of interests, as opt-out by one individual can negatively affect another one. In a somewhat simplistic example, if Team A loses in competition with Team B, Team A can argue that this information negatively affects its members, who might be seen as less competent. Withdrawal of this information from the processing, however, would negatively impact the interests of the winning Team B. The opt-out request, in such circumstances, should not be acted upon.

Specific consideration should be given to the re-use of user input data for further development of the LR or LT tool (e.g., for training an underlying language model). If users are given complete freedom as to the types of data they can input, their data should not by default be re-used for such purposes – rather, they should be given a possibility to opt-in (and then opt-out at a later stage, if they change their mind). In some specific applications, however, this optic can be reversed, if the balance of interests justifies it. For example, if the user input is of low sensitivity (e.g. a query history in a corpus of 17th century theatrical plays), and it can be used to significantly improve the resource or tool, the re-use should be a default, and an opt-out request should only be acted upon if it is well-justified.

2.3 Use Phase

To comply with ethical requirements, LRs and LT tools should also be used responsibly. The questions that any user should ask themself include:

1. **Is the resource or tool suitable for the envisaged use?**

This question is particularly important in the context of generative AI tools, such as Chat GPT. A responsible user should be aware of the drawbacks of AI-generated data, such as the fact that they under-represent dialects and other local specificities of a given language. Furthermore, AI-generated data, if used for training AI tools (which is not uncommon in practice, as a sizable portion of texts on the Internet is likely to be machine-generated) can only reinforce their own shortcomings and create a negative feedback loop. Finally, the use of AI-generated data comes with a risk of overlooking the ‘human factor’, which in certain scenarios is particularly undesirable (e.g., in tools intended for some form of emotional support).

Furthermore, it is important that the user be transparent about the use of AI-generated data.

Regarding all sorts of LR and LT outputs, the user should maintain some ‘healthy scepticism’, rather than blindly rely on the results. LT, even the most advanced (or: especially the most advanced) is known for occasionally ‘hallucinating’ (Alkaissi, McFarlane, 2023), i.e. producing a credible output that is not based on any real-world input. For example, when asked to generate a bibliography for a scientific article, Chat GPT is likely to provide references that look credible *prima facie*, but do not correspond to any real publications (Walters, Wilder, 2023). Such outputs, before they can be used, should be manually verified or cross-referenced with a credible source (e.g., Google Scholar).

2. **Do I know the conditions (terms) of use of the resource or tool?**

It is easy to lose sight of the fact that some LR and LT tools come with conditions (or terms) of use. This is the case of ChatGPT, available only via a dedicated API. These terms of use may prohibit certain uses of the tool or resource, or related outputs. For example, the Open AI’s Terms of use prohibit the use of outputs from their services (like Chat GPT) to develop competing models (Open AI, 2023).

We believe that the respect of such conditions is also an ethical requirement, as they are rooted in the principle of property, even if not based directly on an existing Intellectual Property right.

2.4 Evaluation Phase

In the evaluation phase, reliability of the tool or resource, as well as continuous risk management, seem to be primary concerns. Therefore, the following questions, similar to the one asked at the conception phase, should be answered here:

1. **Who is the person or entity responsible for the tool? Has it changed since the conception phase?**
2. **Is the purpose for which the tool or resource is being or can be used different from its initial purpose?**
3. **Are the actual users of the tool the same as those for whom the tool was initially intended?**
4. **Given the answers to the questions above, what is the potential impact of the tool or resource, as it is used now, on its current users?**

All these questions reflect one idea: the context in which the tool is used can evolve, which requires a new risk assessment. Such a review should be carried out periodically.

3. Ethical Principles for LR and LT

Based on the analysis above, we would like to propose the following list of ethical principles for LR and LT:

1. **Privacy**: stakeholders should be protected against disproportionate intrusion and allowed to keep certain information secret;
2. **Property**: intellectual and cultural property should be handled with respect, in compliance with applicable law, ensuring that any potential harm (evaluated from the owner’s perspective) is outweighed by collective benefit;
3. **Equality**: no group of stakeholders or contributors should be directly or indirectly discriminated against;
4. **Transparency**: LT outputs should be clearly marked as such; stakeholders should be informed about the main principles of, and given a possibility to learn the details about the functioning of LT;
5. **Freedom**: data providers should be free to contribute their data to LR&LT, and, to a reasonably practicable extent, to change their mind at any later stage; human intervention should be necessary and decisive in any process involving the use of LT the outcome of which may seriously impact the user;
6. **Accountability:** the person(s) or entity(-ies) responsible for the resource or tool at different stages of its creation should be clearly identified. The accountability should not be unnecessarily distributed across too many stakeholders;

7. **Risk Anticipation and Mitigation:** any risks related to the use or foreseeable misuse of a LR or LT tool, taking into account its actual use and actual user groups, should be anticipated and, if necessary, mitigated by employing appropriate measures;
8. **Reliability and Limited Confidence:** these principles are like two sides of one coin: a) LRs and LT tools should be built in such a way as to be fit for their intended purpose (Reliability) and b) any results produced with LRs and LT tools should be met with limited confidence and, if appropriate, verified (Limited confidence).

Principles 1-5 restate those that we proposed in our previous contribution. Principles 6-8, which are of more over-arching nature, constitute an original input of this article.

4. Conclusion

Since the last edition of the LREC conference, the debate on ethical issues affecting LRs and LT tools has intensified. Since ethical norms are not a static system, the guiding ethical principles for our field should be periodically revised, to ensure that they maintain their validity and do not become detached from the reality of the field.

In this contribution, we proposed a “checklist”, a list of questions that should be examined at various stages of development of an LR or an LT tool. We do hope it will help in ethics assessments by the data providers, the developers, the users, the evaluators, and potentially even the funders. We also proposed three new guiding principles, which are not intended to replace the principles we previously proposed, but rather to reinforce them by introducing a larger, more overarching perspective on ethics. These new principles are: Accountability, Risk Anticipation and Mitigation, and Reliability and Limited Confidence.

The debate on ethics in our field is bound to continue, and we do hope that this contribution will help structure it, at the very least by proposing a common terminology.

5. Bibliographical References

Alkaissi H., McFarlane S.I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. 2023 Feb 19;15(2):e35179

European Commission (2021). *Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts*. COM/2021/206 final.

Landi, A., Thompson, M., Giannuzzi, V., Bonifazi, F., Labastida, I., Bonino da Silva Santos, L. O., Roos, M. (2020). The “A” of FAIR – As Open as Possible, as Closed as Necessary. *Data Intelligence* 2020; 2 (1-2): 47–55.

Metz, C. and Weise, K. (2023). How ‘A.I. Agents’ That Roam the Internet Could One Day Replace Workers. *The New York Times*, 16 October 2023.

OECD (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449.

Open AI (2023). *Terms of Use*. Updated 13 March 2023. Available at: https://openai.com/policies/terms-of-use (access: 20.10.2023)

Volpicelli, G. (2023). ChatGPT broke the EU plan to regulate AI. *Politico*, 3 March 2023.

Wagner, B. (2023). Accountability by design in technology research. *Computer Law & Security Review*, vol. 37, July 2020.

Walters, W.H., Wilder, E.I (2023). Fabrication and errors in the bibliographic citations generated by ChatGPT. *Scientific Reports* 13, 14045 (2023).

Xiang, Ch. (2023). 'He Would Still Be Here': Man Dies by Suicide After Talking with AI Chatbot, Widow Says. *Vice*, 30 March 2023.

Kamocki, P. and Witt, A. (2020). Privacy by Design and Language Resources. In *Proceedings of the LREC 2020*.

Kamocki, P. and Witt, A. (2022). Ethical Issues in Language Resources and Language Technology – Tentative Categorisation. In *Proceedings of the LREC 2022*.

Zenner, K. (2023). A law for foundation models: the EU AI Act can improve regulation for fairer competition. *OECD AI Policy Observatory*, 20 July 2023.

Legal and Ethical Considerations that Hinder the Use of LLMs in a Finnish Institution of Higher Education

Mika Hämäläinen
Metropolia University of Applied Sciences
Helsinki, Finland
mika.hamalainen@metropolia.fi

Abstract

Large language models (LLMs) make it possible to solve many business problems easier than ever before. However, embracing LLMs in an organization may be slowed down due to ethical and legal considerations. In this paper, we will describe some of these issues we have faced at our university while developing university-level NLP tools to empower teaching and study planning. The identified issues touch upon topics such as GDPR, copyright, user account management and fear towards the new technology.

Keywords: GDPR, privacy, copyright, GenAI, policies, AI ethics, LLM

1. Introduction

In this paper, we will describe our practical experience in building university-level NLP solutions in Metropolia University of Applied Sciences in Finland. We are developing two tools, a Moodle plugin and a tool for planning curricula, both of which rely heavily on Vertex AI[1] (see Kharlashkin et al. 2024), which lets us use PaLM 2 (Anil et al., 2023), a Large Language Model (LLM) provided by Google.

While Vertex AI makes the development process easy, it is not free of legal and ethical hurdles. What makes maters more difficult are the rigid organizational practices related to data safety and user account control that do not scale to the requirements of modern AI.

In 2023, GenAI emerged seemingly out of nowhere (see García-Peñalvo and Vázquez-Ingelmo 2023), at least this was the case for people who were outside of the scientific discipline of NLP. The IT departments of many organizations have been caught off guard with great many staff members asking for ready-made tools like ChatGPT or Copilot (cf. Uren and Edwards 2023). Even these off-the-shelf tools cannot be taken into use in a big organization without planning, let alone custom-made NLP solutions the likes of which we are developing.

This paper presents our experience on the challenges we have faced while striving for a usable AI solution that can provide teachers and study planners alike with the added value of NLP based analysis and content creation.

2. Why can we not host an LLM locally?

The simplest solution to most of our legal issues would be hosting our own LLM instead of relying on Vertex AI. With the fast development of open LLMs (Groeneveld et al., 2024; Chen et al., 2024; Penedo et al., 2023) and their remarkable benchmark performance, together with their availability through tools like Hugging Face Transformers (Wolf et al., 2020), makes this question a valid one that requires serious consideration.

The issue we have with this is not only related to the computational requirements such models have, but it is more related to the fact that our tools must be able to understand and generate Finnish. Based on our trials, many of the open models either do not support Finnish at all or they struggle with the Finnish grammar. All models we have tried and that do support Finnish fail to produce grammatical Finnish. The commercial models PaLM 2 and ChatGPT are capable of producing mostly fluent Finnish due to their sheer size both in terms of training data and parameters. However, even these models make occasional mistakes.

There is one open LLM under development that is tailored for the Finnish language. This LLM is called Poro[2], but, at the time of writing, the model is not yet fully trained and it has not been fine-tuned to handle ChatGPT-like prompting. This means that, for the time being, our only viable solution is to use either PaLM 2 or GPT-4.

[1] https://cloud.google.com/vertex-ai?hl=en

[2] https://huggingface.co/LumiOpen/Poro-34B

LEGAL2024 @LREC-COLING-2024, pages 24–27
20 May, 2024.

3. GDPR and preventing personal data leak

Because we are bound to use an LLM provided by an external provider, the first question that we, as an organization, need to tackle is the GDPR law[3]. OpenAI has already faced legal troubles in one of the member states, namely Italy[4], due to their failure to meet the regulatory requirements on data protection as established by the GDPR law.

This unfortunate situation only leaves us with one alternative: Google Cloud. According to their terms and conditions, Google both promises that our data will not be use for training their LLM[5] and that the output of their model will not violate the copyrights of any author[6]. Of course, it is impossible for us to know the degree to which this is true, but they are the only option we have.

Our Moodle plugin is designed to analyze teachers' slides using an LLM and provide teachers with useful feedback on how to incorporate sustainable development goals into their teaching, what kind of assignments they might give and so on. As slides may very well contain personal information such as students' names or email addresses, we need to anonymize them locally before analyzing them using Vertex AI.

Our anonymization is as simple as running a Finnish named entity recognition model (Luoma et al., 2020) trained on the Finnish BERT model (Virtanen et al., 2019). We use the entity tag "PERSON" to remove all names from the slides. In addition, we use regular expressions to remove emails and phone numbers. We see that this is a necessary step in protecting the privacy of our students and staff regardless of what Google promises us in their terms of service. This approach, however, is not free of problems. Many slides will have citations to authors, which will get removed as well because they are recognized as names.

4. Copyright, work contract and teachers' rights

The first line of organizational resistance we faced was the question of copyrights. Can we analyze teachers' slides using Vertex AI without violating their copyrights? This might sound strange given that teachers are members of our staff. According to the Finnish copyright law[7], copyright is something a person will have when they produce something that is copyrightable. This means that an organization, such as a university, will not automatically get copyrights to the creative work of its staff.

Our organization has a statement of copyright transfer in new work contracts, but this statement was absent in older work contracts. This means that we have several lecturers and senior lecturers who have never given the university any rights for their copyright protected material, such as course slides.

Copyright only protects the form, not the idea behind any creative work. This means that we can, legally, use an LLM to analyze copyrighted material for as long as we don't reproduce that material to a significant degree. Because GenAI as such a new thing and the wording in the Finnish copyright law has never been meant to cover such a use-case, we have taken a more ethical approach and opted for protecting the copyrights more than what the law protects.

In practice, this means that we will not analyze teachers' materials automatically, but we instead let the teachers decide which slides they want to analyze and let them be in charge of deciding whether they want to even use our NLP tool or not.

5. User account management and access rights

Another issue that AI brings to the table is that of access rights. The question of what type of data can be passed to an existing solution like Copilot or ChatGPT is one part of the discussion. Another, larger part is the question whether the organization has the user rights management on such a nuanced level that AI tools can be given only the rights they need and nothing more.

Moodle makes it easy to create an "AI user" that can only access slides. However, our university also uses another learning environment named Peppi[8], which does not have any nuanced access right management. A user can either have access to everything in the system, including student's grades, essays and so on, or individual student or teacher access for a given course.

An individual access means that the NLP tool should be added manually to each course. Even this would not solve this issue because the AI would need to have a teacher access to be able to access slides before they are made available for the students. Teacher rights also entail access to students' grades. Even though we are developing the NLP

[3] https://www.consilium.europa.eu/en/policies/data-protection/data-protection-regulation/

[4] https://techcrunch.com/2024/01/29/chatgpt-italy-gdpr-notification/

[5] https://cloud.google.com/terms/data-processing-addendum/

[6] https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification

[7] https://www.finlex.fi/fi/laki/ajantasa/1961/19610404

[8] https://www.peppi-konsortio.fi

tool by ourselves, we do not want to give rights to grades to any additional systems because this will increase the attack surface (see Schuster and Holz 2013) and potential leak of data.

6. Fear or ethical thinking?

AI ethics is a serious consideration, and there is a growing body of work that highlights issues such as bias on gender (Shrestha and Das, 2022), ethnicity (Garrido-Muñoz et al., 2021) and sexual orientation (Felkner et al., 2022). Especially in the education sector, attempts to grade students or profile them fully automatically are not ethically sustainable. Our goal is not to diminish any of these real ethical considerations, but to highlight the difference between these AI expert driven ethical considerations and those of non-technical people.

We have faced that "dropping the ethics bomb" is a way for certain actors within the organization to hinder the organization from embracing AI. Many times people cannot even verbalize what the exact ethical problems are, but oftentimes it is fear that tools like ChatGPT might produce erroneus output. Or, in the case of image generation systems, such as Adobe Firefly, that using such machine generated images is somewhat morally wrong. Curiously, the teachers who are most strongly against GenAI, are also the ones that demand that all student work be passed through a GenAI detector. Using ChatGPT is seen as unethical, but failing students because of an AI detector's analysis is seen as ethical. Despite the fact that recognizing AI generated text is not an accurate practice (Chaka, 2023).

Much of the fear towards the new technology is thus masked as an ethical consideration. The way non-technical people approach AI ethics in our organization is strikingly different from the way NLP researchers approach the problem.

7. Conclusions

In this paper, I have presented some of the legal and ethical issues we have faced in our organization when embracing LLMs. Despite these problems, many members of the staff are eager about the possibilities our NLP tools bring to teaching and study planning. However, the road to production is long and bureaucratic.

8. Bibliographical References

Chaka Chaka. 2023. Detecting ai content in responses generated by chatgpt, youchat, and chatsonic: The case of five ai content detection tools. *Journal of Applied Learning and Teaching*, 6(2).

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv*.

Virginia K Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2022. Towards winoqueer: Developing a benchmark for anti-queer bias in large language models. *arXiv preprint arXiv:2206.11484*.

Francisco García-Peñalvo and Andrea Vázquez-Ingelmo. 2023. What do we mean by genai? a systematic mapping of the evolution, trends, and techniques involved in generative ai.

Ismael Garrido-Muñoz, Arturo Montejo-Ráez, Fernando Martínez-Santiago, and L Alfonso Ureña-López. 2021. A survey on bias in deep nlp. *Applied Sciences*, 11(7):3184.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *Preprint*.

Lev Kharlashkin, Melany Macias, Leo Huovinen, and Mika Hämäläinen. 2024. Predicting sustainable development goals using course descriptions–from llms to conventional foundation models. *arXiv e-prints*, pages arXiv–2402.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Felix Schuster and Thorsten Holz. 2013. Towards reducing the attack surface of software backdoors. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*, pages 851–862.

Sunny Shrestha and Sanchari Das. 2022. Exploring gender biases in ml and ai academic research through systematic literature review. *Frontiers in artificial intelligence*, 5:976838.

Victoria Uren and John S Edwards. 2023. Technology readiness and the organizational journey towards ai adoption: An empirical study. *International Journal of Information Management*, 68:102588.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish. *arXiv preprint arXiv:1912.07076*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

9. Language Resource References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Jouni Luoma, Miika Oinonen, Maria Pyykönen, Veronika Laippala, and Sampo Pyysalo. 2020. A broad-coverage corpus for Finnish named entity recognition. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4615–4624.

Implications of Regulations on Large Generative AI Models in the *Super-Election Year* and the Impact on Disinformation

Vera Schmitt[1,5], Aljoscha Burchardt[5], Jakob Tesch[2], Eva Lopez[3], Salar Mohtaj[1,5]
Konstanze Neumann[4], Tim Polzehl[5] and Sebastian Möller[1,5]
Technische Universität Berlin[1], Ubermetrics GmbH[2], Deutsche Welle[3], delphai GmbH[4];
German Research Center for Artificial Intelligence[5]
{vera.schmitt, sebastian.moeller}@tu-berlin.de, jakob.tesch@ubermetrics-technologies.com,
eva.lopez@dw.com, {aljoscha.burchardt, salar.mohtaj, tim.polzehl}@dfki.de, konstanze@delphai.com

Abstract

With the rise of Large Generative AI Models (LGAIMs), disinformation online has become more concerning than ever before. Within the *super-election year* 2024, the influence of mis- and disinformation can severely influence public opinion. To combat the increasing amount of disinformation online, humans need to be supported by AI-based tools to increase the effectiveness of detecting false content. This paper examines the critical intersection of the AI Act with the deployment of LGAIMs for disinformation detection and the implications from research, deployer, and the user's perspective. The utilization of LGAIMs for disinformation detection falls under the high-risk category defined in the AI Act, leading to several obligations that need to be followed after the enforcement of the AI Act. Among others, the obligations include risk management, transparency, and human oversight which pose the challenge of finding adequate technical interpretations. Furthermore, the paper articulates the necessity for clear guidelines and standards that enable the effective, ethical, and legally compliant use of AI. The paper contributes to the discourse on balancing technological advancement with ethical and legal imperatives, advocating for a collaborative approach to utilizing LGAIMs in safeguarding information integrity and fostering trust in digital ecosystems.

Keywords: AI Act, DSA, LGAIMs, disinformation

1. Introduction

The World Economic Forum's *Global Risks Report 2024* (Forum, 2024) identifies Artificial Intelligence (AI)-generated disinformation as the second most critical risk, potentially causing significant global crises. In the context of 2024, a year witnessing over 70 elections globally, including major elections such as the U.S. presidential election, India's general elections, and the European Parliament elections, there is increasing concern about the profound influence that AI-generated content may have (Iskandar et al., 2023). In recent years, the field of generative AI has seen impressive advancements. Models such as *ChatGPT-4* (OpenAI, 2023) for text generation, *DALL·E 3* (Nguyen et al., 2024), and Sora (Brooks et al., 2024) for visual content creation, and *Whisper* (Radford et al., 2022) for voice cloning have undergone significant improvements. The AI models discussed in this paper are commonly known as *Foundation Models*, *Large Language Models* (LLMs), or *Large Generative AI Models* (LGAIMs) (Hoffmann et al., 2022) — the terminology we have chosen to use here. LGAIMs have advanced to a stage where they are user-friendly and do not demand deep technical know-how for their utilization. LGAIMs have the potential to liberate professionals to concentrate on important tasks, like direct patient care, potentially leading to a more efficient and fairer distribution of resources (Hacker et al., 2023). As a result, AI is becoming more embedded in our everyday experiences, significantly influencing the transformation of the digital environment, especially with its role in the creation of disinformation. The capacity to generate disinformation, create deepfakes, and disseminate hate speech has greatly escalated, posing serious threats to the integrity of information ecosystems (Hacker et al., 2023; Simon et al., 2023; Longoni et al., 2022; Khamsehashari et al., 2023). The *infodemic* experienced during the Covid-19 pandemic (Balakrishnan et al., 2022) and the military conflicts in Ukraine and Israel (Darwish et al., 2023) exemplify the significant influence that LGAIMs can wield in producing disinformation and shaping public opinion (Monsees, 2023; Satariano and Mozur, 2023). In light of these developments, the impending AI Act, aimed at regulating the use and deployment of AI technologies, assumes critical importance. In the disinformation context, LGAIMs can be used for both generating and detecting mis- and disinformation. The AI Act offers a framework to assess the risk of AI systems and defines obligations depending on the risk category in which the AI system falls. However, the risk assessment and the associated obligations are sometimes not straightforward. Especially when transferring the often generally formulated obligations to concrete technical implementations, much freedom is given concerning the concrete design scope and technical interpretation

LEGAL2024 @LREC-COLING-2024, pages 28–38
20 May, 2024.

of the obligations defined. From the researcher's perspective, the AI Act presents both opportunities and constraints. It offers a structured environment for ethical AI research, emphasizing the need for responsible innovation and the importance of addressing AI's societal impacts. The AI Act offers so-called *Sandboxes*, allowing the fostering of research and innovation. From the deployer's perspective, the Act is a double-edged sword. On the one hand, it offers a much-needed framework for ethical and transparent AI deployment, ensuring that AI technologies are used responsibly and transparently. Deployers must navigate the complex landscape of compliance, grappling with the challenges of integrating ethical considerations into their AI systems without hindering innovation. On the other hand, the Act represents a significant compliance challenge, with stringent regulations potentially hindering the pace of AI development and deployment. From the user's perspective, the AI Act is beneficial in ensuring user rights and safety in the digital age. It promises to safeguard users from the risks associated with AI-generated disinformation, deepfakes, and other forms of digital manipulation. By setting clear standards for transparency, accountability, and reliability, the Act aims to foster trust in AI technologies, enabling users to benefit from AI advancements while being protected from their potential harms. However, the effectiveness of the AI Act depends heavily on its effectiveness in enforcement. When considering the shortcomings of the General Data Protection Directive (GDPR), the enforcement was and is still the major hurdle (see (Schmitt et al., 2023)), where there is no control of GDPR compliance on a technical level as long as there is no complaint from a user. If the enforcement of the AI Act repeats similar mistakes, it will remain ineffective and offer insufficient protection to users. Moreover, the roles of *deployers* and *providers* specified within the AI Act and carrying specific responsibilities have raised discussions and concerns. The definition and responsibilities are vaguely defined, which leaves room for interpretation and may lead to differences on a national level when enforcing the AI Act. Overall, the AI Act and its implications are multifaceted, emphasizing the importance of a balanced approach to AI regulation that considers the perspectives of all stakeholders involved. Thus, within this research, the implications of the AI Act for the use case of mis-and disinformation detection are analyzed from different perspectives: (1) research, (2) provider, (3) deployer, and (4) user perspective. This contribution aims to facilitate the understanding of the AI Act's implications for the stakeholders involved.

2. Background

Different regulations and obligations must be considered when using LGAIMs in the EU. The European Council and Parliament have reached a provisional consensus on the proposed AI Act, establishing uniform regulations for artificial intelligence. This includes Article 13, known as "Transparency and Provision of Information to Users," within the EU AI Act[1]. In this context, the requirements for adequate transparency of AI systems are specified, ensuring that both providers and users can reasonably comprehend the functioning and recommendations of the AI system. Therefore, adherence to transparency obligations is mandatory when utilizing AI systems for disinformation detection within the EU. Furthermore, the voluntary Code of Practice on Disinformation[2] has been crafted collaboratively by various stakeholders from industry, legal, and research sectors to establish a unified approach for addressing disinformation online on an international scale. The AI Act, along with the Code of Practice on Disinformation, mandates transparent and detailed system architecture for AI applications tasked with disinformation detection. The considerable data demands for developing LGAIMs typically mean that creators must depend on publicly accessible internet data for training, a source that is rarely ideal in terms of data quality (Luccioni and Viviano, 2021). Consequently, the output produced by these models can be biased, discriminatory, or detrimental (Nadeem et al., 2020). To prevent or at least lessen this problem, model developers should employ appropriate curation methods (Bai et al., 2022). Although the absence of transparency from most LGAIMs makes it impossible to confirm assertions about handling harmful content, it appears that most LGAIMs depended, or still depend, on human intervention to train an automated content moderation system, aiming to inhibit the generation of abusive content (Frey and Osborne, 2023; Helberger and Diakopoulos, 2023a). However, even if the detection of abusive content were automated and flawless, it would only address part of the issue. The persistent risk is the generation of disinformation, which can be challenging to identify (Goldstein et al., 2023). Nevertheless, LGAIMs can not only be utilized to generate harmful and potentially fake content but also to detect disinformation. Several endeavors are made to fight mis-and disinformation by developing advanced AI models for facilitating its detection. Within the media landscape, AI is progressively employed to perform content verification tasks to detect disinforma-

[1] *Laying down Harmonised Rules on Artificial Intelligence* (AI Act), 15.01.2024.

[2] *Strengthened Code of Practice on Disinformation*, 15.01.2024.

tion. Several research and development projects, such as AI4Media, vera.ai, and news-polygraph face similar questions in the interdisciplinary consortium including partners from research, industry and media what implications existing regulations enforce on the outcome of the projects. Hereby, the question arises of how LGAIMs can be used for this specific use case by complying with the new obligations outlined in the AI Act and Digital Services Act (DSA) in using AI systems for combating disinformation. Given the considerable challenges AI-generated disinformation poses, the legal landscape is evolving to address these complex issues. As AI technologies become increasingly capable of generating persuasive and realistic disinformation, the need for a robust legal framework to mitigate the risks and protect public discourse becomes paramount. Similarly, for the use of LGAIMs, the legal regulations become more pronounced and need to be considered when using AI-based tools for dis- and misinformation detection.

2.1. AI Act

Before delving into the legal implications, an introduction to the AI Act and its foundational concepts is provided. The EU is actively pursuing a broad regulatory effort, the AI Act, designed to create a thorough regulatory framework for AI governance. The European Parliament has endorsed new regulations that focus on enhancing transparency and risk management in creating AI systems across the EU, prioritizing a human-centric and ethical approach. The AI Act encompasses AI applications within both the public and private sectors, targeting systems either sold in the EU market or impacting EU citizens. Its central aim is to provide AI developers, deployers, and users with detailed guidance by outlining requirements and obligations for various AI system applications. Hereby, the adoption of a risk-based approach has been driven by thorough consultations with essential stakeholders, notably the High-Level Expert Group on AI. The risk-based approach balances recognizing AI's inherent benefits and potentials against acknowledging possible dangers and risks from novel AI applications and systems. The regulation adopts an inclusive definition of AI in *Article 3*, covering general AI systems influencing decision-making and opinions by providing content, predictions, recommendations, or decisions. This definition covers a variety of methodologies, including machine learning techniques (such as supervised, unsupervised, reinforcement, and deep learning), logic- and knowledge-based approaches (including inductive logic programming, knowledge representation, and deductive engines), as well as statistical methods like Bayesian estimation and search optimization. Within the framework of the AI Act, a risk-based classification outlines four distinct categories of risks concerning AI systems, with particular emphasis on delineating between *high-risk* and *limited risk* categories. (1) **Unacceptable risk**: This category includes AI systems that pose clear threats to the safety, fundamental rights, and well-being of individuals. Examples encompass state-run social scoring mechanisms and unsafe voice-activated toys explicitly banned from the European market. (2) **High risk:** AI systems necessary to sectors important to human health and safety, such as infrastructure, education, safety components, law enforcement, and public administration, are classified here. Compliance with stringent requirements, as specified in *Chapters 2 and 3* of the AI Act (eu, 2021), is mandatory before these systems can be introduced to the EU market. These requirements cover using high-quality data sets, risk management systems, transparency, accuracy, security and robustness measures, user guidance, human oversight, and conformity evaluations. (3) **Limited risk:** This classification applies to AI applications that necessitate transparency to ensure user interactions with AI are intelligible. It primarily mandates that users be adequately informed when they are interacting with AI systems or AI-generated content, including audio and video manipulations (e.g., deepfakes). (4) **Minimal risk:** AI systems that are supposed to pose a minor risk to humans, such as those used in video games, email spam filters, and certain consumer applications, fall under this category. For these, the Act defines no additional specific regulatory obligations. In light of technological advancements, regulatory bodies have incorporated a provision mandating the continuous evaluation of AI systems' risk classifications. The EU is instructed to consider the "intended purpose of the AI system" during the risk classification process of AI technologies (eu, 2021). This provision underscores the critical issue stemming from the potential of AI systems to bypass or dodge the Act's protective measures. This problem is attributed to the complex interplay among the developers and deployers providing AI systems and the distinct purpose(s) these systems are designed to fulfill (Gutierrez et al., 2022).

2.2. DSA

When discussing regulatory frameworks concerning mis-and disinformation detection, it is also important to consider the DSA. Like almost all new technologies, generative models can be employed for positive uses (such as creating birthday cards) or negative ones (such as starting a *shitstorm* on social media platforms) (Brundage et al., 2018). Specifically, the developers of ChatGPT foresaw the possibility of misuse and trained an in-house AI moderator to detect harmful content, albeit with contentious assistance from contractors in Kenya

(Perrigo, 2023). Nonetheless, individuals determined to use ChatGPT and similar LGAIMs, such as Mixtral and Llama 2, to create deceptive or harmful content will discover methods to elicit such responses. Prompt engineering is evolving into a sophisticated technique for extracting any type of content from LGAIMs, and detecting disinformation becomes more and more challenging despite ongoing industry initiatives to enhance the transparency of models and sources (Deiseroth et al., 2023). In response to the rising challenge of fake news and hateful content, the EU has recently implemented the DSA. However, when the DSA was crafted, LGAIMs were not the center of public discourse. Therefore, the DSA aimed to address illegal content on social networks, which was predominantly generated by human users or the occasional automated X (Twitter) account, rather than tackling the challenges posed by LGAIMs. The DSA appears to be outdated as soon as it was implemented due to two significant limitations in its scope. Firstly, it is applicable only to what is termed intermediary services (as per Articles 2(1) and (2) of the DSA). Article 3(g) of the DSA categorizes these as "mere conduits" (like Internet service providers), "caching," or "hosting" services (such as social media platforms, also referred to in Recital 28 of the DSA). However, it is arguable that LGAIMs do not fit into any of these categories. They differ distinctly from mere conduit or caching services that facilitate internet connections. On the other hand, hosting services are described as entities that store information provided by and at the request of a user (Article 3(g)(iii) DSA). In contrast to traditional social media setups, in the context of LGAIMs, it is the AI model, not the user, that generates the content (Hacker et al., 2023). Therefore, the scope of the DSA mechanisms remains applicable only to the sharing of content generated by LGAIMs on conventional social networks. Mis- and disinformation can also be disseminated effectively and broadly through direct personal communication. Despite the EU legislator's decision to leave closed groups outside the DSA's ambit, this decision necessitates reconsideration in light of the accessibility of LGAIM-generated outputs, which amplify the associated risks. Even the strictest enforcement of DSA regulations, possibly in conjunction with the General Data Protection Regulation (GDPR) mandates for data deletion (Articles 17(2) and 19 GDPR), is insufficient to reverse the damage or often prevent the ongoing spread of problematic content. Despite commendable attempts through the DSA to tackle the spread of disinformation and hate speech, the current EU legislation is inadequate in fully addressing the negative implications of LGAIMs. Thus, a selective expansion of the DSA to LGAIMs is necessary to make them useful for disinformation detection.

3. Risk Assessment of Disinformation Detection

As LGAIMs become more advanced and are also applied for disinformation detection, the risk categorization of LGAIMs needs to be clarified. Given the sensitive nature of disinformation, which frequently entails determinations regarding the flagging, removal, or blocking of information, there exists a potential for infringement upon freedom of expression. Consequently, the deployment and subsequent actions derived from AI systems' classification or prediction outcomes can be classified under the *high-risk* or *limited risk* category, depending on the concrete usage scenario. First, within the AI Act, LGAIMs are defined as General-Purpose AI Systems (GPAIS) designed by the provider to execute universally applicable tasks such as image and speech recognition, generating audio and video, detecting patterns, answering questions, translating, among others; a general-purpose AI system is capable of being utilized across multiple contexts and incorporated into various other AI systems (Art. 3(1b) AI Act). The late inclusion of LGAIMs in the AI Act was a key point of the debate for the final version of the AI Act and was mostly motivated by the emergence and wide adoption of ChatGPT (Hacker et al., 2023). Conceptually, the term *generality* might pertain to various aspects such as their capabilities (like language processing versus visual comprehension or their integration in multimodal models), the range of application areas (such as educational or economic domains), the wide array of tasks they can perform (like summarization versus text completion), or the flexibility in the types of outputs they can generate (such as producing images in black and white or in full color) (Gutierrez et al., 2022). General Purpose AI Systems (GPAIS) fall under high-risk obligations (such as Articles 8 to 15 of the AI Act) if they can be employed as high-risk systems or as parts of such systems (as per Article 4b(1)(1) and 4b(2) of the AI Act). Thus, unless it can be technically guaranteed that misuse is prevented, LGAIMs will generally be classified as high-risk systems under the suggested regulation. Second, even if we would not use GPAIS for disinformation detection, one can easily argue that these systems, through content moderation, impact fundamental rights, in particular freedom of expression and information. As defined in Recital 28a of the AI Act, this is a strong argument for classifying them as high-risk. Third, Annex III lists application areas where systems are classified as high-risk per se. This includes, according to Article 8 (aa):

> *AI systems intended to be used for influencing the outcome of an election or referendum or the voting behavior of natural persons in the exercise of their vote in elections or referenda*.

It can, in our view, easily be argued that the detection and potential deletion of disinformation can influence the outcome of elections (in a positive way, we hope). If now AI systems in the domain of disinformation fall under high-risk, this necessitates their compliance with high-risk obligations, specifically data governance, the creation of an extensive risk management system, transparency obligations, and human oversight as specified by Chapter 2 of the AI Act.

For example, Article 10 on data governance demands that:

> (3) *Training, validation, and testing datasets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose*.

This means that only such GPAIS can be used where the respective data has been documented, which is currently not the case for most commercial models. Moreover, when applied to such open domains as misinformation detection, it is by no means clear what the demand "free of errors and complete" could mean and how this can be proven.

Another obligation of high-risk AI systems in Article 9 on risk management demands that:

> (2) *The risk management system shall be understood as a continuous iterative process planned and run throughout the entire lifecycle of a high-risk AI system, requiring regular, systematic review and updating*.

This means that parallel to the disinformation detection process, a monitoring process needs to be installed and maintained to ensure, e.g., the:

> (a) *identification and analysis of the known and reasonably foreseeable risks that the high-risk AI system can pose to the health, safety, or fundamental rights when the high-risk AI system is used in accordance with its intended purpose*.

As discussed above, in the target domain, this could mean constantly checking if the system does not hinder the freedom of expression. As the requirement mentions the entire lifecycle, it might even include the phase of research (that is usually excluded, see below), which would mean that research and development projects would need additional resources and probably also interdisciplinary cooperation to ensure that at any time, there is a well-defined "intended purpose" and a process to come up with "reasonably foreseeable risks".

4. Implications of Regulatory Frameworks

In the following, the implications of mainly the AI Act and the DSA will be analyzed in more detail for the misinformation use case. Within the news-polygraph, several partners from research, industry, and media are concerned with different challenges concerning legal obligations. Thus, the legal implications will be described from three different perspectives, namely the (1) research perspective, (2) the (provider and) deployer perspective, and (3) the user perspective. Notably, the observations below are not to be understood as a legal exegesis but as an attempt to assess the consequences of the AI Act in the domain of disinformation.

4.1. Research Perspective

The impact of the AI Act on research remains very limited. AI systems and models developed and used solely for scientific research and development purposes are explicitly excluded from the scope of the Act. This exemption acknowledges the distinct nature of research activities from commercial or operational AI applications (Haataja and Bryson, 2022). Moreover, the Act clarifies that AI systems used in the context of product-oriented research, testing, and development activities are not subject to its requirements prior to being placed on the market or put into service. This exclusion aims to encourage exploratory research and innovation without imposing premature regulatory burdens. However, researchers still have to consider ethical principles such as human agency, technical robustness, privacy, transparency, diversity, societal well-being, and accountability. However, they are non-binding and serve as a foundational guide for responsible AI development. Researchers are encouraged to consider these ethical principles in their work, aligning research practices with values that promote trustworthiness and human-centric AI (Helberger and Diakopoulos, 2023b). For research and development activities, this approach underscores the importance of assessing and mitigating potential risks associated with AI systems at an early stage. As research does not take place in the void, researchers developing AI systems that may (later) be classified as high-risk are encouraged to incorporate risk assessment and management practices into their development processes. Within research and development projects such as news-polygraph, it is meaningful and rational to consider the risk assessment and compliance with the respective obligations from an early point in time to design potential resulting products in accordance with the AI Act obligations. Additionally, purely research-based LLMs and LGAIMs are exempt from most regulations and can be developed in so-called *regulatory sandboxes*. However, the AI Act aims to foster ethical considerations and prioritize transparency and the mitigation of biases also in *regulatory sandboxes* (Helberger and Diakopoulos, 2023b). Thus, researchers remain merely unaffected by the AI Act as long as the LGAIMs and LLMs are not put into application and commercial use. Moreover, Researchers and stakeholders are encouraged to engage in activities that promote AI literacy, ensuring that AI technologies are accessible and understandable to a broader audience. This initiative aims to build public trust in AI technologies and foster an informed dialogue about AI's role in society.

4.2. Deployer Perspective

The deployer perspective from a commercial viewpoint is distinct from research. Using fine-tuned LGAIMs as a deployer within the EU for use cases, such as disinformation detection, has three main consequences resulting from the fact that these LGAIMs fall under the high-risk category (Hacker et al., 2023). First, deployers will only be able to use LGAIMs from providers who themselves adhere to the obligations for high-risk applications demanded by the AI-Act if they do not develop the LGAIMs themselves. The specific obligations for providers include comprehensive management for quality assurance and system performance and, as a pre-condition, an assessment of conformity and CE-Marking. While the legislation intends the step towards conformity and establishing standards, the distinction between provider and deployer may raise questions in practice, particularly in the case of Open-Source LLMs. Second, deployers must follow a comprehensive list of obligations. This includes the establishment of a risk management system, transparency obligations, need to be in place, indicating that developers and deployers need to build up an in-depth understanding of potentially risky outputs of LGAIMs and their intended use cases. The question of which training data can be lawfully used cannot be answered by the AI Act alone, as additional regulations such as the Data Act and the GDPR need to be considered in deciding the lawful handling of data. This is especially difficult when using LGAIMs from providers where only limited information about the training data is available. The obligations for high-risk AI systems demand representative, complete, and error-free datasets on which the AI systems are trained on. These criteria are very hard to meet, as no concrete metrics or measures are provided in guiding the assessment of datasets. The assumption that AI systems operate accurately, family and without bias when the aforementioned conditions of the dataset are met is misleading, as also model biases can occur not inherent in the training data. Moreover, the requirements of representative and bias-free datasets can be contradicting. When a representative sample is drawn, e.g., about the social media posts of nurses, there might be a clear gender bias towards female nurses. In such cases, it remains very opaque if representative or bias-free datasets are more important. One of the central obligations following form the high-risk categorization is AI systems' transparency and human oversight. Hereby, emerging research in the realm of eXplainable AI (XAI) has demonstrated its capacity to clarify the opaque *black box* aspects of AI algorithms, enhancing the comprehensibility of AI-driven classifications or outputs (Longo et al., 2023; Speith and Langer, 2023). XAI features not only facilitate the understanding of LGAIMs outputs but also the human oversight of such systems for the disinformation detection use case. However, the concrete interpretation of what meaningful explanations allow for transparency and effective human oversight in specific use cases heavily depends on the background of the users (Schmitt et al., 2024). For companies developing high-risk AI systems, it remains challenging to adopt the obligations to concrete technical measures and metrics as only very limited guidance is given. Additionally, providers of GPAIs with a dual-use character, for example, in the domain of media intelligence, are specifically affected by the regulations as their applications could also be classified as high-risk applications. Here, one business application is the detection of company-related dis- or misinformation with the use of LGAIM-based applications. Such applications may, among others, assist in detecting AI-generated content or tracking the diffusion of disinformation. As the detection of company-based disinformation is a subset, disinformation detection providers may attempt to limit the scope of their applications towards company-related disinformation detection in order to circumvent the obligations for high-risk applications. Nevertheless, companies failing to comply with the obligations outlined for the respective risk category may result in high fines, which can reach up to 35 million € or 7% of the company's worldwide annual turnover. Thus, companies, also within research and development projects, need to undergo the risk assessment of applications developed in such frameworks from an early stage to be aligned with the obligations outlined in the AI Act when products are put on the market.

4.3. User Perspective

From a media organization's perspective, the spread of mis- and disinformation is a significant challenge, and it is expected to become even more so in the coming years, particularly when dealing with synthetic and altered media content. However, LGAIMs are not only potential sources of spreading mis- and disinformation but can also assist journalists in uncovering such content. As reported in a white paper by the EU-funded project AI4media, AI technologies are regarded as highly valuable by most fact-checking and verification specialists (AI4, 2022). The debate over whether LGAIMs in the media should be classified as high-risk and subjected to the strictest regulatory measures is closely linked with ongoing discussions about the influence of algorithm-driven platforms and, more broadly, the effects of AI utilization in media on fundamental rights like freedom of speech and privacy rights. As fundamental rights might be affected when using LGAIMs to detect disinformation and harmful content, LGAIMs can be categorized as high-risk AI systems and need to follow the respective obligations (Helberger and Diakopoulos, 2023b). Therefore, among others, effective human oversight, transparency obligations, and a risk management system need to be ensured when applying LGAIMs for mis- and disinformation detection. When using LGAIMs for disinformation detection, the next regulatory framework relevant to their application from a user's perspective is the Digital Services Act (DSA). The user perspective is relevant to consider to gain a more in-depth understanding of the implications of the AI Act on advanced transparency and human oversight of AI systems used for mis-and disinformation detection. While some use cases for applying LGAIMs in the journalistic verification process may be obvious, others may appear less relevant at first glance. Overall, LGAIMs and AI systems can be used differently in the journalistic context, which is also partially covered by tools developed within the news-polygraph research and development project.

LGAIMs-driven tools are widely accepted in the field of Human Language Technologies. These tools, such as plainX[3] allow for the transcription and translation of video and audio content. While these technologies are not primarily designed to detect mis- and disinformation, they are undoubtedly useful for journalists to learn what content in a foreign language is about and whether it is the same as it claims to be about. Videos that intentionally mistranslate the original foreign language speech through incorrect voiceovers or subtitles are often used for entertainment purposes. There are tools available, such as the Caption Generator, that enable users to create content with fictional subtitles for popular videos, such as 'Dimitri Reacts'[4]. However, there are also many examples of critical videos with fake subtitles. For example, Full Fact reported on several videos addressing the ongoing conflict in Israel and the Gaza Strip. One social media video suggests that a Palestinian woman said in Arabic, 'We are prisoners of Hamas,' which is a deliberately incorrect translation[5]. Other videos wrongfully claim to show North Korean leader Kim Jong Un making a speech about the Israel-Gaza conflict or Putin and Erdogan warning America over its support for Israel[6]. These examples show that even AI tools used for translation can result in harmful outputs when AI system predictions are wrong. Wrong translations can result in misinterpretation of the meaning. This can lead to the blocking of such information or printing it as truthful content when no expert-level language knowledge is available to prove the AI-generated translations. However, such AI systems apply rather to the category of *limited risks* and need to comply with minor transparency obligations. Additionally, different AI tools are already used to support journalists in their fact-checking tasks. The InVid WeVerify[7] Chrome plugin provides an advanced forensic toolbox for image verification suspected of being manipulated. Moreover, approaches such as Retrieval-Augmented Generation (RAG) can be used to integrate external knowledge sources for knowledge enrichment for certain fact-checking tasks. For example, the Database of Known Fakes (DBFK)[8] provides a useful integration of external knowledge in multiple languages relevant for checking context information about a specific claim or entity. When using such tools, transparency is highly important to journalists as they need to understand the reasoning behind a specific AI model output for content verification. Thus, independent of the risk category, journalists require sufficient and meaningful transparency to rely on the AI model output. As most of the AI tools applied in the fact-checking and content-verification process apply to the high-risk category, they must integrate transparency measures and effective human oversight. Previous research has

[3] https://www.plainx.com/, last accessed 03.04.2024.

[4] https://www.captiongenerator.com/make-a-dimitri-finds-out-video, last accessed 03.04.2024.

[5] https://fullfact.org/online/fake-subtitles-video-palestinian-woman/, last accessed 03.04.2024.

[6] https://fullfact.org/online/fake-kim-jong-un-north-korea-israel-gaza/, last accessed 03.04.2024.

[7] https://weverify.eu/

[8] https://shorturl.at/kBDHL, last accessed 03.04.2024.

shown that natural language explanations can be easily perceived by humans but also create false trust in the AI system when the predictions or classifications are wrong (Schmitt et al., 2024). Therefore, the high-risk obligations model's faithfulness and robustness are highly important for sensitive tasks such as content verification. Moreover, meaningful explanations heavily depend on the users' prior knowledge and background. Therefore, explanations need to be incorporated to provide explanations on different levels of abstractions that users with varying degrees of expert knowledge can comprehend. From a user's perspective, the obligations defined in the AI Act are very beneficial if implemented adequately. The transparency measures, explanations given, and modes of collaboration for ensuring human oversight need to be designed carefully to allow for the effective integration of human knowledge and human oversight, especially in domains where human rights might be affected.

When combining the three perspectives for the news-polygraph research and development project, the research institutes involved in the project need to consider the obligations defined for high-risk AI systems to prepare the tools for the deployer partners adequately. The deployer partners have to establish adequate procedures for risk assessment (also continuously), data governance structure, technical documentation, accuracy, robustness, and security measures. In collaboration with the user partners, the research and industry partners must develop sufficient means of transparency and meaningful explanations to allow for meaningful collaboration between journalists and AI systems for an overall improved performance on the content verification task.

5. Critique

When analyzing the different obligations within the AI Act, such as transparency, complete and representative datasets for model training, accountability, and fairness, the concrete implications of specific use cases remain opaque, as does a clear definition of process steps such as "research" versus "entire lifecycle of a high-risk AI system", which are subject to very different regulatory measures. Some guidance is given on the risk assessment of AI systems conducted by providers and deployers by themselves, but there is still room for interpreting the risk categorization depending on the provider's/deployers' needs. Due to the comprehensive list of obligations defined in the high-risk category, it can be assumed that deployers will avoid categorizing their AI systems as high-risk AI systems. Deployers need to be fully aware of the consequences the choice of LGAIMs as a technology for production, for example, in media intelligence applications, may entail even if the contribution of the LGAIM to the overall functionality is limited, e.g., if the LGAIM is only used for a final language checking of other system's output in a hybrid setting. They need to carefully select providers of LGAIM based on an assessment of eligible certification and existing quality assurance practices, as the failure to do so could result in massive fines of up to 7% of the annual turnover. As the AI Act requires a constant exchange between deployers and developers of LGAIMs, deployers should assign clear responsibilities for these tasks to their respective managers. While LGAIM-based applications provide various opportunities for improved services to uncover company-based disinformation, the deployment will come at the cost of adhering to the regulations imposed by the AI Act. Deployers may find themselves in a situation where they want to contribute as part of their Corporate Social Responsibility campaign an ad-hoc report about the spread of disinformation in light of an upcoming election and may choose to produce this report without the use of LGAIMs in order to circumvent the regulations imposed by the AI Act or disregard such reports at all. In light of the early stage of implication, deployers will need to follow the developments around the implementation of the AI Act and the legal interpretation made for weakly specified terms in the AI Act across Europe closely, for example, in court rulings or administrative regulations. Moreover, as described in Section 4.3, LGAIMs can be valuable in identifying dis- and misinformation. Journalists require such tools, and several are already available or in development. However, it is crucial to explain these tools' functionalities, outcomes, and constraints. Journalists often work under time constraints while also striving for high credibility. As a result, journalists need to have a certain level of technical skills and AI literacy to be able to recognize the strengths and limitations of the tools they are using. Additionally, journalists must be able to determine whether the use of LGAIMs-based tools complies with the DSA, particularly when processing sensitive data. This may include leaked data or information containing personal data. For example, if data requires verification, LGAIMs that use the inserted information for training should not be applied.

The implementation of GDPR has revealed that without clear technical guidelines and the absence of monitoring mechanisms at both national and EU levels, the regulation may not achieve its intended effectiveness. Previous research (Schmitt et al., 2023) indicates that while GDPR has enhanced certain practices in personal data management, it falls short of establishing precise technical criteria for detecting non-compliance. Despite platforms, applications, and services declaring GDPR adherence through privacy policies and consent forms, these claims often lack verifiable technical substantiation. Similarly, without verification processes to assess compliance with the AI Act from a technical standpoint, this regulation risks being as ineffectual as GDPR, yielding only marginal improvements in ethical AI system practices.

Overall, the regulations must be interpreted and understood depending on specific use cases in which LGAIMs are calibrated. Therefore, we recommend 1) setting minimum standards for LGAIMs and not classifying all LGAIMs as high-risk AI systems, 2) defining high-risk rules specific for LGAIMs employed and used in high-risk scenarios, and 3) establishing standards of adequate transparency, human oversight, and risk management to comply with the rules outlined in the AI Act.

6. Conclusion

In conclusion, the deployment and utilization of GLAIMs for disinformation detection within the complex landscape of the forthcoming AI Act and DSA offer both significant opportunities and difficult challenges. The paper has examined the multifaceted implications of the AI Act, highlighting the nuanced obligations these frameworks impose on research, deployer, and user perspectives in the context of mis- and disinformation detection. Central to the discourse is recognizing LGAIMs as potentially high-risk systems when applied to disinformation detection, necessitating rigorous compliance with a longer list of obligations such as risk management, (training data) transparency, and human oversight. This designation underscores the critical need for deployers and developers to ensure that LGAIMs are not only effective in detecting and mitigating disinformation but also aligned with ethical standards and legal requirements aimed at safeguarding public discourse and protecting fundamental rights. Moreover, we highlight the challenges and ambiguities in interpreting the AI Act's provisions, offering clear standards and guidelines that facilitate the responsible use of LGAIMs in combating disinformation. Time will tell to what extent the issues we consider will remain in the implementation of the AI Act. In summary, this paper provides a targeted analysis of the legal and ethical landscape surrounding the use of LGAIMs for disinformation detection, offering insights into the complexities of navigating regulatory frameworks. It underscores the imperative for a collaborative effort among stakeholders to ensure that the deployment of LGAIMs is both effective in countering disinformation and compliant with evolving legal standards, thereby contributing to the integrity and resilience of information ecosystems in the digital age.

7. Acknowledgements

The research was conducted in the scope of the BMBF project news-polygraph (03RU2U151x).

8. Bibliographical References

2021. Proposal regulation: laying down harmonised rules artificial intelligence.

2022. Use case 1: Deepfake detection - white paper. Technical report, AI4Media. Accessed: 2024-04-04.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Vimala Balakrishnan, Ng Wei Zhen, Soo Mun Chong, Gan Joo Han, and Tan Jiat Lee. 2022. Infodemic and fake news–a comprehensive overview of its global magnitude during the covid-19 pandemic in 2021: A scoping review. *International Journal of Disaster Risk Reduction*, page 103144.

Tim Brooks, Bill Peebles, Connor Homes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Wing Yin Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators.

Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*.

Omar Darwish, Yahya Tashtoush, Majdi Maabreh, Rana Al-essa, Ruba Aln'uman, Ammar Alqublan, Munther Abualkibash, and Mahmoud Elkhodr. 2023. Identifying fake news in the russian-ukrainian conflict using machine learning. In *Advanced Information Networking and Applications: Proceedings of the 37th International Conference on Advanced Information Networking and Applications (AINA-2023), Volume 3*, pages 546–557. Springer.

Björn Deiseroth, Mayukh Deb, Samuel Weinbach, Manuel Brack, Patrick Schramowski, and Kristian Kersting. 2023. Atman: Understanding transformer predictions through memory efficient attention manipulation. *arXiv preprint arXiv:2301.08110*.

World Economic Forum. 2024. Global risks 2024: At a turning point.

Carl Benedikt Frey and Michael Osborne. 2023. Generative ai and the future of work: A reappraisal. *Brown Journal of World Affairs*, pages 1–12.

Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative language models and automated influence operations: Emerging threats and potential mitigations. *arXiv preprint arXiv:2301.04246*.

Carlos Ignacio Gutierrez, Anthony Aguirre, Risto Uuk, Claire C Boine, and Matija Franklin. 2022. A proposal for a definition of general purpose artificial intelligence systems. *Available at SSRN 4238951*.

Meeri Haataja and Joanna J Bryson. 2022. Reflections on the eu's ai act and how we could make it even better. *TechREG™ Chronicle*, (March 2022).

Philipp Hacker, Andreas Engel, and Marco Mauer. 2023. Regulating chatgpt and other large generative ai models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1112–1123.

Natali Helberger and Nicholas Diakopoulos. 2023a. Chatgpt and the ai act. *Internet Policy Review*, 12(1).

Natali Helberger and Nicholas Diakopoulos. 2023b. The european ai act and how it matters for research into ai in media and journalism. *Digital Journalism*, 11(9):1751–1760.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

Dudi Iskandar, Indah Surywati, and Geri Suratino. 2023. Public communication model in combating hoaxes and fake news in ahead of the 2024 general election. *International Journal of Environmental, Sustainability, and Social Science*, 4(5):1505–1518.

Razieh Khamsehashari, Vera Schmitt, Tim Polzehl, Salar Mohtaj, and Sebastian Moeller. 2023. How risky is multimodal fake news detection? a review of cross-modal learning approaches under eu ai act constrains. In *Proc. 2023 ISCA Symposium on Security and Privacy in Speech Communication*, pages 47–51.

Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, et al. 2023. Explainable artificial intelligence (xai) 2.0: A manifesto of open challenges and interdisciplinary research directions. *arXiv preprint arXiv:2310.19775*.

Chiara Longoni, Andrey Fradkin, Luca Cian, and Gordon Pennycook. 2022. News from generative artificial intelligence is believed less. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 97–106.

Alexandra Sasha Luccioni and Joseph D Viviano. 2021. What's in the box? a preliminary analysis of undesirable content in the common crawl corpus. *arXiv preprint arXiv:2105.02732*.

Linda Monsees. 2023. Information disorder, fake news and the future of democracy. *Globalizations*, 20(1):153–168.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

OpenAI. 2023. Gpt-4 technical report.

B Perrigo. 2023. The 2 dollar per hour workers who made chatgpt safer. `https://time.com/6247678/openai-chatgpt-kenya-workers/`.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Adam Satariano and Paul Mozur. 2023. The people onscreen are fake. the disinformation is real. *International New York Times*, pages NA–NA.

Vera Schmitt, James Nicholson, and Sebastian Möller. 2023. Is your surveillance camera app watching you? a privacy analysis. In *Science and Information Conference*, pages 1375–1393. Springer.

Vera Schmitt, Luis-Felipe Villa-Arenas, Nils Feldhus, Joachim Meyer, Robert P. Spang, and Sebastian Moeller. 2024. The role of explainability in collaborative human-ai disinformation detection. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

Felix M Simon, Sacha Altay, and Hugo Mercier. 2023. Misinformation reloaded? fears about the impact of generative ai on misinformation are overblown. *Harvard Kennedy School Misinformation Review*, 4(5).

Timo Speith and Markus Langer. 2023. A new perspective on evaluation methods for explainable artificial intelligence (xai). In *2023 IEEE 31st International Requirements Engineering Conference Workshops (REW)*, pages 325–331. IEEE.

Selling Personal Information: Data Brokers and the Limits of US Regulation

Denise DiPersio
Linguistic Data Consortium, University of Pennsylvania
3600 Market Street, Suite 810, Philadelphia, PA 19104 USA
dipersio@ldc.upenn.edu

Abstract

A principal pillar of the US Blueprint for an AI Bill of Rights is data privacy, specifically, that individuals should be protected from abusive practices by data collectors and data aggregators, and that users should have control over how their personal information is collected and used. An area that spotlights the need for such protections is found in the common practices of data brokers who scrape, purchase, process and reassemble personal information in bulk and sell it for a variety of downstream uses. Such activities almost always occur in the absence of users' knowledge or meaningful consent, yet they are legal under US law. This paper examines how data brokers operate, provides some examples of recent US regulatory actions taken against them, summarizes federal efforts to redress data broker practices and concludes that as long as there continues to be no comprehensive federal data protection and privacy scheme, efforts to control such behavior will have only a limited effect. This paper also addresses the limits of informed consent on the use of personal information in language resources and suggests a solution in an holistic approach to data protection and privacy across the data/development life cycle.

Keywords: personal information, privacy, data broker

1. Introduction

A principal pillar of the US Blueprint for an AI Bill of Rights is data privacy, specifically, that individuals should be protected from abusive practices by data collectors and data aggregators, and that users should have control over how their personal information is collected and used. An area that spotlights the need for such protections is found in the common practices of data brokers who scrape, purchase, process and reassemble personal information in bulk and sell it for a variety of downstream uses. Such activities almost always occur in the absence of users' knowledge or meaningful consent, yet they are legal under US law. This paper examines how data brokers operate, provides some examples of recent US regulatory actions taken against them, summarizes federal efforts to redress data broker practices and concludes that as long as there continues to be no comprehensive federal data protection and privacy scheme, efforts to control such behavior will have only a limited effect. This paper also addresses the limits of informed consent around the use of personal information in language resources and suggests a solution in an holistic approach to data protection and privacy across the data/development life cycle.

2. What Is Personal Information?

Acknowledging that there is no legal framework governing the use of 'personal information', the Blueprint for an AI Bill of Rights does not attempt to define the term. It instead focuses on the ways industry and government use individuals' data, particularly in 'senstitve' domains that include health, employment, education, criminal justice and personal finance. Similarly, as shown below, a significant part of the discussion about data brokers refers to their use of 'sensitive' geolocation data.

In US human subjects data collections, researchers refer to 'personally identifiable information' (PII) as something that must be protected. But the Common Rule – the federal regulation governing human subjects research – does not define that term.[1] Some US government agencies have developed their own definitions of PII. The US General Services Administration (GSA), the body responsible for managing federal property and providing contracting options for government agencies, defines PII broadly as 'information that can be used to distnguish or trace an individual's identity, either alone or when combined with other information that is linked or linkable to a specific individual.' [2] This definition also recognizes that PII must be assessed on a case-by-case basis since it is not characterized by a single category or technology. [3]

By contrast, the global standard as expressed in Article 4(1) of the General Data Protection Regulation (GDPR) schemes in the European Union and the United Kingdom expands the notion of 'personal data' to factors including a person's economic, cultural, social and physical identity.

[1] The Common Rule speaks in terms of *private information* and *identifiable private information*, both of which refer to non-public data or behavior. 19 CFR 46.102(e), (4), (5).

[2] GSA Rules of Behavior for Handling Personally Identifiable Information (PII), https://www.gsa.gov/directives-library/gsa-rules-of-behavior-for-handling-personally-identifiable-information-pii-2#.

[3] The author is familiar with a case where an earlier version of the GSA definition was adopted by another agency and applied to a language resource data collection conducted by the Linguistic Data Consortium.

LEGAL2024 @LREC-COLING-2024, pages 39–46
20 May, 2024.

This paper means to refer to personal information in the broadest sense, ranging from information provided by individuals in the course of their normal interactions with web platforms and applications (including social media and mobile phone use), to linguistic data collected under research protocols and accessible in published language resources. This includes personal data and sensitive personal data as described in the Blueprint for an AI Bill of Rights, private information as referenced in the Common Rule, PII such as defined by the GSA and personal data within the meaning of the GDPR. Effort will be made to distringuish among these descriptions below.

3. The Data Broker Ecosystem

Anyone participating in the digital world leaves a personal information footprint: from their browser history, website visits and related activities like credit card transactions, to their messaging content and behavior, and beyond. As part of those transactions and interactions, the companies and platforms accessed by users repurpose the information left behind to further monetize it, usually without user knowledge or consent. This can occur in several ways, including through third-party apps containing the data broker's software development kit (SDK), in the broker's own mobile apps, and from information the broker purchased from other brokers and data aggregators. Results range from targeted marketing recommendations to providing the information in bulk to third party data brokers who distribute, combine, process and resell it downstream in various forms. The most pernicious of these are data sets that contain geolocation data which either in its original form, or when manipulated and combined with other data, reveals personal information about a host of habits, including entertainment choices, travel history, and visits to sensitive locations (hospitals, reproductive clinics, places of worship), the latter of which can result in threatening behavior toward identified individuals.

3.1 The Current Regulatory Landscape

In the United States, it is legal to buy data through data brokers. And because there is no general US data protection and privacy law, there is no federal mechanism to scrutinize the privacy implications in data broker transactions. Nevertheless, one can identify some key areas where those transactions violate general privacy principles.

The US regulations governing human subjects research provide that informed consent must be obtained prior to using personally identifiable information. However, interacting with web platforms and related applications is not typically considered to constitute human subjects research. Thus, most platforms and apps are not required to, and do not, provide for meaningful consent in the first instance, nor do they disclose that they have the right to sell user data to unidentified third parties for unknown purposes. Even if they did, those terms are typically buried in a long document to which the user clicks consent. Research shows that the majority of users will not read these documents.[4] Similarly, data broker claims that users have "opted-in" to the ecosystem because they share their information on an app fails as well because as indicated above, users have not been meaningfully informed about the downstream uses of their data. Finally, even if a single data set does not disclose individual information, that information can often be easily reidentified when it is combined with other data (Gebhart & Richman, 2023). Research has shown that reidentification is possible from only a few data points (Sweeney 2000; de Montjoye, et al., 2015).

Buying and selling personal information under the EU GDPR and the UK GDPR is covered under the rules for processing personal data. This means that there must be a legal basis to process personal data, that the data can be used only for the purpose for which it was collected, that the purpose is disclosed, that consent is obtained, and that consent can be withdrawn at any time.[5] These rules apply to data brokers even though they may not have collected the personal data originally if the use by the data broker is different from the use for which consent was obtained.

To the extent that US data platforms and data brokers collect, purchase or sell information from EU citizens that would otherwise be subject to GDPR requirements, it seems clear that their practices do not meet GDPR personal data processing standards.

3.2 Problematic Practices and Efforts to Redress Them

Data brokers generally promote themselves as agents of information that operate for good. They boast that their resources can boost business marketing campaign effectiveness, assist academic researchers searching for equitable solutions to a multitude of problems and help the government manage public crises. And they assert that they accomplish those goals while properly protecting individual privacy rights.

For example, **Veraset** claims billons of data points and location date from over 150 countries 'trusted' by more than 100 data scientists.[6] **Cubebiq** touts its

[4] Cameron Dell, The Sad Truth of the FTC's 'Historic' Privacy Win, https://www.wired.com/story/ftc-xmode-outlogic-location-data-settlement/ (citing research that a person needs around 76 working days to review the privacy policies they interact with in one year).

[5] Ivan Lyaskivskij, GDPR requirements to selling of personal data', https://legalitgroup.com/en/gdpr-requirements-to-selling-of-personal-data-ccpa-vs-gdpr-on-insurance-and-trade/; see also Information Commissioner's Office, What common issues might come up in practice?, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/the-right-to-be-informed/what-common-issues-might-come-up-in-practice/ (construing UK GDPR).

[6] Veraset, https://www.veraset.com/about/data-industry.

work with Oxford University and the US Center for Disease Control to provide population 'mobility insights' during the COVID-19 pandemic. [7] **Kochava** promises that anything can be measured: 'Any Channel, Any Device, Any Audience'. [8]

However, as indicated above, the principal way data brokers attempt to defuse privacy-related criticisms is to invoke the notion of a user 'opt-in'. This practice has been the focus of recent disclosures and US regulatory actions against data brokers.

SafeGraph (2022). It was discovered that this firm purchased data from the Life360 app – designed to connect family location information – that included location data for US Planned Parenthood clinics -- which it in turn offered for sale. The company removed its family planning center data in response to protests. (Gebhart & Richman, 2023). This disclosure also resulted in a 2023 class action lawsuit against Life360 claiming that users' location data was sold without permission.[9]

Kochava (2024). The US Federal Trade Commission (FTC), an agency that regulates unfair trade practices, filed a lawsuit against this firm in 2022 for selling geolocation data from mobile devices tracing individual movements to and from sensitive locations. The court dismissed the complaint, but allowed the FTC to amend it with specific examples of consumer harm. In February 2024, the court denied a motion to dismiss the amended complaint which means that the case will proceed. The FTC seeks an injunction to stop Kochava from selling sensitive data without user consent. [10]

X-Mode/Outlogic (2024). The FTC settled its complaint against this firm in January 2024 by entering into a consent order under which, among other things, Outlogic will be prohibited from sharing or selling any sensitive location data; it must also destroy all non-deidentified, sensitive location data previously collected. The company must establish clear and simple user procedures for withdrawing consent, for obtaining the identity of organizations who bought their data, and for removing their data from the company database and recipients' databases. Finally, no recipients of Outlogic's data sets must be able to associate the data with locations relating to LGBTQ+ services, locations of political or social demonstrations or protests, or the location of a specific individual. [11]

3.3 Connections to Research, Law Enforcement and Government

In addition to their commercial customers, data brokers sell their data sets and related resources (tools, APIs) to academic institutions, law enforcement organizations and government agencies. These transactions support the corporate message that data broker services benefit society. But is that the case?

Broker data sets are typically described in journal publications about academic research as data that is 'anonymized' or 'privacy-compliant', which is not always true (Gebhart & Richman, 2023). Those descriptions are perpetuated in open research data sharing mechanisms where safe use is assumed. It also raises the question, posed by Gebhard & Richman, whether this makes researchers 'accomplices' to the practices of data brokers (Ibid., 2023).

In 2023, the US Office of the Director of National Intelligence issued a report showing that agencies including the Federal Bureau of Investigation, the Internal Revenue Service, and the Department of Homeland Security, among others, purchased databases from data brokers, thus avoiding the need to obtain a warrant, a court order, or a subpoena (Ayoub & Goitein, 2024). The US Constitution's Fourth Amendment requires the government to obtain a warrant to access material in which individuals have a reasonable expectation of privacy. Some agencies claim that the Fourth Amendment does not apply to data sold to the government. (Ibid.). This practice seems likely to continue for the time being. Pending legislation would ban government purchases of communications data only. In addition, the Blueprint for an AI Bill of Rights exempts from its coverage government agencies engaged in national security and law enforcement activities.

Data brokers also sell personal information to customers outside the United States. No law regulates or prevents those transactions, notwithstanding the risk that such data can be used against US interests.

3.4 What Americans Think About How Their Personal Information Is Collected and Shared

Americans are becoming increasingly concerned about the privacy of their personal information. In a 2019 study by the Pew Research Center, a

[7] Cubeiq's Data for Good Program: Where We've Been and Where We're Going, https://www.cuebiq.com/resource-center/resources/cuebiqs-data-for-good-program-where-weve-been/.

[8] Kochava, https://www.kochava.com/.

[9] Jon Keegan, Life360 Sued for Selling Location Data, The Markup, https://themarkup.org/privacy/2023/06/01/life360-sued-for-selling-location-data.

[10] Ashley Belanger, Data broker allegedly selling de-anonymized info to face FTC lawsuit after all, https://arstechnica.com/tech-policy/2024/02/data-broker-selling-de-anonymized-info-to-face-ftc-lawsuit-after-all/.

[11] Federal Trade Commission, FTC Order Prohibits Data Broker X-Mode Social and Outlogic from Selling Sensitive Location Data, https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-order-prohibits-data-broker-x-mode-social-outlogic-selling-sensitive-location-data.

nonpartisan organization that conducts opinion polling and other research, over 80% of respondents indicated that they did not have control over data collected about them by companies or the US government and that the risks associated with company-collected data outweighed the benefits. (Auxier & Rainie, 2019). Most did not understand how companies (59%) or the government (78%) used data collected from them.[12] Respondents generally preferred more government regulation but were resigned to the idea that their online activity is being tracked and their personal data collected. (Ibid.).

A 2023 Pew study focused on data privacy revealed that Americans had grown more pessimistic about how their personal information is used. Over 70% had growing concerns over how the government uses the personal data it collects, and they do not trust companies to use their data responsibly. (Faverio, 2023). Even when they make the right decisions to protect their personal information, most believed that their actions do not make a difference in the way companies or social media executives protect their privacy. (Ibid.). As in 2019, the majority of respondents support more government regulation of how personal data is used. (Ibid.; Auxier & Rainie, 2019).

4. Recent Regulatory Developments

4.1 Executive Branch Actions

In the gap left by the lack of a comprehensive law addressing threats to data protection and personal privacy in the digital space, as well as international pressure, the US Executive Branch has taken steps to set down principles and rules designed to address the threat to individuals from the growing scope of AI-powered technologies and systems.

In 2022, the White House Office of Science and Technology Policy issued a Blueprint for an AI Bill of Rights based on five principles: safe and effective systems, algorithmic discrimination procedures, data privacy, notice and explanation, and human alternatives, consideration, and fallback. The data privacy pillar acknowledges the abuses in the way personal data is collected and used by stipulating that data should be collected and used for a particular, stated purpose and context, that consent to collect and use that data should be obtained and the conditions of consent respected, and that any data used in 'sensitive domains' should be subject to further review and potential restraint. Despite the violations of legal process committed by various government agencies using personal data from brokers, the AI Bill of Rights exempts from its coverage government agencies involved in law enforcement and national security.

Building on the AI Bill of Rights, President Biden issued an Executive Order in 2023 on safe, secure, and trustworthy artificial intelligence. The order urges Congress to pass data privacy legislation and identifies actions to enhance privacy protection, such as developing technologies for that purpose, reviewing data collection practices, and establishing federal guidelines. The Order requires technology companies to share with the government the safety testing results of their AI models, a move that has been criticized as stifling innovation and raising the specter of government misuse of such information. Critics also claim that the order usurps legislative authority in the way it outlines a broad, multi-agency effort without prior enabling legislation.

The Executive Branch took a step toward addressing data broker transactions in 2024 in an Executive Order that curtails data brokers' ability to sell sensitive information to non-US customers in, or vendors selling data in, 'countries of concern' (China, Russia, North Korea, Iran, Cuba and Venezuela).[13] This will be accomplished by regulations developed by the US Department of Justice. The order was meant to address in part the disclosures about US government data purchases although it does not prevent the government from purchasing or using such data, nor does it stop data broker sales to non-covered countries.

4.2 Relevant Pending Legislation

Among the many pending legislative bills relating to data protection, privacy, and artificial intelligence, among other things, there are two initiatives with some relationship to data broker activity. The first attempts to prohibit the US government from purchasing communications-related data from brokers. The second is designed to protect consumer privacy by broadly defining personal information. To date Congress has taken no significant action on either.

The **Fourth Amendment is Not for Sale Act (H.R. 4639)** was originally introduced in 2021 and reintroduced in 2023. A response in part to the US Supreme Court's 2018 decision in *Carpenter v. United States*[14] which held that a warrant is needed to obtain an individual's cell phone data, it bars the US government from purchasing communications information, including location data, from third parties that collect or process that information as well as any

[12] 77% of the study respondents had heard 'at least a little bit' about ad targeting (Auxier & Rainie, 2019). The study did not specify data brokers as among "company" data collectors. Based on the information presented in this paper, it can be argued that most respondents would have been unaware or only vaguely aware of the existence of data brokers and the potential downstream uses of their data beyond ad targeting.

[13] The White House, Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern, https://www.whitehouse.gov/briefing-room/presidential-actions/2024/02/28/executive-order-on-preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related-data-by-countries-of-concern/ .

[14] *Carpenter v. United States*, 585 U.S. ___, 138 S.Ct. 2206 (2018).

such information collected by deceptive means through unauthorized access to a device or online account.

The **American Data Privacy and Protection Act (H.R. 8152)** was introduced in 2022 and is meant to provide comprehensive protection for consumer privacy. Personal information is broadly defined to include anything that identifies, is linkable or 'reasonably' linkable' to an individual. Additional protections are extended to sensitive information. Entities covered under the bill do not include the US government, however.

5. Language Resources, Personal Information and Privacy

Personal information is a key component of language resources that support machine learning and natural language processing tasks. Handling personal information during the data collection, processing and data sharing phases is subject to various laws, regulations and ethical best practices. The language resource and evaluation community has largely respected the need to obtain informed consent and to protect personally identifiable information in human subjects collections. This is in contrast to the mostly unrestrained behavior exhibited by data brokers. US and European regulations and their limits are briefly reviewed below, followed by a discussion about ways in which the field is adopting a more holistic approach to data protection and privacy that shows promise.

5.1 Limits of Informed Consent

Informed consent is the linchpin for collecting data from humans for research in the United States. This means that a person must be given sufficient information about the study and about how their data or information will be collected, used and shared. If they agree to participate, they signify their consent, typically in writing or electronically. Similarly, the EU and UK GDPR schemes require consent that describes, among other things, the specific purpose and lawful basis for the collection.

The community typically preserves individual privacy under human subjects research regulations by assigning random identifiers to participants which are stored with the research data; participants' personal information (e.g., their name), is stored separately. The data may also be anonymized or otherwise de-identified after collection and before the material is broadly shared. This is the usual standard for published language resources containing data obtained from human subjects.[15]

However, human subjects regulations do not adequately address the normal interactions between humans and the digital world since those are not typically considered to constitute human subjects research (at least under US law). Even when consent is provided for in click-through terms and conditions, users typically cannot easily find it, nor are the terms clearly explained. In other words, 'consenting' under these circumstances does not rise to the level of informed consent.

Activities such as uploading content to public websites (text, audio, image, video) can implicate personal information in a number of ways, either directly, by containing traditional identifiers like name and other contact information, or indirectly, by containing biometric information, for example. Such multimodal data is highly desired for machine learning and natural language processing applications. Most sites containing such information require that users obtain the consent of the individual uploaders to copy, process and/or share such material. But this is a requirement that is honored more in breach than observance.[16]

Sharing language resources that have not been subject to a legal, ethics and/or privacy review and that are not properly documented in that respect can lead to the continued reuse of problematic data and/or the models and systems developed from it. This is not a trivial concern given the vast number of options for data sharing, many of which provide little or no oversight with respect to the resources posted there.

5.2 An Evolving Holistic Approach to Data Protection and Privacy

As society has become increasingly aware of the ways in which individual personal information can be used without their knowledge, the idea of a broad notion of privacy, separate from copyright protection, is emerging. It encompasses all of the types of data collected about individuals in the digital space and the potential ways in which that data can be used, processed and shared, including problematic downstream effects on algorithm development and system performance. The Blueprint for an AI Bill of Rights endorses a comprehensive approach to data protection and privacy along these lines. This approach is also consistent with provisions in the EU and UK GDPR as well as in various national data protection and privacy laws.

Gaining traction in the community is the thought that data protection, ethics and privacy should be considered and re-considered at all stages of the data/development life cycle: from the research plan, through data collection, milestones, testing, and deployment. This is not a new idea. The concept of

[15] Biometric data, such as a person's voice or image, can also be considered an identifier, or the informed consent may limit the way in which that information can be shared. Thus, published resources may include masked speech, blurred faces, or data to which other methods have been applied to protect privacy. The details about such methods and their efficacy are beyond the scope of this paper.

[16] This attitude is bolstered in large part by the prevailing view of US courts that using certain web data for a machine learning use case constitutes a fair use under the exception to US copyright law. (DiPersio, 2018).

‘Privacy by Design’ originated in the 1970s and has garnered renewed attention since the late 1990s in the US and EU (particularly post-GDPR). (Kamocki & Witt, 2020). Attempts have been made to articulate what a privacy-designed project looks like. One example is an ‘ethos life cycle’ showing six stages of a data science workflow – problem identification, data discovery, exploratory data analysis, modeling, interpretation and conclusions, and communication. (Boenig-Liptsin, et al., 2022). In another example, potential sources of bias are identified across the cycle; they include historical, representation, measurement, aggregation, learning, evaluation, and deployment biases. (Suresh & Guttag, 2021). A third example focused on personal data describes a software tool that allows individuals to control their data during a study and choose the data they contribute to researchers. (Clos, et al., 2022). In all of these instances, researchers continue to be involved in thinking about data protection and privacy through the entire data collection, development, data sharing and deployment process.

6. Future Outlook

As long as there continues to be no comprehensive US data protection and privacy regulatory scheme, the pattern of piecemeal enforcement seen to date will persist. The FTC, an agency with limited powers to regulate unfair trade practices, has carried the principal burden of protecting individual privacy. This is seen most recently in the actions against data brokers Kochava and X-Mode/Outlogic. Many view the consent order against the latter a significant achievement. Yet, some think that the penalty should have been more severe and ultimately will not change data broker behavior.

The steps taken by the Executive to articulate AI’s collective harms (and benefits) are encouraging and to a large extent, they reflect the concerns of most Americans as the Pew studies demonstrate. One can surmise that such activity was motivated in part by a desire to appear in step with the rest of the world. Indeed the 2023 Executive Order was issued just one month before the UK AI Safety Summit. Another goal was likely to highlight the US Congress’ failure to act. Overall, however, these Executive actions will have a limited effect.

The outlook for Congressional action on the pending bills discussed above is bleak. Political differences have made it difficult to enact even the most non-controversial measures. Those differences will be exacerbated in 2024, an election year. Moreover, the strong technology lobby has consistently opposed regulation despite their public statements acknowledging the need for increased data protection and transparency.

This is not good news for the many Americans that support the enactment of laws protecting their personal data and their privacy. Companies face uncertainties even as they continue to develop AI applications. Some have created internal processes for using data, taking into account existing state and federal laws, best practices, and assumptions about future regulation based on the current discourse. At a time when the EU is moving forward with the AI Act, which will bring needed transparency to the development and use of artificial intelligence, the United States remains a passive observer.

7. Conclusion

This paper examines the role of data brokers in collecting, aggregating and selling personal information, usually without users’ knowledge and consent and attempts to demonstrate the need for effective measures regulating data broker behavior. The recent Executive Branch actions and orders around AI and data brokers’ non-US transactions are encouraging, but their effect is limited. This paper also addresses the limits of informed consent on the use of personal information in language resources and suggests a solution in an holistic approach to data protection and privacy across the data/development lifecycle.

8. Ethics Statement

This paper describes ethical issues implicated by the practices of data brokers and data aggregators as well as general ethical issues around data protection and privacy. It does not present any output, formula or suggestion that can be implemented in an unethical manner.

9. Bibliographical References

Auxier, B. and Rainie, L. (2019). Key takeaways on Americans’ views about privacy, surveillance and data-sharing, https://www.pewresearch.org/short-reads/2019/11/15/key-takeaways-on-americans-views-about-privacy-surveillance-and-data-sharing/, accessed 1 March 2024.

Ayoub, E., and Goitein, E. (2024). Closing the Data Broker Loophole, https://www.brennancenter.org/our-work/research-reports/closing-data-broker-loophole, accessed 28 February 2024.

Belanger, A. (2024). Data broker allegedly selling de-anonymized info to face FTC lawsuit after all, https://arstechnica.com/tech-policy/2024/02/data-broker-selling-de-anonymized-info-to-face-ftc-lawsuit-after-all/, accessed 28 February 2024.

Boenig-Liptsin, M., Tanweer, A. and Edmundosn, A. (2022) Data Science Ethos Lifecycle: Interplay of Ethical Thinking and Data Science Practice. *Journal of Statistics and Data Science Education*, 30(3) : 228-240.

Bousquette, I. (2024). AI Is Moving Faster Than Attempts to Regulate It. Here’s How Companies Are Coping., https://www.wsj.com/articles/ai-is-moving-faster-than-attempts-to-regulate-it-heres-how-companies-are-coping-7cfd7104 accessed 28 March 2024.

Carpenter v. United States, 585 U.S. ___, 138 S.Ct. 2206 (2018).

Clos, J., McClaughlin, E., Barnard, P., Nichele, E., Knight, D., McAuley, D. and Adolphs, S. (2022).

PriPA : A Tool for Privacy-Preserving Analytics of Linguistic Data. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 73–78, Marseille, France. European Language Resources Association.

Cubeiq. (2024). Cubeiq's Data for Good Program: Where We've Been and Where We're Going, https://www.cuebiq.com/resource-center/resources/cuebiqs-data-for-good-program-where-weve-been/, accessed 1 March 2024.

Dell, C. (2024). The Sad Truth of the FTC's 'Historic' Privacy Win, https://www.wired.com/story/ftc-xmode-outlogic-location-data-settlement/, accessed 29 February 2024.

deMontjoye, Y., Radaelli, L., Singh, V.K. and Pentland, A. R. (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* 347 (6221): 536-539.

Department of Health and Human Services. (2019). Protection of Human Subjects. 45 CFR Part 46.

DiPersio, D. (2018). A US Perspective on Selected Legal and Ethical Issues Affecting the Development of Language Resources and Related Technology. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18), W21, Legal Issues and Ethics*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

DiPersio, D. (2022). Data Protection, Privacy and US Regulation. In *Proceedings of the Workshop on Ethical and Legal Issues in Human Language Technologies and Multilingual De-Identification of Sensitive Data In Language Resources within the 13th Language Resources and Evaluation Conference*, pages 9-16, Marseille, France. European Language Resources Association.

EU General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), OJ 2016 L 119/1.

Faverio, M. (2023). Key findings about Americans and data privacy, https://www.pewresearch.org/short-reads/2023/10/18/key-findings-about-americans-and-data-privacy/, accessed 21 March 2024.

Federal Trade Commission. (2024). FTC Order Prohibits Data Broker X-Mode Social and Outlogic from Selling Sensitive Location Data, https://www.ftc.gov/news-events/news/press-releases/2024/01/ftc-order-prohibits-data-broker-x-mode-social-outlogic-selling-sensitive-location-data, accessed 28 February 2024.

Gebhart, G. and Richman, J. (2023). Science Shouldn't Give Data Brokers Cover for Stealing Your Privacy, *Scientific American*, https://www.scientificamerican.com/article/science-shouldnt-give-data-brokers-cover-for-stealing-your-privacy/ accessed 28 February 2024.

General Data Protection Regulation ((EU) 2016/679) (EU GDPR) as it forms part of the law of England and Wales, Scotland and Northern Ireland by virtue of section 3 of the European Union (Withdrawal) Act 2018 and as amended by Schedule 1 to the Data Protection, Privacy and Electronic Communications (Amendments etc) (EU Exit) Regulations 2019 (SI 2019/419).

General Services Administration. (2019). GSA Rules of Behavior for Handling Personally Identifiable Information (PII), https://www.gsa.gov/directives-library/gsa-rules-of-behavior-for-handling-personally-identifiable-information-pii-2# accessed 29 March 2024.

Information Commissioner's Office, What common issues might come up in practice?, https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/individual-rights/the-right-to-be-informed/what-common-issues-might-come-up-in-practice/ accessed 28 February 2024.

Kamocki, P. and Witt, A. (2020). Privacy by Design and Language Resources. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3423–3427, Marseille, France. European Language Resources Association.

Keegan, J. (2023). Life360 Sued for Selling Location Data, https://themarkup.org/privacy/2023/06/01/life360-sued-for-selling-location-data, accessed 1 March 2024.

Kochava. (2024). Empowering Marketers and Publishers, https://www.kochava.com/, accessed 1 March 2024.

Lyaskivskij, I. (2024). GDPR requirements to selling of personal data, https://legalitgroup.com/en/gdpr-requirements-to-selling-of-personal-data-ccpa-vs-gdpr-on-insurance-and-trade/, accessed 28 February 2024.

Suresh, H. and Guttag, J. (2021). A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Proceedings of EEAMO '21: Equity and Access in Algorithms, Mechanisms, and Optimization (EEAMO '21)*. ADM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3465416.3483305.

Sweeney, L. (2000). Simple Demographics Often Identify People Uniquely. Pittsburgh, Pennsylvania: Carnegie Mellon University, Data Privacy Working Paper3.

The White House. (2022). Blueprint for an AI Bill of Rights, https://www.whitehouse.gov/ostp/ai-bill-of-rights/, accessed 9 February 2024.

The White House. (2024). Executive Order on Preventing Access to Americans' Bulk Sensitive Personal Data and United States Government-Related Data by Countries of Concern, https://www.whitehouse.gov/briefing-room/presidential-actions/2024/02/28/executive-order-on-preventing-access-to-americans-bulk-sensitive-personal-data-and-united-states-government-related-data-by-countries-of-concern/ accessed 2 March 2024.

Thomson Reuters. (2023). How President Biden’s executive order on AI impacts the legal sector, https://legal.thomsonreuters.com/blog/how-president-bidens-executive-order-on-ai-impacts-the-legal-sector/, accessed 20 February 2024.

Timelex. (2020). What Should One Keep In Mind When Selling, Purchasing, Or Licensing Personal Data, https://www.timelex.eu/en/blog/what-keep-in-mind-selling-purchasing-licensing-personal-data accessed 28 February 2024.

Veraset. (2024). Your trusted partner for location data, https://www.veraset.com/about/data-industry, accessed 1 March 2024.

What Can I Do with this Data Point? Towards Modeling Legal and Ethical Aspects of Linguistic Data Collection and (Re-)use

Annett Jorschick[†‡], **Paul T. Schrader**[§‡], **Hendrik Buschmeier**[†‡]
[†] Bielefeld University, Faculty of Linguistics and Literary Studies
[§] Bielefeld University, Faculty of Law
[‡] Bielefeld University, SFB 1646 ‘Linguistic Creativity in Communication’
Bielefeld, Germany
{annett.jorschick|paul.schrader|hbuschme}@uni-bielefeld.de

Abstract

Linguistic data often inherits characteristics that limit open science practices such as data publication, sharing, and reuse. Part of the problem is researchers’ uncertainty about the legal requirements, which need to be considered at the beginning of study planning, when consent forms for participants, ethics applications, and data management plans need to be written. This paper presents a newly funded project that will develop a research data management infrastructure that will provide automated support to researchers in the planning, collection, storage, use, reuse, and sharing of data, taking into account ethical and legal aspects to encourage open science practices.

Keywords: linguistic data collection, language resources, open data, informed consent, legal tech

1. Introduction

A growing emphasis on transparency in research processes and improved quality assurance in scientific research has underscored the importance of ‘open science’ and the publication of accompanying research data. New standards for data sharing and reuse have been established, e.g., in the development of the FAIR data principles (Wilkinson et al., 2016). However, linguistic data often has inherent characteristics that makes sharing difficult from a legal and ethical perspectives.

Linguistic data encompass a wide range of data types characterized by their diversity. These data comprise different modalities (from written texts, to spoken audio, to multimodal video recordings), different settings (e.g., generic data collections, field studies, case studies, experimental setups), different topics of interest (e.g., examining and modeling of language families, production and recognition processes, language acquisition, diagnostics and therapy of speech disorders), and involve different groups of participants (e.g., ‘standard’ speakers and listeners, people with speech disorders, vulnerable speakers such as children). Crucially, linguistic data often contains personal information or comes from vulnerable speaker groups and thus requires careful handling in collection, storage, sharing, and reuse practices. Furthermore, anonymization of linguistic data is not always possible, as researchers may be particularly interested in aspects that are inherently personal (e.g., multimodal behaviors such as gestures, facial expressions, or gaze direction shown in video recordings of participants interacting in face-to-face dialogue). However, beyond the issues of transparency, quality assurance and reproduciblity, being able to share linguistic data would be very useful more generally. The reuse of linguistic data is desirable because it is a valuable resource. Collecting, preparing, and annotating linguistic data is time-consuming and labor-intensive given the meticulous collection methods. More importantly, from a scientific (and also cultural) perspective, linguistic data are valuable because they provide a unique record of linguistic diversity (across languages, speakers, geography and time), the existence and preservation of which is a prerequisite for various areas of linguistics.

Given the sensitivity of specific linguistic data, it is critical to address the legal aspects of data collection and use, and to ensure that data publication is covered by participant consent. Because legal, ethical, and privacy considerations are often intertwined, they should be taken seriously from the outset. Addressing these complexities requires careful planning of data collection, including the preparation of appropriate consent forms and technical and organizational data security measures. However, it is often challenging for researchers to anticipate all relevant considerations (especially also those for re-use), adequately address the regulatory framework, and effectively manage repetitive procedures. There is a need to explore the potential of automation to support researchers in meeting these challenges.

In this paper, we describe the infrastructure project INF (‘User-oriented research infrastructure assisting linguistic data collection and (re-)use’) of the newly funded Collaborative Research Center SFB 1646 ‘Linguistic Creativity in Communication’ at Bielefeld University, Germany. The project will develop and implement a research data management infrastructure for the collection, storage, use, reuse, and sharing of different types of linguistic data, along with automated support for the ethical and legal considerations and challenges involved. Each data 'point' (which in this case could range from a single response to all the data generated by a participant in an experiment) is thoroughly characterized (via metadata) to enable researchers to quickly determine permissible operations such as analysis, automated processing, and sharing. To provide maximum support to researchers, all steps in the data life cycle and data organization are automated as much as possible. This includes automated generation of customized consent forms, open science guidelines, and methodological support checklists. By automating these aspects, researchers will be equipped with the tools and resources they need to navigate the complex landscape of data management and (re)use with confidence and ease. This not only facilitates compliance with ethical and legal standards, but also fosters a culture of transparency and reproducibility within the scientific community.

A number of projects have developed approaches to support the ethics application process and the creation of consent forms. Hannesschläger et al. (2020) and van den Heuvel et al. (2020) describe tools that facilitate the creation of GDPR-compliant (GDPR, 2016) consent forms (see sec. 2.1 for a detailed review of the relevant legal norms). 'Ethiktool' (Bendixen et al., 2023) is a computer program that supports researchers in creating applications to internal ethics review boards (IRBs) as well as participant information.

The diversity of linguistic data collection efforts, however, may present particular challenges, which complicate the use of standardized text modules for consent forms, ethics applications, or data management plans. Moreover, consent forms composed of automated components are often lengthy and legal texts are difficult to comprehend, particularly for participants who struggle with language. This is a fundamental problem, as participants need to be well-informed to provide informed consent. This complexity is further exacerbated when considering data reuse and sharing. The results of the replication crisis (Open Science Collaboration, 2015) demonstrate that sustainable research and open science practices are not merely optional methods but essential additional research goals. As explained above, adherence to the FAIR principles (Wilkinson et al., 2016) is particularly relevant for linguistic data.

The project aims to address these methodological, legal and technical issues through an interdisciplinary effort – comprising (psycho-)linguists (for the expertise in linguistic data collection and preparation), legal experts (for adherence to regulatory frameworks), and computational scientists (for modeling and building the technical infrastructure).

2. Concept

The project will consider the ethical and legal regulations (as will be outlined in sec. 2.1), as well as the technical framework (see sec. 2.2), to model the life cycle of linguistic data in a research data management infrastructure composed of three fundamental modules: (1) a data collection setup wizard guiding researchers to setup studies, which automatically generates customized consent forms, potentially guiding the creation of ethics applications, and data management plans; (2) a computational platform where all information about studies and their associated data are collected; (3) a search engine allowing data queries and sharing in a way that is consistent with the individualized consent (opt-in/out) provided by participants.

The data life cycle begins as early as the planning phase of the linguistic study. As outlined above, the creation of appropriate informed consent forms and the data management plan necessitate information about the use, sharing, and potential reuse of the data. To assist researchers in this initial stages of study planning, we will develop a 'Wizard' tool that guides users through the setup and design process, aiming to maximize the application of the FAIR principles (Wilkinson et al., 2016) wherever feasible. The main outcome of this wizard is a highly customized informed consent form tailored to the specific study and adapted to the requirements of it's participants, utilizing combinable text blocks, that are ethically and legally sound and coherent.

The underlying technical platform stores information collected from each study (including individual consent information), converts details into a standardized metadata format (Broeder et al., 2012), and subsequently integrates the information gathered during the study design phase with the collected data in order to model the permissible operations that can be done with a data point based on participants' consent.

To facilitate data reuse and collaboration, a search engine will be implemented to query and retrieve available data resources. It will enable researchers to easily identify and access relevant data 'points' for their specific research needs. Advanced search functionalities and metadata indexing will enable users to filter and refine search results based on various detailed criteria, such as data type, topic, accessibility status, and permissions for usage.

To achieve this goal, a careful analysis of the various legal norms (see sec. 2.1) will identify potential contradictions and redundancies between requirements that researchers may have. In addition, the objective requires a comprehensive review of legal doctrines, their interpretations, and relevant judicial precedents. Based on this, text blocks are created that can be assembled into individualized consent forms. It is important to ensure the use of easy-to-understand language to enhance participant's comprehension, while avoiding excessive detail that may discourage participation due to length. In addition, as text blocks are combined within the consent form, it is critical to avoid redundancy, so the computational generation process must cross-check building blocks as they are uniquely assembled and resolve redundancies to maintain clarity and conciseness in the final document.

2.1. Legal and Ethical Regulations

The creation of a wizard tool to support researchers through the automated creation of data protection declarations and subsequent storage in a data management system requires a precise analysis of the legal and ethical framework conditions. Data collection in the field of empirical research is subject to various regulations.

Since the Helsinki Declaration in 1964 (World Medical Association, 2013) and the Belmont report in 1978 (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, 1979), ethical considerations have become an increasingly important aspect of empirical research. In cooperation with the Ethics Committee of Bielefeld University, we will use the German Psychological Society's guidelines (DGPs, 2022) as the institutional basis for modeling consent forms. These guidelines are used as a framework for ethics applications in many universities in Germany. Moreover, the German higher education framework act (Hochschulrahmengesetz; HRG, 1999) encompasses provisions that specifically address research activities conducted within academic institutions, ensuring that research adheres to legal standards and ethical principles.

In addition to ethical considerations, ensuring compliance with various legal frameworks is essential for maintaining data security, proper data storage, facilitating data reuse, and regulating data sharing practices. Data protection regulations include, in particular, the European General Data Protection Regulation, which has been in force since 2018 (GDPR, 2016). As an European regulation, it is directly legally valid in the member states and also applies to public institutions. The GDPR is supplemented by the national BDSG (BDSG, 2017) and the DSG-NRW (DSG-NRW, 2018). The BDSG only applies if the law of the European Union, in particular the GDPR, does not apply directly (§ 1 V BDSG). A further restriction of the scope of application of the law can be found in § 1 I S. 1 Nr. 2 BDSG: The data protection law of the federal states (DSG-NRW) applies to the processing of personal data by public organizations of the federal states. However, the GDPR forms the overarching legal framework. As part of project INF, linguistic research data is to be collected, stored and processed. In certain cases, data processing without explicit consent may be lawful if it is necessary for the performance of a task carried out in the public interest (Art. 6 I lit. e) GDPR). This may apply in accordance with § 17 I DSG-NRW for data processing in the scientific field. Under certain circumstances, consent may not be required. The rights of access (Art. 15 GDPR), rectification (Art. 16 GDPR), restriction of processing (Art. 18 GDPR) and objection (Art. 21 GDPR) may also be restricted if the exercise of these rights is likely to render impossible or seriously impair the realization of the research or statistical purposes and the restriction of these rights is necessary for the fulfillment of these purposes (§ 17 V DSG-NRW). However, the consent of the study participants is generally required (Art. 6 I lit. a) GDPR). This consent is subject to certain legal requirements. Above all, it must be given freely, specifically and for a particular purpose (Art. 6 I lit. a) GDPR). Any use of the data outside of this purpose limitation is generally not permitted and may only take place with the renewed consent of the data subject.

In the scientific field, the purpose of the processing of personal data cannot usually be fully specified at the time of data collection. In order to take sufficient account of the special requirements of data collection in scientific research, Art. 5 I lit. b) GDPR provides for a relaxation of this rigid purpose limitation. However, data subjects still have the option of restricting their consent to certain research areas or parts of research projects (Recital 33 GDPR). Special requirements apply to consent in relation to data processing of particularly sensitive groups such as children (Art. 8 GDPR) or the processing of particularly sensitive data such as health data (Art. 9 GDPR). For example, when collecting health data, the consent to be given must contain an explicit reference to the processing of sensitive data. This information should enable the data subject to make an individual risk assessment. In general, it should be noted that consent can be withdrawn at any time and the possibility of withdrawal must be pointed out (Art. 7 III GDPR). With regard to the legally compliant use of the planned wizard, the legal requirements described here must be included in the module to be programmed. Particular attention is paid to the specific requirements for data collection in the field of scientific research.

2.2. Technical Framework

The establishment of a robust research data management infrastructure as outlined above involves several key stages. First, the integration of the framework into the infrastructure of Bielefeld University has to be considered, while ensuring compatibility with national and international data and metadata standards and repositories. Second, security standards have to be implemented to ensure privacy and integrity of (meta)data. Finally, the source code of the infrastructure will be made openly available to encourage adoption and facilitate collaboration within the scientific community.

In order to ensure a integration across the university and compatibility between different systems, a careful consideration of existing services and interfaces is essential. This includes harmonizing research data management processes with existing platforms at Bielefeld University (such as RDMO, Gitlab.UB, PUB). These platforms support various aspects of research data management, from planning and documentation to version control and (data) publication. By integrating these interfaces, researchers will be able to streamline data management workflows and improve accessibility and reproducibility within the university.

In establishing effective research data management practices, it's important to consider both regional and global standards and initiatives. This includes adherence to metadata schemata for interoperability (Broeder et al., 2012) and engagement with platforms such as the Registry of Research Data Repositories. Additionally, infrastructures such as CLARIN, CLARIN-D and CMDI enhance accessibility and usability of linguistic resources. By aligning with these initiatives, the linguistic research data published in the platform/infrastructure will increase its visibility and impact on an international scale.

Security guidelines are important for software development because they ensure the integrity, confidentiality, and availability of data and systems. Standards such as ISO/IEC 27001 (ISO/IEC, 2022) provide a framework for establishing, implementing, maintaining, and continually improving an information security management system. This includes defining security policies, conducting risk assessments, implementing controls, and monitoring and auditing security measures. By following these standards and incorporating security best practices into software development processes, we can mitigate security risks and protect sensitive information from potential threats and vulnerabilities.

3. Embedding and Perspectives

The Collaborative Research Center SFB 1646 'Linguistic Creativity in Communication' comprises 16 research projects using different empirical methods to collect a diverse set of linguistic data. These methods include historical data analysis, corpus collection in various modalities (e.g., auditory, multimodal), and experimental investigations with diverse speaker groups. This rich research environment will allow project INF to comprehensively model linguistic data collection and usage practices, and to generate a wide range of use cases for the building block inventory for consent form generation. At the same time, the research center will be able to immediately benefit from (and influence) the creation of the infrastructure developed within INF that enables its open science and open data objectives.

In this first funding phase of the Collaborative Research Center, INF will focus on the implementation of the data management platform and its integration into the infrastructure of Bielefeld University. Data sharing with national and international infrastructures is planned for the second funding phase and will only be pursued after a thorough evaluation of the software security.

Acknowledgments

INF (537522983) is supported by the Deutsche Forschungsgemeinschaft (DFG) via the Collaborative Research Center SFB 1646 'Linguistic Creativity in Communication' (512393437).

Bibliography

BDSG. 2017. Gesetz zur Anpassung des Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680 (Datenschutz-Anpassungs- und -Umsetzungsgesetz EU – DSAnpUG-EU). *Bundesgesetzblatt Teil I*, 2017(44):2097–2132.

Alexandra Bendixen, Thomas G.G. Wegner, and Wolfgang Einhäuser. 2023. Facilitating ethics application and review for interdisciplinary human-participant research via software-based guidance and standardization. In *1st International Conference on Hybrid Societies*, Chemnitz, Germany.

Daan Broeder, Menzo Windhouwer, Dieter van Uytvanck, Twan Goosen, and Thorsten Trippel. 2012. CMDI: A component metadata infrastructure. In *Proceedings of the LREC 2012 Workshop on Describing Language Resources with Metadaa*, pages 1–4, Istanbul, Turkey.

DGPs. 2022. Berufsethische Richtlinien DGPs / BDP. Föderation Deutscher Psychologenvereinigungen.

DSG-NRW. 2018. Gesetz zur Anpassung des allgemeinen Datenschutzrechts an die Verordnung (EU) 2016/679 und zur Umsetzung der Richtlinie (EU) 2016/680 (Nordrhein-Westfälisches Datenschutz-Anpassungs- und Umsetzungsgesetz EU – NRWDSAnpUG-EU). *Gesetz- und Verordnungsblatt (GV. NRW.)*, 2018(12):243–268.

GDPR. 2016. Regulation (EU) 2016/679 of the European Parliament and of the Council. *Official Journal of the European Union*, L 119:1–88.

Vanessa Hannesschläger, Walter Scholger, and Koraljka Kuzman. 2020. The DARIAH ELDAH consent form wizard. In *DARIAH Annual Event 2020: Scholarly Primitives*.

HRG. 1999. Hochschulrahmengesetz. *Bundesgesetzblatt Teil I*, 1999(3):18–34.

ISO/IEC. 2022. ISO/IEC 27001:2022 – Information security, cybersecurity and privacy protection – Information security management systems – Requirements. Standard 27001:2022, International Organization for Standardization (ISO), Geneva, Switzerland.

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. 1979. The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research. *Federal Register*, 44(76):23192–23197.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716.

Henk van den Heuvel, Aleksei Kelli, Katarzyna Klessa, and Satu Salaasti. 2020. Corpora of disordered speech in the light of the GDPR: Two use cases from the DELAD initiative. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3317–3321, Marseille, France.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

World Medical Association. 2013. World Medical Association Declaration of Helsinki: Ethical principles for medical research involving human subjects. *JAMA*, 310:2191–2194.

Data-Envelopes for Cultural Heritage: Going beyond Datasheets

Mrinalini Luthra, Maria Eskevich
Huygens Insitute, Koninklijke Nederlandse Akademie van Wetenschappen (KNAW)
Oudezijds Achterburgwal 185, 1012 DK Amsterdam
{mrinalini.luthra, maria.eskevich}@huygens.knaw.nl

Abstract

Cultural heritage data is a rich source of information about the history and culture development in the past. When used with due understanding of its intrinsic complexity it can both support research in social sciences and humanities, and become input for machine learning and artificial intelligence algorithms. In all cases ethical and contextual considerations can be encouraged when the relevant information is provided in a clear and well structured form to potential users before they begin to interact with the data. Proposed data-envelopes, basing on the existing documentation frameworks, address the particular needs and challenges of the cultural heritage field while combining machine-readability and user-friendliness. We develop and test data-envelopes usability on the data from the Huygens Institute for History and Culture of the Netherlands.

Keywords: machine-readable datasheets, cultural heritage, data ethics, transparency, auditability, FAIR

1. Introduction

The digitisation of historical collections presents opportunities for research and education, transforming how we understand and access the past, and define how the future for the past can be collectively shaped (Trouillot, 2015; McGillivray et al., 2020). However, this digital transformation is accompanied by significant ethical, legal, and practical challenges, especially as historical datasets become critical resources for not only academic scrutiny but also serve as fuel for advanced computational models. The complexity of these challenges necessitates a robust framework to guide the use of cultural heritage (CH) data, ensuring their accessibility, transparency, and ethical (re)use.

In response to these challenges, this paper presents the following contributions: i) we highlight the complexity of CH data, featuring the unique ethical and contextual considerations they entail on the example of materials that are offered by Huygens Institute; ii) we evaluate and compare existing dataset documentation frameworks, examining their suitability for CH datasets; iii) we introduce the "data-envelope"–a machine-readable adaptation of existing dataset documentation frameworks, to tackle the specificities of CH datasets. Its modular form is designed to serve not only the needs of machine learning (ML), but also and especially broader user groups varying from humanities scholars, governmental monitoring authorities to citizen scientists and the general public. Importantly, the data-envelope framework emphasises the legal and ethical dimensions of dataset documentation, facilitating compliance with evolving data protection regulations and enhancing the accountability of data stewardship in the cultural heritage sector. We discuss and invite the readers for further conversation on the topic of ethical considerations, and how the different audiences should be informed about the importance of datasets documentation management and their context.

2. Diversity in Cultural Heritage Data

In this section, we delve into the multifaceted nature of CH data, emphasising the specific ethical and contextual considerations that it necessitates. By examining data from the Huygens Institute for History and Culture of the Netherlands[1], part of the KNAW Humanities Cluster[2] we illustrate three key aspects of CH data: the extensive historical range of the collections, the unique contexts of their creation and aggregation, and the intricate data structures within these datasets. This institution is selected for its representative practices and data interaction types that are common within the CH sector in the Netherlands, suggesting that solutions identified here may be applicable more broadly.

The collections managed by the Huygens Institute showcase the evolution of CH data from physical to digital realms. Initially, data were selected and published in book form, starting in 1902 (Kooijmans and de Valk, 1985). This historical approach laid the groundwork for contemporary digital projects such as GLOBALISE[3], Oorlog voor de Rechter[4] [War in Court], and REPUBLIC[5]. These initiatives reflect the shift towards digital accessibil-

[1] https://www.huygens.knaw.nl/en/

[2] KNAW is an abbreviation of the Koninklijke Nederlandse Akademie van Wetenschappen [Royal Dutch Academy of Arts and Sciences]

[3] https://globalise.huygens.knaw.nl

[4] https://oorlogvoorderechter.nl

[5] https://republic.huygens.knaw.nl

LEGAL2024 @LREC-COLING-2024, pages 52–65

ity and the ongoing efforts to process and release data.

The nature of CH datasets, often spanning over centuries, is distinguished not only by their historical depth but also by their collection and selection processes. These processes, historically influenced by various biases, shape the datasets' structure and content. Digital historians and scholars, equipped with a deep understanding of the field's evolution and ongoing debates, approach these datasets critically, mitigating biases through careful analysis (Tasovac et al., 2020; Maryl et al., 2023). This scholarly perspective informs the data's structure, metadata quality, and its application, diverging from the requirements commonly associated with machine learning and artificial intelligence (AI) disciplines (Heger et al., 2022). Therefore, dataset documentation could benefit from integrating such rich contextual information, ensuring CH data is utilised responsibly and effectively in technological applications (Jo and Gebru, 2020).

Research in (digital) humanities utilises complex data structures and/or interconnected datasets to deepen historical understanding and introduce new insights into past events. On the one hand, scholars navigate numerous challenges, including handling low-resource languages, accommodating spelling variations, and correcting text recognition errors (Koolen et al., 2023). The diversity of document types and domains, coupled with language evolution and noisy inputs, further complicates analysis. On the other hand, the information about the same entity, such as a migrant person, can be scattered across different registries, archives, and other official documents as well as informal records collected by civil society organisations, churches, and other non-governmental organisations, thus varying in form, structure, and availability (Arthur et al., 2018). Moreover, the research might combine both analysis of the content of particular types of documents such as letters, and the way the communication was evolving through the network analysis (Hotson, 2019). To effectively utilise these rich historical resources, data brokers must provide comprehensive, accessible information on data limitations and considerations, ensuring users can fully engage with the historical context.

2.1. Retrodigitised Editions

Building on the exploration of the complexities of CH data, this subsection explores the specific case of retrodigitised editions. Historians frequently engage with these editions, which are historical documents compiled and commented on in book form, later digitised for broader access (Kooijmans and Th.S. Bos, 1985; Tollebeek, 1994). This process exemplifies the transformation of CH data across formats (varying from the actual physical instance to a plethora of data representations), highlighting the necessity for clear documentation on annotation and content transformation decisions. Such detailed documentation is crucial for users to understand the historical context and the interpretative layers added through digitisation, further illustrating the challenges and limitations of existing documentation frameworks mentioned above.

2.2. GLOBALISE: Commodities Dataset

The GLOBALISE project, focused on leveraging AI to transcribe and extract data from the Dutch East India Company (VOC) archives (Petram and van Rossum, 2022), underscores the limitations of current dataset documentation standards in digital cultural heritage. For instance, the documentation of the commodities dataset (Pepping et al., 2023)—detailing classifications and a thesaurus of commodities traded in the early modern Indian Ocean World—highlights these gaps. Existing templates fail to adequately capture the complexity of the provenance inherent in such datasets, derived from primary sources and enriched through the multiple secondary sources. Furthermore, they fall short in addressing the linguistic diversity and temporal scopes, both being crucial aspects for accurately documenting digital cultural heritage data.

2.3. Potential Legal and Ethical Issues with Huygens Institute Resources

Institutions such as the Huygens Institute combine running projects with managing access to legacy datasets (more than 200 in this case) which brings a lot of potential legal and ethical issues along the way:

- *Copyright:* In principle, within this institute copyright is less of an open issue, even though different copyright regulations apply to the datasets. In a lot of cases the copyright stays with the institute, as it publishes or has published the materials (Kooijmans and Th.S. Bos, 1985).

- *Licenses:* As Large Language Models (LLMs) increasingly rely on structured datasets for training, it is crucial to consider the potential risks associated with using cultural heritage data. Given the historical intricacies and biases inherent in cultural heritage data, there is a danger that LLMs trained on such datasets may inherit these biases. When applied in contemporary contexts, these models may perpetuate discriminatory practices and reinforce historical prejudices. Moving forward, it is important to develop strategies for mitigating the potential risks of bias amplification when making cultural heritage datasets available for LLM training (Hicks, 2017; Thylstrup, 2019; Noble, 2018). Additionally, cultural heritage institutions need to navigate and rethink the landscape of intellectual property rights and openness in the era of generative AI. These institutions may need to adopt a more nuanced approach, differentiating between private users, researchers, and commercial entities, while also renegotiating license agreements and addressing technical challenges related to copyright protection in the context of AI (Lehmann, 2024).

- *Privacy:* There is a number of projects that make public and digitally accessible personal information which requires special attention and contextualisation. For example, the project “Oorlog voor de Rechter" (War in Court) aims at disclosure of archival documents about collaboration during Second World War[6], and another project, “Child Separation", works with the information about extraction of children from their indigenous context and putting them into foster care (Mak et al., 2020).
- *Information security:* Historical documents reflect different aspects of the past within the country, and when accessed and processed without proper contextualisation, this information might provoke wrongful assumptions, statements, and even prosecutions.
- *Ethical and Emotional:* Such considerations are particularly poignant in cases involving information about living individuals, such as data related to wartime collaboration or child adoption, which require sensitive handling to mitigate potential harm or distress (Wood et al., 2014).

3. Harmonising Machine Learning, Cultural Heritage, and Legal Insights

In the rapidly evolving digital landscape, the documentation of CH datasets emerges as a critical juncture where machine learning practices, cultural heritage stewardship, and legal compliance intersect. This section delves into the existing documentation frameworks, underscoring the limitations within machine learning paradigms, the unique complexities of cultural heritage data, and the increasing importance of aligning with legal standards. Through this examination, we highlight the imperative for a nuanced, comprehensive approach to dataset documentation that is answered by the proposed data-envelope framework.

3.1. ML perspective

Dataset documentation, often referred to as "datasheets", first introduced by Gebru et al. (2021) advocates for the inclusion of comprehensive documentation alongside machine learning dataset publications. Such documentation is envisioned to serve multiple critical functions: facilitating informed decision-making regarding dataset application, enhancing transparency concerning the datasets' composition and creation, and establishing clear guidelines for dataset development (Gebru et al., 2021; Pushkarna et al., 2022; Library of Congress, 2021; Roman et al., 2023).

3.2. CH perspective

The complexity inherent in (digital) cultural heritage data transcends the technical dimensions typically addressed by machine learning documentation standards. These datasets are situated within diverse social, cultural, and historical contexts, often encompassing multiple perspectives and interpretations (Cameron and Kenderdine, 2007) as demonstrated in Section 2. The temporal and spatial complexity of the data adds another layer of challenge, as does the presence of uncertainties and incompleteness. Furthermore, cultural heritage data is often subject to copyrights, traditional knowledge, and intellectual property considerations (Torsen and Anderson, 2010). The collaborative nature of knowledge production in this domain necessitates careful attribution and recognition of contributors (Srinivasan et al., 2010; Powell, 2016).These factors collectively underscore the need for documentation practices that can adequately capture and convey the nuances and complexities of cultural heritage data (Candela et al., 2023).

A recent paper by the Datasheets for digital cultural heritage Working Group, set up within the Europeana Research Community and EuropeanaTech Community, has made a first attempt to documenting datasets from the cultural heritage sector (Alkemade et al., 2023). However, these initial steps, while pioneering, reveal gaps in usability, machine-readability, and the depth of coverage on critical issues like provenance, ethical, and legal considerations.

3.3. Legal Perspective

The legal landscape around data use and governance is undergoing significant transformation on both international and national levels. Legislations such as the EU Data Act[7], EU Data Governance

[6] https://oorlogvoorderechter.nl/

[7] https://digital-strategy.ec.europa.eu/en/policies/data-act

Parameter	Datasheets	Data Cards	Open Datasheets	Datasheets for DCH	Data-Envelope
Structure	Questionnaire format	Structured Summaries	JSON-based metadata	Tailored for DCH data	Modular with detailed sections
Machine Readability	Not primary focus		Yes, fully supported	Not primary focus	Yes, Designed for machine readability
Provenance	Not explicitly/sufficiently considered				Extensively covered
Target Audience	ML/AI researchers			ML/AI researchers, CH Institutions	CH Institutions, ML community, legal institutions, broader public
FAIR	Not directly addressed				Designed with FAIR in mind, with specific section devoted to datasets' adherence to FAIR principles
Positionality	Not emphasized, only mentioned for annotators				Explicit focus on creators, contributors, annotators' positionality

Act[8], and EU Artificial Intelligence (AI) Act[9] on the EU level and Archiefwet [Law about archiving in the Netherlands][10] introduce complex requirements for dataset documentation, transparency, and accountability[11]. However, in practice the current lack of standardised, machine-readable documentation frameworks complicates the actual compliance and auditing processes. Our contribution lies in the development of a comprehensive machine-readable documentation framework, which enables automated auditing of datasets, particularly in areas concerning data collection, sharing, and (re)use. By bridging the gap between legal requirements and technical documentation, the proposed data-envelope facilitates compliance with regulatory mandates, thereby enhancing transparency and accountability in data governance practices.

3.4. Advancing Documentation Practices

Table 3.1 outlines a comparison between different dataset documentation frameworks (Gebru et al., 2021; Pushkarna et al., 2022; Alkemade et al., 2023; Roman et al., 2023). The data-envelope offers a machine-readable structured data alongside qualitative narrative elements, thereby ensuring versatility. This approach not only supports the development of AI models but also addresses the educational and research necessities of cultural heritage, underpinning the importance of a well-rounded, accessible data documentation method. The data-envelope's particular emphasis on positionality (Harding, 2003; Haraway, 2016; Mignolo and Walsh, 2018) and adherence to FAIR principles (Wilkinson et al., 2016; Harrower et al., 2020) demonstrates its comprehensive approach to dataset documentation and accessibility.

4. Data-Envelopes for Datasets

We introduce the "data-envelope", intended to provide clear guidance for the creation and documentation of CH datasets, ensuring that their complexity and context are effectively communicated and preserved for current and future use.

4.1. Contextual Wrapper for Datasets

At its core, the data-envelope is conceptualised as a contextual wrapper for datasets. Going beyond existing documentation frameworks (Gebru et al., 2021; Pushkarna et al., 2022), the data-envelope encases the dataset within a comprehensive context that elucidates the cultural, historical, and social dimensions of the data. By situating data within this contextual framework, the data-envelope empowers users to comprehend not just the 'what' but also the 'why' behind the data they engage with. This method guarantees that any interpretations

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020PC0767

[9] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206

[10] https://wetten.overheid.nl/BWBR0007376/2022-05-01

[11] https://www.europarl.europa.eu/RegData/etudes/STUD/2022/729541/EPRS_STU(2022)729541_EN.pdf

and utilisations of the dataset are rooted in an appreciation of its origins and importance, thereby encouraging more informed and thoughtful applications (Mignolo and Walsh, 2018).

In the current data interaction model, depicted in Figure 1, the CH sector oversees the creation and population of datasets, metadata, and datasheets primarily within its own confines. Subsequently, AI/ML algorithms typically ingest only the data and some metadata to generate models and tools, often stripping away valuable context.

The proposed model, illustrated in Figure 2, introduces the data-envelope as a pivotal innovation. Here, it acts as a central hub, harmonising access to comprehensive information and documentation for both CH users and the AI/ML community. This new paradigm aims to enrich AI/ML algorithms with a fuller context, enhancing the quality and applicability of the resulting models and tools.

The axis in both figures represents the amount of contextual information that the users are provided with when having access to the materials: when confronted with the trained model or a working tool they usually have way less context and explanation than when looking at the data and metadata itself. Under the current data interaction model, end-users engaging with AI/ML outputs encounter a notable deficit in context and explanation. In contrast, the data-envelope model facilitates direct access to extensive background information on datasets for more informed use.

4.2. Modular Structure

The data-envelope is structured into modular sections, each designed to encapsulate different facets of the dataset in a systematic manner. The philosophy behind this five-level structure is to provide a comprehensive yet organised representation of the dataset. By separating the information into distinct levels, users can quickly locate the specific details they need without being overwhelmed by a monolithic documentation. The five-level structure, visualised in Figure 3, is elaborated on below, highlighting the basic ideas, philosophy, and differentiation from other templates. Further details about each level of the data-envelope are provided in the Section 8 (Appendix A), offering a more granular view of the specific contents and considerations within each section.

4.2.1. Basic Information/What Goes on the Data-Envelope

This section is dedicated to outlining the core details of both the data-envelope and the dataset it encompasses. It goes beyond traditional documentation practices by introducing a dual-versioning system: one for the dataset and another for the data-envelope itself. Recognising the dynamic nature of dataset documentation, this approach allows for the data-envelope to evolve independently of the dataset, adapting to the changing needs and standards of data management over time.

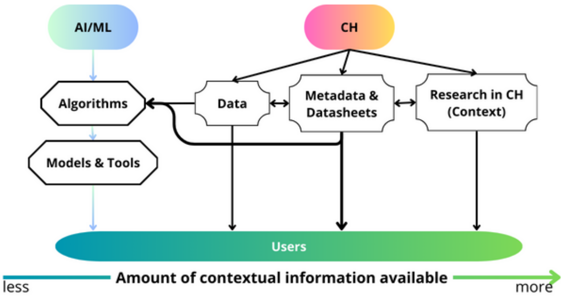

Figure 1: Current data interaction model with CH output (data, metadata, and research) in the context AI/ML development.

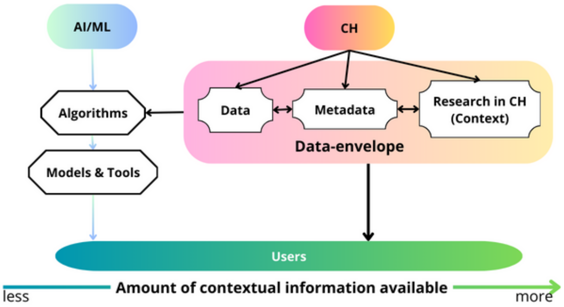

Figure 2: Proposed data interaction model with CH output (via data-envelope) in the context AI/ML development.

Additionally, this segment includes comprehensive contact information for individuals involved in various stages of the project. From conceptualisation and technical implementation to administration and more, users are provided with direct avenues to connect with experts for specific inquiries. This not only enhances the accessibility and transparency of the dataset but also fosters a collaborative environment where users can seek guidance, clarification, or further information as needed.

4.2.2. Basic Dataset Metadata

The Basic Dataset Metadata section conforms to the Data Catalog Vocabulary (DCAT) standards to guarantee compatibility with machine-readable formats (World Wide Web Consortium, 2014).[12] It catalogues key dataset information such as title, identifier, version, and a detailed description, along with the genre and topic classification. This section also outlines the dataset's geographical and temporal scope, essential for situating cultural heritage data within specific contexts.

[12] We refer to the most recent version: DCAT-3.

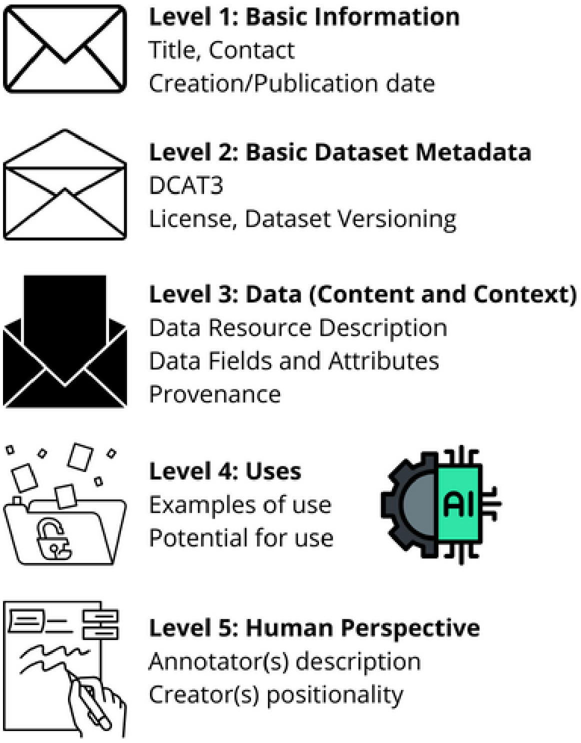
Level 1: Basic Information
Title, Contact
Creation/Publication date
Level 2: Basic Dataset Metadata
DCAT3
License, Dataset Versioning
Level 3: Data (Content and Context)
Data Resource Description
Data Fields and Attributes
Provenance
Level 4: Uses
Examples of use
Potential for use
AI
Level 5: Human Perspective
Annotator(s) description
Creator(s) positionality

Figure 3: Data-envelope structure

Details about the dataset's inception and release provide insight into its relevance. Acknowledgement of contributors affirms transparency and credits those involved. Information on distribution, access, licensing, and maintenance is meticulously presented, equipping users with knowledge about usage conditions. Furthermore, a dedicated subsection ensures compliance with FAIR data principles, emphasising Findability, Accessibility, Interoperability, and Reusability (Wilkinson et al., 2016; Devaraju et al., 2021; Singh et al., in press). This commitment to FAIR principles ensures that the datasets are well-documented and suitable for broader use, aligning with global data management standards.

4.2.3. Data Content and Context

This section addresses the inclusion of diverse resources within the dataset, such as thesauri, reference data, and annotations. It is comprehensive, covering languages, encoding formats, resource creation dates, subjects of the data, modality, and descriptive statistics. It also describes data fields and attributes, presents sensitivity assessments, and provides examples to illustrate common errors and redundancies. Additionally, it details the annotation and labelling processes.

Furthermore, it has an extensive section on data provenance, connecting to additional documentation such as datasheets for sources and handwritten text recognition outputs, annotation instructions, and any other documentation, where available, providing users with supplementary information. Lastly, this section concludes with ethical reviews, social impact assessments, and bias considerations.

4.2.4. Uses

This section encourages dataset creators to introspect and articulate both recommended and discouraged uses of the dataset. It invites consideration of various application contexts, offering a platform for detailed descriptions and linkages to related datasets, publications, and models. This proactive reflection on the dataset's appropriate and inappropriate uses fosters responsible utilisation and helps users understand the boundaries within which the dataset is intended to operate.

4.2.5. Human Perspective

Positionality, rooted in Sandra Harding's (2003) standpoint theory, emphasises that personal backgrounds—encompassing gender, ethnicity, socioeconomic status, and more—influence an individual's knowledge and actions. This idea challenges the belief in objective, absolute truths within scientific research, instead suggesting that knowledge is created within a web of personal and social experiences (Haraway, 2016). Feminist epistemologists have thus argued that acknowledging and integrating positionality into the research can lead to more comprehensive and nuanced understandings (Mignolo and Walsh, 2018; Harding, 2013).

In dataset documentation, embracing positionality is vital for various reasons. Firstly, it illuminates the biases and assumptions that may influence data collection and analysis. Secondly, it provides transparency, allowing users to understand the context in which the dataset was created and to consider how this context may affect their use of the data. Thirdly, it promotes inclusivity by recognising the diverse standpoints of dataset creators and subjects, encouraging a multiplicity of perspectives in data interpretation. While positionality of annotators is becoming common practice (Geva et al., 2019), it is yet uncommon to see mention of positionality of the curators of datasets. The data-envelope will have a dedicated section on positionality of the institutions, projects, and persons involved in dataset creation.

An illustrative example is the work of Dutch linguist Jo Daan. In her seminal 1963 study at the Meertens Institute, Daan did not merely catalog dialects; she contextualised the data within the social dynamics of the speakers (Daan and Meertens, 1963). Her approach to documenting language patterns was inherently tied to the positionality of the communities she studied, pioneering a path in linguistic research that considered the complex interplay of language with social identity and culture. This historical example underscores the depth and richness that positionality can bring to dataset documentation, and why it is increasingly becoming a best practice in the field.

4.3. Machine-Readable Implementation

The development of the data-envelope template is underway, aiming to transform it into a user-friendly, fillable form accessible on a static website. This innovative approach is designed to streamline the process of documenting datasets by allowing users to input detailed information directly into the form. Once completed, the form will enable the download of documentation in formats that are both human-readable and machine-readable.

Inspiration for this model comes from successful implementations such as CFFINIT[13], developed by the Netherlands e-Science Center, which facilitates the creation of citations for software and datasets. Similarly, Microsoft's introduction of the 'Open Datasheet' form, which outputs information in JSON format, exemplifies the potential of such tools in promoting standardised, machine-readable dataset documentation (Roman et al., 2023).

The ultimate goal is for these machine-readable documents to seamlessly integrate with Open Science repositories, like Zenodo[14], facilitating the automatic population of metadata fields. This integration would significantly advance the FAIRness of datasets, making them more discoverable and usable across the scientific community (Wilkinson et al., 2016). Although the practice of automatically integrating machine-readable datasheets into repositories is not yet commonplace, it embodies a progressive strategy to ensure that datasets are not only easily accessible but also thoroughly documented.

Balancing Metrics and Narratives in Cultural Heritage Datasets

The use of metrics and measures in cultural heritage datasets is a topic of ongoing debate. Cultural heritage institutions have a long history of qualitative item and collection descriptions, with minimal reliance on numbers. Historians and humanists are often skeptical of quantitative measures, recognizing their dependence on social context (Urton and Llanos, 1997). In contrast, the machine learning community places great value on descriptive statistics, digitization metrics, and annotation analysis (Alkemade et al., 2023). Resolving this divergence requires a case-by-case approach, selecting metrics based on their value and relevance to the dataset's intended purpose. Dialogue between domain experts, researchers, and tech-savvy individuals is crucial in determining appropriate metrics.

Moving forward, as the authors further develop the data-envelope template, they will consider incorporating controlled vocabularies for sensitive content categories and mitigation measures. This approach aims to facilitate the communication of crucial information in a standardized, machine-readable format while allowing for the inclusion of both quantitative and qualitative information as deemed appropriate for each specific dataset. By striking a balance between metrics and narratives, the data-envelope template seeks to promote transparency, accountability, and ethical considerations in the documentation of cultural heritage datasets.

5. Conclusion, Future Work and Challenges

This paper advocates for a paradigm shift in how we document, use, and understand cultural heritage datasets through the introduction of the data-envelope framework. By addressing the limitations of existing documentation practices and proposing a solution that caters to both technical requirements and broader societal needs, we invite the academic community and stakeholders in the cultural heritage sector to engage in a critical dialogue about the future of dataset documentation. Our work underscores the importance of a multidisciplinary approach to data governance, one that recognises the intricate web of legal, ethical, and practical considerations surrounding the stewardship of cultural heritage in the digital age.

While initially conceived to address the specific challenges of CH data, we argue that the data-envelope framework holds potential for broader applicability across diverse datasets. As many contemporary datasets are inherently socially and historically constructed, our documentation template serves as a valuable tool for enhancing transparency and understanding across various data domains.

The data-envelope template, as presented in the appendix, is a comprehensive framework designed to capture the intricacies of (Digital) Cultural Heritage datasets. As we continue to refine the template through collaborative iterations with diverse research groups within the Huygens Institute, we are actively engaged in a bottom-up approach to finalise the template to fit the needs of diverse projects, datasets, creators, and users. This iterative process involves gathering feedback, identifying common challenges, and adapting the template to ensure its flexibility and applicability across a

[13] https://citation-file-format.github.io/cff-initializer-javascript

[14] https://zenodo.org

wide range of cultural heritage contexts.

5.1. Ethical Considerations and Novelty

The ethical dimensions of this work are twofold. Firstly, the data-envelopes incorporate explicit statements about data bias and (re)use policies, addressing critical ethical concerns in the (re)use of historical datasets. Secondly, by harmonising the differing perspectives of data scientists and legal experts, proposed data-envelopes serve as a bridge between technical and legal frameworks, facilitating a more ethical and legally compliant use of historical data.

5.2. Technical Implementation and Embedding Data-Envelopes into the Data Life-Cycle

The scientific novelty of our approach lies in its emphasis on machine-readability, which not only enhances transparency and trust but also allows for the data-envelopes to be easily harvested and utilised as by the institutions internally, as well as by data marketplaces and repositories on the (inter)national level. We envisage that filling in and updating data-envelopes can become part of the standard research procedures, as they complement already established practice of creating data management plans.

5.3. Standardisation

To ensure the interoperability and widespread adoption of the data-envelope framework, we recognise the importance of aligning our template with existing standards and best practices in the cultural heritage sector. This includes considering the compatibility of the data-envelope with metadata standards such as Dublin Core (Weibel et al., 1998) and CIDOC-CRM (Doerr, 2005), as well as ensuring compliance with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) for research data management (Wilkinson et al., 2016).

By actively engaging with the cultural heritage community and relevant standardisation bodies, we aim to develop a data-envelope template that aligns with existing standards while still addressing the unique challenges of cultural heritage datasets. This standardisation effort will not only facilitate the integration of the data-envelope into existing data management workflows but also promote its adoption across various cultural heritage institutions and projects.

5.4. User-friendliness and Collaborative Documentation

As we continue to engage with the research community and refine the data-envelope template, our primary goal is to achieve a balance between comprehensive documentation and practical implementation. To ensure the template's accessibility and ease of use, we will present the data-envelopes in the form of user-friendly, fillable forms accompanied by clear explanations for each section and field. These explanations will include illustrative examples and outline the purpose of each section, empowering dataset creators to provide accurate and relevant information.

Recognising the collaborative nature of dataset creation and documentation within the cultural heritage domain, we have designed the data-envelope template to facilitate teamwork and collective input. The template will allow multiple team members to contribute to the forms simultaneously, with features such as real-time collaboration, version control, and the ability to save progress as they work through the various sections. This collaborative approach not only streamlines the documentation process but also ensures that the final data-envelope benefits from the diverse expertise and perspectives of the entire research team.

To maximize the benefits of the data-envelope framework, we strongly advise implementing this documentation process at the outset of a research project. By conducting a thorough structural analysis of the dataset during the planning phase, researchers can effectively define their work plans, allocate resources, and identify potential data ethics issues early on. This proactive approach not only saves time and effort in the long run but also promotes a culture of responsible data stewardship from the very beginning of the research lifecycle.

6. Acknowledgements

This work resulted from a joined initiative in the context of two projects carried out at the Huygens Institute: Werk aan Uitvoering (WaU) and GLOBALISE, which is financed by the Dutch Research Council (NWO) in the research programme Research Infrastructure (project number 175.2019.003). We thank the GLOBALISE team, particularly the historical contextualisation team, for their valuable discussions during the initial development of data-envelopes. We also benefitted greatly from Leon's insights on machine readability, Kay's work on the previous datasheet version, and the inspiration drawn from existing datasheets. The Data Management department at Humanities Cluster KNAW, especially Lotte, Douwe, Harm, and Lodewijk, provided crucial feedback in refining the data-envelope.

7. Bibliographical References

Henk Alkemade, Steven Claeyssens, Giovanni Colavizza, Nuno Freire, Jörg Lehmann, Clemens Neudeker, Giulia Osti, Daniel van Strien, et al. 2023. Datasheets for digital cultural heritage datasets. *Journal of Open Humanities Data*, 9(17):1–11.

Paul Longley Arthur, Jason Ensor, Marijke van Faassen, Rik Hoekstra, and Nonja Peters. 2018. Migrating people, migrating data: Digital approaches to migrant heritage. *Journal of the Japanese Association for Digital Humanities*, 3(1):98–113.

Emily M Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Fiona Cameron and Sarah Kenderdine. 2007. *Theorizing digital cultural heritage: A critical discourse*.

Gustavo Candela, Nele Gabriëls, Sally Chambers, Milena Dobreva, Sarah Ames, Meghan Ferriter, Neil Fitzgerald, Victor Harbo, Katrine Hofmann, Olga Holownia, et al. 2023. A checklist to publish collections as data in GLAM institutions. *Global Knowledge, Memory and Communication*.

Nicole Contaxis, Jason Clark, Anthony Dellureficio, Sara Gonzales, Sara Mannheimer, Peter R. Oxley, Melissa A. Ratajeski, Alisa Surkis, Amy M. Yarnell, Michelle Yee, and Kristi Holmes. 2022. Ten simple rules for improving research data discovery. *PLOS Computational Biology*, 18(2):1–11.

Jo Daan and Pieter Jacobus Meertens. 1963. Toelichting bij de taalatlas van noord- en zuinederland. Technical report, Bijdragen en Medelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappente Amsterdam.

Anusuriya Devaraju, Mustapha Mokrane, Linas Cepinskas, Robert Huber, Patricia Herterich, Jerry de Vries, Vesa Akerman, Hervé L'Hours, Joy Davidson, and Michael Diepenbroek. 2021. From conceptualization to implementation: Fair assessment of research data objects. *Data Science Journal*, 20(1):1–14.

Martin Doerr. 2005. The cidoc crm, an ontological approach to schema heterogeneity. Schloss-Dagstuhl-Leibniz Zentrum für Informatik.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *arXiv preprint arXiv:1908.07898*.

Donna Haraway. 2016. 'situated knowledges: The science question in feminism and the privilege of partial perspective'. In *Space, gender, knowledge: Feminist readings*, pages 53–72. Routledge.

Sandra Harding. 2003. How standpoint methodology informs philosophy of social science. *The Blackwell guide to the philosophy of the social sciences*, pages 291–310.

Sandra Harding. 2013. Rethinking standpoint epistemology: What is "strong objectivity"? In *Feminist epistemologies*, pages 49–82. Routledge.

Natalie Harrower, Maciej Maryl, Timea Biro, Beat Immenhauser, and ALLEA Working Group E-Humanities. 2020. Sustainable and fair data sharing in the humanities: : Recommendations of the allea working group e-humanities. Technical report, ALLEA.

Amy K Heger, Liz B Marquis, Mihaela Vorvoreanu, Hanna Wallach, and Jennifer Wortman Vaughan. 2022. Understanding machine learning practitioners' data documentation perceptions, needs, challenges, and desiderata. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–29.

Mar Hicks. 2017. *Programmed inequality: How Britain discarded women technologists and lost its edge in computing*. MIT press.

Rik Hoekstra and Marijn Koolen. 2019. Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 52(2):79–94.

Howard Hotson, editor. 2019. *Reassembling the Republic of Letters in the Digital Age*. Universitätsverlag Göttingen, Göttingen.

Eun Seo Jo and Timnit Gebru. 2020. Lessons from archives: Strategies for collecting sociocultural data in machine learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 306–316.

Kees Kooijmans and Johannes Petrus de Valk. 1985. “Eene dienende onderneming". De Rijkscommissie voor Vaderlandse geschiedenis en haar Bureau 1902-1968. pages 203–271.

Kees Kooijmans and C.E. Keij J.G. Smit Th.S. Bos, A.E. Kersten, editors. 1985. *Bron en publikatie. Voordrachten en opstellen over de ontsluiting van geschiedkundige bronnen*. Bureau der Rijkscommissie voor Vaderlandse Geschiedenis, Den Haag.

Marijn Koolen, Rik Hoekstra, Joris Oddens, Ronald Sluijter, Rutger Van Koert, Gijsjan Brouwer, and Hennie Brugman. 2023. The value of preexisting structures for digital access: Modelling the resolutions of the dutch states general. *J. Comput. Cult. Herit.*, 16(1).

Jörg Lehmann. 2024. Orientation in turbulent times.

Library of Congress. 2021. labs-ai-framework/Experiment/Data-Processing-Plan-template-2021-12-01-draft.docx at main · LibraryOfCongress/labs-ai-framework.

Geertje Mak, Marit Monteiro, and Elisabeth Wesseling. 2020. Child separation: (post-)colonial policies and practices in the netherlands and belgium. *Bijdragen en Mededelingen Betreffende de Geschiedenis der Nederlanden*, 135(3-4):4–28. Introduction to special issue.

Maciej Maryl, Marta Błaszczyńska, Ilaria Bonincontro, Beat Immenhauser, Szilvia Maróthy, Eveline Wandl-Vogt, and Joris J. van Zundert. 2023. Recognising digital scholarly outputs in the humanities. Technical report, ALLEA.

Barbara McGillivray, Beatrice Alex, Sarah Ames, Guyda Armstrong, David Beavan, Arianna Ciula, Giovanni Colavizza, James Cummings, David De Roure, Adam Farquhar, Simon Hengchen, Anouk Lang, James Loxley, Eirini Goudarouli, Federico Nanni, Andrea Nini, Julianne Nyhan, Nicola Osborne, Thierry Poibeau, Mia Ridge, Sonia Ranade, James Smithies, Melissa Terras, Andreas Vlachidis, and Pip Willcox. 2020. The challenges and prospects of the intersection of humanities and data science: A White Paper from The Alan Turing Institute.

Walter D Mignolo and Catherine E Walsh. 2018. *On decoloniality: Concepts, analytics, praxis*. Duke University Press.

Safiya Umoja Noble. 2018. Algorithms of oppression: How search engines reinforce racism. In *Algorithms of oppression*. New York university press.

Kay Pepping, Henrike Vellinga, Manjusha Kuruppath, Leon Van Wissen, and Matthias Van Rossum. 2023. GLOBALISE Thesaurus - Commodities.

Lodewijk Petram and Matthias van Rossum. 2022. Transforming historical research practices–a digital infrastructure for the voc archives (globalise). *International journal of maritime history*, 34(3):494–502.

Timothy B Powell. 2016. Digital knowledge sharing: forging partnerships between scholars, archives, and indigenous communities. *Museum Anthropology Review*, 10(2):66–90.

Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data cards: Purposeful and transparent dataset documentation for responsible ai. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1776–1826.

Anthony Cintron Roman, Jennifer Wortman Vaughan, Valerie See, Steph Ballard, Nicolas Schifano, Jehu Torres, Caleb Robinson, and Juan M. Lavista Ferres. 2023. Open datasheets: Machine-readable documentation for open datasets and responsible ai assessments.

Navroop K Singh, Shuai Wang, Angelica Maineri, and Tycho Hofstra. in press. Aligning data management plans with community standards using fair implementation profiles.

Ramesh Srinivasan, Katherine M Becvar, Robin Boast, and Jim Enote. 2010. Diverse knowledges and contact zones within the digital museum. *Science, technology, & human values*, 35(5):735–768.

Toma Tasovac, Sally Chambers, and Erzsébet Tóth-Czifra. 2020. Cultural Heritage Data from a Humanities Research Perspective: A DARIAH Position Paper. DARIAH’s response to European Commission’s evaluation and possible revision of the Commission Recommendation of 27 October 2011 on Digitisation and Online Accessibility of Cultural Material and Digital Preservation (REC 2011/711/EU).

Nanna Bonde Thylstrup. 2019. *The politics of mass digitization*. MIT Press.

Jo Tollebeek. 1994. *De ijkmeesters : opstellen over de geschiedschrijving in Nederland en België*. Bakker, Amsterdam.

Molly Torsen and Jane Anderson. 2010. Intellectual property and the safeguarding of traditional cultures.

Michel-Rolph Trouillot. 2015. *Silencing the past: Power and the production of history*. Beacon Press.

Gary Urton and Primitivo Nina Llanos. 1997. *The social life of numbers: A Quechua ontology of numbers and philosophy of arithmetic*. University of Texas Press.

Stuart Weibel, John Kunze, Carl Lagoze, and Misha Wolf. 1998. Dublin core metadata for resource discovery. Technical report.

Mark D Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E Bourne, et al. 2016. The fair guiding principles for scientific data management and stewardship. *Scientific data*, 3(1):1–9.

Stacy Wood, Kathy Carbone, Marika Cifor, Anne Gilliland, and Ricardo Punzalan. 2014. Mobilizing records: re-framing archival description to support human rights. *Archival Science*, 14:397–419.

World Wide Web Consortium. 2014. Data Catalog Vocabulary (DCAT).

8. Appendix A. Data-Envelope Structure

Currently, we envisage that the information in Levels 1 and 2 should constitute the default minimum requirements, while still allowing for flexibility in certain fields, such as the role of the project creators for legacy data. Level 3 should be completed to the best of the dataset creators' knowledge, providing essential context and provenance information. Levels 4 and 5 demand the most introspection from the dataset creators, challenging them to step back from the dataset and consider a variety of potential uses and issues. Consequently, these levels would have the least mandatory nature, but we strongly emphasise the importance of maximising their completeness to ensure responsible and ethical use of the datasets.

- **Level 1: Basic Information on Data-Envelope**
  - Title
  - Contact details for each relevant contact person (Name, ORCID, Role in Project, Email)
  - Data-envelope Creation Dates
  - Data-envelope Publication Date
- **Level 2: Basic Dataset Metadata**
  - Snapshot
    - ⋆ Dataset Title
    - ⋆ Version of dataset
    - ⋆ Dataset URL
    - ⋆ Description
    - ⋆ Genre
    - ⋆ Topic Classification
    - ⋆ Geographical Coverage
    - ⋆ Temporal Coverage
  - Dates
    - ⋆ Dataset Creation Dates
    - ⋆ Dataset Publication Date
  - Contributors
    - ⋆ Publishing Organisation (Name, ROR ID, Organization Type)
    - ⋆ Contributor (Name, ORCID, Organization Name, ROR ID, Role)
    - ⋆ Funding Sources for each funding source: Institution(s) (Name of Institution, ROR ID, Funding or Grant Summary(ies), Relevant Links)
  - Distribution
    - ⋆ Dataset Link: own dedicated website or if hosted on sites such as Zenodo, Dataverse
    - ⋆ DOI
    - ⋆ Repository
    - ⋆ Download (URL, File Type(s) and Size)
    - ⋆ Citation Information
  - Access/Licenses
    - ⋆ Licensing Information for every license (Identifier, URL)
    - ⋆ Access Level (Description, URL, Contact Information)
    - ⋆ If Access level: restricted, then (Purpose of access controls, Highlight any restrictions or limitations, Access Prerequisites)
  - Dataset Version and Maintenance
    - ⋆ Version Details (Current Version, Last Updated, Release Date)
    - ⋆ Maintenance Status (Regularly Updated, Actively Maintained, Limited Maintenance, Deprecated)
    - ⋆ Maintenance Plan (Versioning, Updates, Errors, Feedback)
    - ⋆ Next Planned Update(s), if known (Version Affected, Next data update, Next Version)

- **Level 3: Data (Content and Context)**
  - Data Resource Description
    - Name of Resource
    - Description
    - Path, Format, Size, Date
    - Data Subject(s) (Sensitive data about people, Non-sensitive data about people, Data about natural phenomena, Data about places and objects, Synthetically generated data, Data about systems or products and their behaviour)
    - Language(s)
    - Encoding
    - Data Modality (Image Data, Text Data, Tabular Data, Audio Data, Video Data, Time Series, Graph Data, Geospatial Data, Multimodal)
    - Descriptive Statistics (Size of Dataset, Number of Fields, Labelled Classes, Number of Labels, Average labels per instance, Algorithmic labels, Human Labels)
  - Data Fields and Attributes
    - Data Fields Summary
    - Use of Linked Open Data, Controlled Vocabulary, Multilingual Ontologies/Taxonomies
    - Description of every data field in the resource (Data Field Name, Data Field Type, Description of the Field, Sensitivity, Notable Feature(s), Attributes)
    - Data Point Example
    - Atypical Datapoint
    - Any errors, sources of noise, or redundancies in this resource
    - Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, other datasets)
  - Annotation & Labeling
    - Annotation Workforce Type (machine vs human (from experts to non-experts, crowdsourcing, etc)
    - Annotation Characteristics (Description, Number of unique annotations, Total number of annotations, Average number of tokens/annotation, Total tokens annotated, Inter Annotator Agreement (or other relevant metric)
  - Social Impact, Sensitivity, and Biases
    - Does the resource contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?
    - Does the resource contain data that might be considered confidential?
    - Known Biases in the resource
    - Sensitive Human Attributes
    - Unintentionally Collected Attributes
    - Any ethical review processes conducted (e.g., by an institutional review board)?
  - Data Provenance for each source used
    - Name
    - Path
    - Description
    - Creators of Source (name, affiliation, organization and contact if available)
    - Year of publication
    - Language
    - Temporal Scope
    - Geographical Scope
    - Notable Features
    - Datasheet/data-envelope
    - Data Selection Criteria:
  - Digitisation Pipeline

- **Level 4: Uses:** purpose of potential use, domain(s), motivating factors and problem space(s)
  - Uses
    - Dataset Use(s): safe for production use or for research use; conditional use-some unsafe applications; only approved use
    - Links to Related Datasets, Publications, and Models
    - Suitable Use Case(s)
    - Unsuitable Use Case(s)
    - Is there a repository that links to any or all papers or systems that use the dataset?
  - Use with other data
    - Safety Level: safe to use with other data, conditionally safe to user with other data, should not be used with other data
    - Known safe dataset(s) or data type(s)
    - Best Practices
    - Known unsafe dataset(s) or Data Type(s)
    - Limitation(s) and Recommendation(s)
  - Use in ML or AI Systems
    - Dataset Use(s): training, testing, validation, development or production use, fine tuning
    - Notable Feature(s)
    - Known Correlation(s)
    - Data splits
  - Sampling
    - Safety Level: safe to sample, conditionally safe to sample, should not be sampled
    - Acceptable Sampling Method(s)
    - Best Practice(s)
    - Risk(s) and Mitigation(s)
- **Level 5: Human Perspective**
  - Annotator Description(s) per each annotation type
    - Task type, e.g. survey, video annotation, text annotation, image annotation
    - Number of unique annotators
    - Expertise of Annotators
    - Description of annotators
    - Compensation
    - Language distribution of annotators
    - Age distribution of annotators
    - Geographic distribution of annotators
    - Gender distribution of annotators
    - Socio-economic distribution of annotators
    - Summary of annotation instructions
    - Summary of gold questions
    - Annotation platforms
  - Creators

Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data

Maryam Al Emadi, Wajdi Zaghouani
Hamad Bin Khalifa University, Qatar
{mmalemadi, wzaghouani}@hbku.edu.qa

Abstract

Freedom of speech on online social media platforms, often comes with the cost of hate speech production. Hate speech can be very harmful to the peace and development of societies as they bring about conflict and encourage crime. To regulate the hate speech content, moderators and annotators are employed. In our research, we look at the effects of prolonged exposure to hate speech on the mental and physical health of these annotators, as well as researchers with work revolving around the topic of hate speech. Through the methodology of analyzing literature, we found that prolonged exposure to hate speech does mentally and physically impact annotators and researchers in this field. We also propose solutions to reduce these negative impacts such as providing mental health services, fair labor practices, psychological assessments and interventions, as well as developing AI to assist in the process of hate speech detection.

Keywords : Hate Speech, Violent content, Moderators, Annotators, Data Annotation, Mental Health consequences

1. Introduction

“Warning! Today’s presentation contains harmful and toxic materials that are offensive.” This is what appeared on the screen of a researcher when he wanted to present his research about hate speech detection on social media platforms. Hate speech is any verbal attack against a certain group of people with a specific characteristic such as gender, race, ethnic group, religion, or political preference (Dreißigacker et al., 2024).

With the rise in the use of social media platforms, the generation and creation of abusive and hateful content such as texts, pictures, videos, or memes is evident as content creation on these platforms has gained great freedom (Roberts, 2016). These contents are prolonged as they stay forever on the platforms (Oksanen et al., 2021) which increases their consequences on individuals and causes societal implications. Consequently, researchers aim to mitigate the amount of toxicity in these mediums and create a safe place to share different beliefs and ideas. As a result of some flexible policies and the failure of machine learning to mitigate them, harmful content is being posted on a day-to-day basis by users. To reduce the amount of harmful content, tech companies need to rely on human decisions rather than completely depending on machine learning. In response to the spread of hate speech, tech companies have relied on employing content moderators from low-wage countries (Gillespie, 2018) to continually screen the user-generated content (UGC) posted on social media and decide whether they comply with the platforms’ policies and rules or not (Roberts, 2016).

Content moderators play an important role in maintaining digital civility. (Gilliespie, 2018) They get exposed to hate and violent content for long hours daily to identify the harmful content and decide whether specific content complies with the platform’s policy or not, or whether it is acceptable or not (Roberts, 2016). However, the process of moderation itself which exposes the moderators who work as the “gatekeepers of digital civility” to a barrage of disturbing material leads to psychological and emotional consequences. (Newton, 2019) Just like individuals who are victims of hate speech on social media, or even more, moderators face psychological and emotional consequences that can both affect their mental and physical well-being.

Content moderators confront a huge number of challenges such as the prolonged exposure to hate speech, violent content, and other forms of harmful content. This exposure can take a toll on their mental health and physical well-being. Psychological damage, post-traumatic stress disorder (PTSD), anxiety, depression, and insomnia are all effects of long-term exposure to harmful and abusive UGC (Das et al., 2020).

Acknowledging these consequences, researchers like (Das, Dang, & Lease, 2020) suggested a way of getting accurate decisions from moderators in mitigating harmful content in social media and complying EU Service Digital Act and considering the risks on moderators. They emphasize the importance of blurring the contents to reduce the negative effects on moderators’ welfare. However, this approach cannot ensure accurate decisions all the time. Therefore, other solutions tended to test interventions such as gray scaling to achieve the same goal of minimizing emotional impact (D'cruz, Noronha, 2020), which was effective to a certain limit only. Other solutions were imposed such as limiting the time of the exposure, frequent breaks, rotation of duties, on-cite psychological support, and interdisciplinary collaboration between government, tech companies, and mental health experts to set rules that mitigate harm.

Although many research studies were established to discuss the implications of abusive content on individuals, society, and moderators, less is known about the experiences of annotators and the implications of hate speech content on them. Researchers in the field of hate speech detection studies mostly think about ways to detect hate speech, and how to get accurate, transparent, and unbiased results to build ethical datasets to mitigate hate speech on social media. They also discuss the issue of the way annotators perceive toxicity on social media and how their different characteristics influence their decisions (Sap et al., 2022; Waseem, 2016). However, they do not see or acknowledge the emotional consequences on the annotators or themselves as researchers in the field of hate speech detection.

Annotators are a group of individuals who spend their days labeling and tagging hateful content such as videos, photos, memes, and texts for research purposes and for “good” (Kudan, 2022). Unlike moderators, the annotators’ job is more difficult as they must read and see each content carefully to be able to label them with different labels, to train machine learning algorithms to detect hate speech and other abusive content (Kudan, 2022). Therefore, the nature of their work leads to more harm to their mental and physical health. The task of annotation, particularly when it involves labeling harmful and offensive content, carries with it a profound psychological toll that merits closer examination.

Consequently, annotators are mostly expected to suffer from vicarious trauma. This phenomenon occurs when individuals are indirectly exposed to traumatic material through their work, leading to symptoms that mirror those experienced by direct trauma survivors (Pearlman & Saakvitne, 1995). For data annotators, the daily confrontation with content depicting graphic violence, hate speech, sexual abuse, and other forms of human cruelty can lead to a host of distressing symptoms, including intrusive thoughts, hyperarousal, and avoidance behaviors, which are hallmark indicators of Post-Traumatic Stress Disorder (PTSD) (Craig & Sprang, 2010).

In the development and annotation of datasets aimed at detecting hate speech, the authors of this paper have faced the significant challenge of being exposed to hate speech content. This exposure was an essential yet challenging part of their work while curating datasets for various studies. For instance, in their work on the "UPV at the Arabic Hate Speech 2022 Shared Task" (De Paula et al., 2022), they analyzed offensive language and hate speech using transformers and ensemble models. Their subsequent research on hate speech detection in Arabic languages further underscores the complexity of this issue (Magnossão de Paula et al., 2023). Additionally, the creation of a multi-label hate speech annotated Arabic dataset highlighted the nuanced aspects of hate speech across different contexts (Zaghouani et al., 2024). Their collaborative efforts extended to developing the MARASTA corpus, focusing on multi-dialectal Arabic cross-domain stance (Charfi et al., 2024), and analyzing Facebook comments to gather insights on stance, sentiment, and emotion in response to Tunisia's July 25 measures (Laabar & Zaghouani, 2024).

Moreover, constant exposure to these types of content compounds the risk, creating an environment where adequate psychological protection seems virtually impossible (D'cruz & Noronha, 2020). The resulting emotional numbing, a defense mechanism against overwhelming distress, further complicates the annotators' ability to disengage from the trauma of their work, impacting their personal lives and relationships.

Furthermore, the stigma associated with discussing mental health issues, particularly in professional contexts, can deter annotators from seeking the help they need. This silence perpetuates a cycle of suffering, with many feelings isolated in their experiences and uncertain of where to turn for support (Rosenbaum et al., 2018).

Highlighting these challenges faced by annotators in addressing hate speech on social media platforms, the need for greater awareness for those employees who are constantly engaged in such work is needed. As well as implementing effective strategies to minimize the psychological and physiological harm impacts imposed on annotators.

Despite the serious consequences and implications of data annotation on the annotators’ well-being, research addressing this issue is not found. This calls us to establish our comprehensive study titled “Emotional Toll and Coping Strategies: Navigating the Effects of Annotating Hate Speech Data”.

In this study, we will not only focus on the impact of hate speech on annotators, but rather on researchers as well whose lives are evolving around hate speech related topics as they constantly read about them.

2. Methodology

Our research begins by thoroughly and comprehensively exploring the essential responsibilities shouldered by annotators in the process of hate speech annotation. This investigation necessitates an in-depth review of existing literature pertaining to the roles and obligations of annotators engaged in tasks related to hate speech detection. Through an exhaustive examination of available scholarly works, our objective is to grasp the guidelines and protocols that are instituted for annotators prior to the commencement of the annotation process.

More specifically, our focus lies in elucidating the guidelines imparted to annotators to ensure the accurate labeling and categorization of hate speech content. This endeavor involves a thorough analysis of the training methodologies employed to instill adherence to these guidelines among annotators, as well as the mechanisms for assigning suitable labels to the annotated data. Additionally, we explore the impediments faced by annotators in adhering to these guidelines, particularly within the confines of the time constraints imposed for completing annotation tasks.

By amalgamating insights gleaned from the literature, we discern pivotal tasks undertaken by annotators, the nature of their training, and the repercussions of time constraints on the quality of annotated data. This methodological approach facilitates a comprehensive understanding of the factors that influence annotators' performance and the emotional toll incurred in the process of hate speech annotation.

Annotators often operate within demanding and stressful environments, compelled to annotate vast quantities of hate speech content within specified time frames. Each piece of content necessitates meticulous review and labeling, exerting significant cognitive effort. Consequently, we investigate the physical and digital work environments of annotators, including their work schedules, breaks, technological tools utilized, and the support systems provided by their employers to mitigate associated challenges.

Furthermore, our study evaluates the impact of imposed deadlines and productivity targets on annotators' well-being, seeking to strike a balance between the quality and quantity of their work. Moreover, we explore the effects of exposure to harmful content on the mental health of both researchers and annotators involved in such endeavors. Through the administration of a survey targeting both cohorts, we aim to quantify the impact of prolonged exposure to harmful content on their mental well-being and elucidate the diverse ramifications that impede their daily activities.

The paper also raises the ethical questions about the psychological toll of data annotation and the regulatory complexities that are continuously evolving. It discusses the moral imperative to protect the well-being of these workers, necessitating the implementation of comprehensive mental health support systems, regular psychological assessments, and accessible interventions designed to address the unique challenges faced by annotators (Roberts, 2016). It aims to explore the partnership with mental health organizations to provide support for annotators with issues related to their constant exposure to such harmful content.

Furthermore, fostering a workplace culture that prioritizes mental health, encourages open discussions about emotional well-being, and actively destigmatizes mental health issues is crucial. Such measures not only support annotators in managing the psychological impacts of their work but also contribute to a more compassionate and ethical approach to data annotation (Armstrong et al., 2018).

In this paper, we provide policies that can actively destigmatize mental health issues among employees as well as addressing the ethical considerations surrounding the work of annotators which aims to protect their well-being. To achieve that, the study engages in a broader discussion about the responsibility of tech companies, policymaker, and the global community in recognizing the psychological and emotional consequences of such work on annotators and how to support their mental health.

3. Discussion

In suggesting a solution to mitigate harm to annotators, the study recommends that tech companies provide fair labor practices and on-site mental health support, transparency, and accountability about the nature of data annotation, and developing ethical AI technologies to assist annotators in data annotation. Tech companies have an inherent ethical responsibility to ensure that the working conditions of data annotators meet high standards of fairness and respect for human dignity. Given the psychologically taxing nature of annotating harmful and offensive content, companies must go beyond traditional labor practices to implement comprehensive mental health support systems. These systems should include access to psychological counseling, mental health days, and programs designed to mitigate the impact of vicarious trauma (Armstrong et al., 2018; Craig & Sprang, 2010).

Moreover, the ethical obligation extends to providing a work environment that fosters open communication about mental health challenges without fear of stigma or reprisal. Implementing regular mental health assessments and training for managers and supervisors on recognizing and addressing signs of psychological distress among their teams can create a supportive atmosphere conducive to employee well-being (D'cruz & Noronha, 2020).

Transparency about the nature of data annotation work and the potential psychological risks associated with it is another critical ethical responsibility. Tech companies must ensure that annotators are fully informed about the content they will encounter and understand the available support mechanisms. This transparency should also extend to the public and regulatory bodies, with companies openly disclosing their practices and the measures they take to safeguard annotator well-being (Roberts, 2016).

Accountability mechanisms, such as independent audits of working conditions and mental health support provisions, can further ensure that companies adhere to their ethical obligations. These measures not only protect annotators but also build trust among stakeholders, including users, regulators, and the broader public (Gillespie, 2018).

The development of AI systems for content moderation raises profound ethical questions about the reliance on human-annotated data. Tech companies must grapple with the dual imperatives of advancing technological innovation and ensuring that this progress does not come at the expense of human well-being. Ethical AI development practices require a commitment to minimizing the reliance on human annotation of harmful content wherever possible, exploring alternative methods that reduce exposure to such content, and investing in research aimed at improving AI's ability to understand context and nuance without extensive human input (Gorwa, Binns, & Katzenbach, 2020).

The ethical obligations of tech companies in the realm of data annotation are multifaceted and complex. As the digital world continues to evolve, the need for responsible, ethical practices in the development and maintenance of AI systems becomes increasingly paramount. By prioritizing the health and well-being of data annotators, fostering transparency and accountability, and pursuing ethical AI development, tech companies can navigate the challenges of data annotation while upholding their moral responsibilities to their employees and society at large.

The paper also touches on legal and regulatory considerations for tech companies in data annotation. It suggests evolving legal frameworks necessitate a proactive approach to compliance. This involves not only implementing robust content moderation systems but also ensuring that the processes of data annotation — a critical component in the development of these systems — align with legal standards regarding worker rights and data privacy.

Furthermore, due to the regular harm that is imposed on annotators, the study believes that legal considerations extend beyond compliance to encompass the ethical implications of data annotation work, particularly regarding the protection of annotators from harm.

In addition to adhering to legal requirements, there is a growing call for tech companies to engage in self-regulation and the development of industry standards for data annotation. This involves creating transparent, accountable practices that ensure the ethical treatment of annotators and the responsible development of content moderation technologies. Industry standards could include guidelines for annotator well-being, data privacy, and the accuracy and fairness of annotated datasets used to train AI systems (Roberts, 2016).

4. Conclusion

The task of data annotation is a difficult task that has prolonged consequences, which emerges as a poignant emblem of the hidden costs associated with building safer online environments. It underscores a significant yet unappreciated human cost in the effort to create a safer online environment with less harm to its users. Recognizing the long-lasting effects on annotators mental health is paramount to developing sustainable and ethical practices in the field of data annotation. The exploration of this critical yet often overlooked aspect of digital infrastructure reveals profound ethical, psychological, and regulatory challenges that demand our immediate attention and action.

The psychological toll on data annotators, highlighted through the lens of vicarious trauma and the elevated risks of mental health issues such as PTSD, anxiety, and depression, underscores a pressing moral imperative (Craig & Sprang, 2010). These individuals, who serve as the first line of defense against the proliferation of harmful content, endure significant emotional and psychological strain, necessitating a robust framework of support (D'cruz & Noronha, 2020). The ethical obligations of tech companies in this context extend beyond mere compliance with legal standards to encompass a duty of care that honors the humanity and dignity of each annotator (Roberts, 2016). Which includes providing comprehensive mental health support, fostering a workplace culture that prioritizes their well-being and implementing fair labor practices.

Furthermore, the evolving legal and regulatory landscape presents both challenges and opportunities for safeguarding the well-being of data annotators. Legislation such as the Digital Services Act in Europe represents a critical step towards holding tech companies accountable for the content on their platforms and, by extension, for the conditions under which data annotators work. However, these regulations must be carefully crafted to ensure they do not inadvertently exacerbate the pressures on annotators, instead fostering an environment that prioritizes their mental health and well-being (Keller, 2020). Therefore, an interdisciplinary collaboration needs to be established between tech companies, policymakers, and mental health experts to come up with regulations that can effectively protect the public users and the annotators.

Looking ahead, the future of data annotation and content moderation lies in the delicate balance between leveraging technological advancements and preserving the essential human element. The potential of artificial intelligence and machine learning offer promising avenues for reducing the burden on human annotators by automating aspects of content moderation. However, these technologies are not a panacea. The nuances of human communication and the contextual understanding necessary for evaluating content underscore the irreplaceable value of human judgment (Gorwa, Binns, & Katzenbach, 2020). Therefore, those innovations should aim to support not replace the critical work of annotators, ensuring that technologies enhance rather than diminish human well-being.

In conclusion, the discourse surrounding the data annotation of harmful and offensive content invites us to reflect on the broader implications of our digital age. It compels us to consider not only the technological and economic dimensions but also the human cost of creating and maintaining digital spaces. As we navigate this complex terrain, we must forge a path that respects the contributions of data annotators, addresses the ethical challenges inherent in their work, and envisions a future where technology serves to enhance human well-being. The integrity and safety of our digital spaces depend on our collective ability to recognize, support, and protect those who labor in the shadows to keep them clean.

5. Limitations and Future Work

Our study was significantly constrained by both temporal limitations and ethical considerations. The sensitive nature of our research topic, which involved examining the potential impacts of sensitive data annotation on the mental and physical well-being of annotators, necessitated a careful approach to participant engagement. However, the lack of Institutional Review Board (IRB) approval emerged as a major impediment, limiting our ability to conduct in-depth interviews and, consequently, restricting the scope of our investigation. The unavailability of ethical clearance precluded the collection of direct testimonies from annotators, thus curtailing our understanding of the effects of their work.

The primary reason for the absence of IRB clearance was the stringent time constraints under which our study operated. The time-sensitive nature of the research process did not allow for the completion of the extensive and rigorous IRB approval procedures, thereby hindering our capacity to engage directly with annotators through interviews or surveys.

This limitation not only highlights the ethical complexities associated with research on mental health topics but also stresses the necessity for future studies to meticulously navigate the ethical review process. The experience underscores the critical importance of obtaining IRB approval to ensure a comprehensive exploration of the research subject.

To address the aforementioned constraints and augment the methodological rigor of our subsequent inquiries, securing Institutional Review Board (IRB) clearance will be our foremost priority. Achieving this will facilitate the execution of in-depth, qualitative interviews with a carefully selected cohort of data annotators. This strategic approach is intended to yield a more nuanced understanding of the intricacies and repercussions inherent in data annotation processes. We posit that conducting individualized, qualitative interviews will be essential for eliciting profound insights into the annotators' personal experiences, their strategies for managing work-related stress, and their perceptions of support from their employers. Such an investigative framework will enable the collection of detailed personal narratives, thereby illuminating the experiences of those involved in sensitive data annotation and the subsequent effects on their mental and physical well-being. This methodological enhancement is expected to significantly contribute to the body of knowledge concerning the occupational health aspects of data annotation work.

6. Acknowledgments

This publication was made possible through the financial support of NPRP grant NPRP13S-0206-200281 from the Qatar National Research Fund and the Qatar Research Development and Innovation Council (QRDI). The findings and opinions expressed herein are solely the responsibility of the authors.

7. Bibliographical References

Armstrong, C., Davis, J., Holder, K., Knowles, K., & Patel, K. (2018). The Trauma Floor: The Secret Lives of Facebook Moderators in America. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Armstrong, G., Blashki, G., Jorm, A. F., Kitchener, B. A., & Crisp, D. A. (2018). An analysis of stigma and suicide literacy in responses to suicides broadcast on social media. Asian Journal of Psychiatry, 37, 25-31. https://doi.org/10.1016/j.ajp.2018.07.017

Charfi, A., Ben-Sghaier, M., Atalla, A., Akasheh, R., Al-Emadi, S., & Zaghouani, W. (2024). MARASTA: A multi-dialectal Arabic cross-domain stance corpus. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

De Paula, A. F. M., Rosso, P., Bensalem, I., & Zaghouani, W. (2022). UPV at the Arabic hate speech 2022 shared task: Offensive language and hate speech detection using transformers and ensemble models. In Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection (pp. 181-185).

Craig, C. D., & Sprang, G. (2010). Compassion satisfaction, compassion fatigue, and burnout in a national sample of trauma treatment therapists. Anxiety, Stress & Coping, 23(3), 319–339. https://doi.org/10.1080/10615800903085818

Das, A., Dang, V., & Lease, M. (2020). Fast, Accurate, and Healthier: Interactive Blurring Helps Moderators Reduce Exposure to Harmful Content. In Proceedings of the 7th AAAI Conference on Human Computation and Crowdsourcing (HCOMP 2019). https://www.ischool.utexas.edu/~ml/papers/das_hcomp20.pdf

D'cruz, P., & Noronha, E. (2020). Navigating the extended reaches of online harm: The experience of online content moderators. Work, Employment and Society, 34(3), 456-475. https://doi.org/10.1177/0950017020914877

D'cruz, R., & Noronha, E. (2020). Testing Stylistic Interventions to Reduce Emotional Impact of Content Moderation Workers. Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 8, 128–137. https://ojs.aaai.org/index.php/HCOMP/article/view/5270

Dreißigacker, A., Müller, P., Isenhardt, A., & Schemmel, J. (2024, February 10). Online hate speech victimization: consequences for victims' feelings of insecurity. Crime Science. https://doi.org/10.1186/s40163-024-00204-y

European Commission. (2020). Digital Services Act: Ensuring a safe and accountable online environment. https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package

Ghosh, D., & Guha, R. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. In Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (pp. 112–117). Association for Computational Linguistics. https://aclanthology.org/W16-5618.pdf

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape social media. Yale University Press.

Gillespie, T. (2018). Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press.

Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society, 7(1). https://doi.org/10.1177/2053951720919770

Keller, D. (2020). Internet platforms: Observations on speech, danger, and money. Hoover Institution, Aegis Series Paper No. 1907.

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. Harvard Law Review, 131, 1598-1670.

Koops, B. J. (2014). The trouble with European data protection law. International Data Privacy Law, 4(4), 250-261. https://doi.org/10.1093/idpl/ipu023

Kudan, G. (2022). Unpublished manuscript.

Kudan. (2022, December 8). The role of a data annotator in machine learning. Toloka. Retrieved from https://toloka.ai/blog/what-does-a-data-annotator-do/

Laabar, S., & Zaghouani, W. (2024). Multi-dimensional insights: Annotated dataset of stance, sentiment, and emotion in Facebook comments on Tunisia's July 25 measures. In Proceedings of the Second Workshop on Natural Language Processing for Political Sciences co-located with the 2024 International Conference on Computational Linguistics, Language Resources and Evaluation.

Magnossão de Paula, A. F., Bensalem, I., Rosso, P., & Zaghouani, W. (2023). Transformers and ensemble methods: A solution for hate speech detection in Arabic languages. arXiv preprint arXiv:2303.

Newell, J. M., MacNeil, G. A. (2010). Professional burnout, vicarious trauma, secondary traumatic stress, and compassion fatigue. Best Practices in Mental Health, 6(2), 57-68.

Newton, C. (2019). The toll of the online content moderator. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2019). The Trauma Floor: The secret lives of Facebook moderators in America. The Verge. Retrieved from https://www.theverge.com/2019/2/25/18229714/co gnizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2020). Inside the traumatic life of a Facebook moderator. The Verge. https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona

Newton, C. (2020, January 28). What tech companies should do about their content moderators' PTSD. The Verge. https://www.theverge.com/interface/2020/1/28/21082642/content-moderator-ptsd-facebook-youtube-accenture-solutions

Oksanen, A., Celuch, M., Latikka, R., Oksa, R., & Savela, N. (2021, November 23). Hate and harassment in academia: the rising concern of the online environment. Higher Education. https://doi.org/10.1007/s10734-021-00787-4

Pearlman, L. A., & Saakvitne, K. W. (1995). Trauma and the therapist: Countertransference and vicarious traumatization in psychotherapy with incest survivors. W. W. Norton.

Pearlman, L. A., & Saakvitne, K. W. (1995). Trauma and the Therapist: Countertransference and Vicarious Traumatization in Psychotherapy with Incest Survivors. W.W. Norton & Company.

Roberts, S. T. (2016). Commercial Content Moderation: Digital Laborers' Dirty Work. In The Intersectional Internet: Race, Sex, Class, and Culture Online (pp. 147-159). Peter Lang.

Roberts, S. T. (2016). Content Moderation. Social Media + Society, 2(2), 2056305116644493. https://www.academia.edu/31637827/Content_Moderation

Robertson, A. (2019, June 19). BODIES IN SEATS At Facebook's worst-performing content moderation site in North America, one contractor has died, and others say they fear for their lives. The Verge. https://www.theverge.com/2019/6/19/18681845/facebook-moderator-interviews-video-trauma-ptsd-cognizant-ta

Rosenbaum, L., et al. (2018). Caring too much: Compassion fatigue and mental Smith, J. K. (Year). Title of the article. Journal Name, Volume(Issue), Page range. Retrieved from URL

Rosenbaum, L., Shieber, S., & McNally, R. J. (2018). The National Stressful Events Survey PTSD Short Scale (NSESSS). Psychological Assessment, 30(3), 405–415. https://doi.org/10.1037/pas0000562

Sap, Swayamdipta, Vianna, Zhou, Choi, & Smith, A. (2022). Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. https://aclanthology.org/2022.naacl-main.431.pdf

Waseem. (2016). Are You a Racist or Am I Seeing Things? Annotator Influence on Hate Speech Detection on Twitter. https://aclanthology.org/W16-5618.pdf European Commission. (2020). Code of Practice on Disinformation. https://ec.europa.eu/digital-single-market/en/code-practice-disinformation

Zaghouani, W., Mubarak, H., & Biswas, M. R. (2024). So hateful! Building a multi-label hate speech annotated Arabic dataset. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024).

User Perspective on Anonymity in Voice Assistants – A comparison between Germany and Finland

Ingo Siegert[1], **Silas Rech**[2], **Tom Bäckström**[2], **Matthias Haase**[3]
[1]Mobile Dialog Systems, Otto von Guericke University, Magdeburg, Germany,
[2]Speech Interaction Technology, Aalto University, Finland,
[3]Department of Engineering and Industrial Design,
University of Applied Sciences Magdeburg-Stendal, Germany
siegert@ovgu.de, silas.rech@aalto.fi, tom.backstrom@aalto.fi, matthias.haase@h2.de,

Abstract

This study investigates the growing importance of voice assistants, particularly focusing on their usage patterns and associated user characteristics, trust perceptions and concerns about data security. While previous research has identified correlations between the use of voice assistants and trust in these technologies, as well as data security concerns, little evidence exists regarding the relationship between individual user traits and perceived trust and security concerns. The study design involves surveying various user attributes, including technical proficiency, personality traits, and experience with digital technologies, alongside attitudes toward and usage of voice assistants. A comparison between Germany and Finland is conducted to explore potential cultural differences. The findings aim to inform strategies for enhancing voice assistant acceptance, including the implementation of anonymization methods.

Keywords: Voice Assistants, Trust, Privacy Concerns, Technology Commitment

1. Introduction

Voice assistants have become integral to our daily lives, particularly in commercial contexts, garnering a substantial user base (Kinsella, 2020; Kleinberg, 2018; Osborne, 2016). Their simplicity and natural communication style, without the need for additional peripherals, have contributed to their widespread adoption. However, as they infiltrate sensitive domains, socio-ethical considerations regarding their use are gaining prominence.

Despite their increasing popularity, concerns persist regarding the privacy of user input data, particularly speech data stored and processed on cloud platforms (Krüger and Siegert, 2020; Leschanowsky et al., 2023). This skepticism, stemming from fears of potential misuse for unauthorized purposes, is impeding the widespread adoption of voice assistants in public and healthcare interactions (Wienrich et al., 2021).

Despite the growing availability of speech-based technology, a significant portion of the population either uses these technologies minimally or not at all. The reasons behind this discrepancy remain unclear, leading many studies to simply categorize individuals as either users or non-users (Sinha et al., 2022). Moreover, a 2019 study revealed a rapid increase in the proportion of non-users of voice assistants in Germany over a two-year period, with privacy concerns identified as a major factor for non-usage (Splendid Research GmbH, Januar 2019).

To comprehensively understand the factors influencing both usage and non-usage, it is imperative to examine the perspectives of both users and non-users. While existing research often cites a lack of trust in voice assistants or the companies behind them, a deeper exploration into how users and non-users perceive and rationalize their mistrust, their views on speaker anonymization, and the factors influencing their attitudes and behaviors remains largely unexplored.

Several studies have linked the non-usage of voice assistants to issues of trust and privacy/data security concerns (Olson and Kemery, 2019; Brill et al., 2019; Dhagarra et al., 2020; Vimalkumar et al., 2021). A more recent survey by Bitkom in Germany highlighted data security as the primary concern among participants, with 59% expressing worry over their data, 53% fearing eavesdropping by third parties, and 35% being reluctant to transmit background speech over the internet (Bitkom, 2022). Despite these apprehensions, the survey also revealed a general willingness to utilize voice assistants, with only 22% of participants expressing reluctance to control devices via voice commands.

However, a nuanced understanding of why some individuals harbour data security concerns while others do not remain elusive. Nonetheless, existing research underscores the positive impact of perceived usefulness and competence on trust and attitudes toward voice assistants (Pitardi and Marriott, 2021). Hereby, most studies do not differentiate between mobile and stationary systems. But, distinguishing between stationary and mobile Voice User Interfaces (VUIs) is essential due to the differing contexts and usage patterns associated with each platform. Stationary VUIs, such as smart speakers, are typically used in fixed locations within homes or workplaces, providing hands-free access to in-

LEGAL2024 @LREC-COLING-2024, pages 73–78
20 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0

formation and services. In contrast, mobile VUIs, integrated into smartphones or wearable devices, offer on-the-go access to voice-activated features, enabling seamless interaction while moving. Consequently, the current study intentionally makes this differentiation.

As part of this discussion, a study conducted in early 2023 specifically investigated individual reasons for the use and non-use of voice-assisted technologies in Germany by surveying both users and non-users. The study also examined whether anonymizing user information and speech input helps reduce barriers towards voice assistants (Haase et al., 2023). The current paper aims to explore whether and how there are differences in response behaviour among two selected European countries. For comparison, Finland was chosen as it shares the same legal framework (data protection, AI regulation) as an EU country but has a significantly higher level of digitization than Germany. According to the Digital Economy and Society Index (DESI), which tracks the progress of EU member states in four key areas including human capital, connectivity, integration of digital technology, and digital public services, Finland has a digitization index of 69.6 (1st place), whereas Germany, ranked 13th, has an index of only 52.88 (European Commission, 2022).

The aim of this study is to increase the understanding of individual reasons for the use and non-use of voice-assisted technologies by surveying users and non-users and whether anonymization of user's information and speech input helps to reduce restraints towards voice assistants. Hereby, the use and non-use of VUIs in a stationary and mobile scenario will be analyzed between the participants from Germany and Finland.

2. Methods

The study employs a data collection method similar to that of (Haase et al., 2023) for the initial step. It investigates the relationships between attitudes toward voice assistants (distinguishing mobile and stationary systems), technology commitment (including acceptance, competence, and control beliefs), individual personality traits, and actual usage or non-usage. To achieve this, data is gathered through an online survey questionnaire.

Utilizing quantitative methods, the research design focuses on analyzing correlations between different user variables and the adoption or rejection of voice assistants.

Recruitment: Both surveys aimed to gather a diverse sample in terms of age (spanning from 18 to 81) and gender. Additionally, they collected information on participants' education level, technology usage, and familiarity with modern information and communication technologies.

The recruitment for the German survey was carried out by students from the Human-Technology Interaction and Rehabilitation Psychology programs at the University of Applied Sciences Magdeburg-Stendal, as well as through various mailing lists managed by the researchers. Utilizing the snowball method, the students encouraged others, peers, friends, and family members, to participate in the survey. The first part of data collection took place from January 16 to January 29, 2023. For the Finnish survey, recruitment was conducted by the first and second authors through mailing lists, social media post by the city of Espoo and LinkedIn posts. The second part of data collection started on October 10, 2023, and was terminated on March 10, 2024.

Survey and evaluation methods: The survey was conducted using the SoSci survey platform hosted at Otto von Guericke University, Magdeburg (Leiner, 2019). This platform ensures end-to-end SSL encryption and secure data storage. Servers are located in a certified and secured data center in Germany, adhering to the General Data Protection Regulation (GDPR). The Ethics Committee of the Department of Applied Human Sciences at Magdeburg University of Applied Sciences approved the Germon version of the study. The Finnish counterpart survey was approved by the research ethics committee at Aalto University. Participants provided informed consent, acknowledging the study's objectives, voluntary participation, right to withdraw, and their rights under the GDPR.

Survey content: Both surveys covered sociodemographic variables, Big-Five personality dimensions (BFI-L) (Rammstedt and John, 2005), current technology usage, and perceived hedonic and utilitarian benefits, trust in voice assistants and general privacy concerns were included. Table 1 gives an overview of both survey contents. For most items, the same (English) questionnaires as those used in the German study were employed. However, in the sociodemographic variables, the question regarding educational attainment was adapted to the Finnish system. Since there is no translated and validated version or anything comparable for the construct of technology readiness, we have opted to use the Affinity for Technology Interaction (ATI) Scale for the Finnish questionnaire instead of the Technology Commitment (Neyer et al., 2012). Consequently, the areas of experiencing technology competence and technology control experience are unfortunately omitted. For all other scales, we refer the reader to the paper on the German study (Haase et al., 2023). In total, the

Table 1: Overview of the different instruments of the survey, differences to the German questionnaire apart from language are highlighted. Details can be found in the text. The last row denotes whether the instrument is used for the (G)erman and/or (F)innish questionnaire.

Section	# Items	Content	Reference	
Sociodemographic Variables	5	age, gender, education degree, current employment, place of residence	–	G,F
Big-Five Personality	21	Short version of the Big-Five Inventory (BFI-K)	(Rammstedt and John, 2005)	G,F
Technology Commitment	12	Brief Measure of Technology Commitment Commitment	(Neyer et al., 2012)	G
Affinity for Technology Interaction	9	Just measures the technological affinity	(Franke et al., 2019)	F
Technology Usage	6	computer/smartphone usage per week, use of voice assistants, frequency of use	–	G,F
Hedonic and Utilitarian Benefits	5	Hedonic (enjoyment, entertainment value, fun in accomplishing tasks) Utilitarian (convenience in organizing time, facilitation of tasks)	(McLean and Osei-Frimpong, 2019)	G,F
Trust	3	truthfulness of statements, trustworthiness, trust in developing companies	(Pitardi and Marriott, 2021; Olson and Kemery, 2019)	G,F
Privacy concerns	5	confidentiality doubts, hesitations about conducting transactions via voice assistants, worries about personal data storage, and reluctance to share personal information	(Pitardi and Marriott, 2021; Olson and Kemery, 2019)	G,F

German questionnaire comprises 57 items and the Finnish survey covered 54 items, due to the different questionnaires regarding technology experience, Technology Commitment vs. Affinity for Technology Interaction.

Hypotheses and Analysis: Both questionnaires are analyzed based on the following hypotheses, with each device type tested independently:

Trust & Concerns Regarding Privacy

H1 Individuals with lower trust in VUIs use them to a lesser extent than those with higher trust.

H2 Individuals with stronger privacy concerns use VUIs to a lesser extent.

Hedonic & Utilitarian Benefits

H3 Individuals who perceive lower utilitarian benefits from VUIs also use them to a lesser extent.

H4 Individuals who experience lower hedonic pleasure in using VUIs also use them to a lesser extent.

Relationship between Trust, Provacy & Hedonic, Utilitarian Benefits

H5 There is a relationship between perceived hedonic and utilitarian benefits and trust in VUIs.

H6 There is a relationship between perceived hedonic and utilitarian benefits and concerns regarding privacy with respect to VUIs.

Technology Readiness & Technological Experience

H7 Individuals who report overall lower technology readiness use VUIs to a lesser extent.

Personality

H8 Statistical correlations can be found between the Big Five dimensions of Neuroticism, Conscientiousness, Openness, and VUI usage.

For data analysis, SPSS 29 was utilized. Hypotheses 1 to 4 were tested using a point-biserial correlation, given that usage vs. non-usage represents a dichotomous variable. Hypotheses 5 to 8 were examined using a Spearman correlation, as the variables under consideration are each regarded as interval-scaled.

3. Sample Description

A total of 581 people finished the German survey and 46 people finished the Finish survey.

Not surprising, the average age of the participants is relatively young (German: M=34.5 years, Finland M=35.4 years), see Table 2 for the age-group distribution. In terms of gender distribution, in the German survey the majority is female (female: 55.6%, male: 43.5%9, diverse 0.7%) while in the Finnish survey a gender equality was achieved (female 47.8%, male: 50.0%, diverse: 2.2%). Thus, the participant samples are comparable, except a slight shift between the age-groups of Gen Z (18-25) and Millennials (26-35).

Table 2: Age distribution of survey participants

	German	Finland
	[%]	[%]
18 to 25	46.5	17.4
26 to 35	16.9	41.3
36 to 45	11.9	26.1
46 to 55	11.4	10.9
56 to 65	7.4	4.3
66 to 75	5.0	0.0
Unknown	1.4	0.0

4. Results

Hypothesis 1 was confirmed for the German sample for both stationary ($r_{pb} = .174$, $p < 0.001$) and mobile usage ($r_{pb} = .256$, $p < 0.001$), that individuals with lower trust in VUIs use them to a lesser extent. In the Finnish sample, this holds true only for the use of stationary devices ($r_{pb} = .296$, $p = 0.023$); for mobile VUIs, the analysis is not significant ($r_{pb} = .213$, $p = 0.077$), but only marginally outside significance.

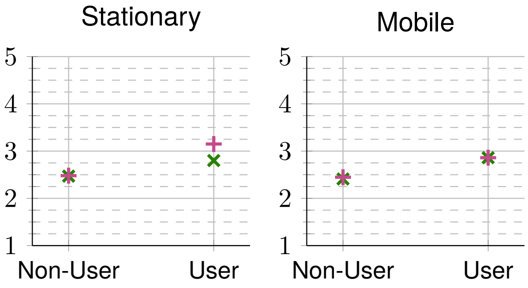

Figure 1: Mean Difference between trust and usage of stationary and mobile VUIs for German (×) and Finnish (+) participants (H1).

Hypothesis 2 was also confirmed for the German participants, both for stationary ($r_{pb} = -.245$, $p < 0.001$) and mobile ($r_{pb} = -.273$, $p < 0.001$) usage. For the Finnish questionnaire, interesting differences emerge between stationary and mobile usage. While there is no correlation between privacy concerns and the use of VUIs for stationary usage ($r_{pb} = .021$, $p = 0.444$), this is highly pronounced for mobile devices ($r_{pb} = -.374$, $p = 0.005$).

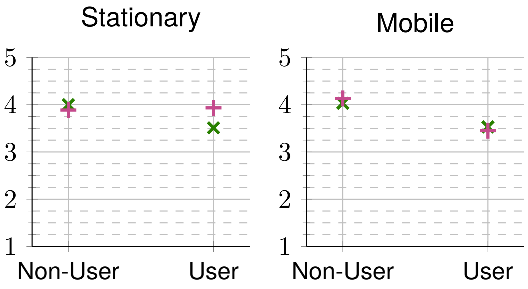

Figure 2: Mean Difference between privacy concerns and usage of stationary and mobile VUIs for German (×) and Finnish (+) participants (H2).

Regarding the perceived hedonic and utilitarian benefits (H3 & H4), quite contrary observations were made between German and Finnish participants. For German participants, individuals who perceive lower benefits or joy in using VUIs also use them to a lesser extent. This applies to both stationary (Hedonic: $r_{pb} = -.332$, $p < 0.001$ Utilitaristic $r_{pb} = -.286$, $p < 0.001$) and mobile (Hedonic: $r_{pb} = -.380$, $p < 0.001$ Utilitaristic $r_{pb} = -.319$, $p < 0.001$) voice assistants. However, this does not hold true for Finnish participants, as no significant differences can be found neither for stationary devices (Hedonic: $r_{pb} = .205$, $p = 0.086$ Utilitaristic $r_{pb} = .119$, $p = 0.215$) nor for mobile devices Hedonic: $r_{pb} = .197$, $p = 0.095$ Utilitaristic $r_{pb} = .238$, $p = 0.056$).

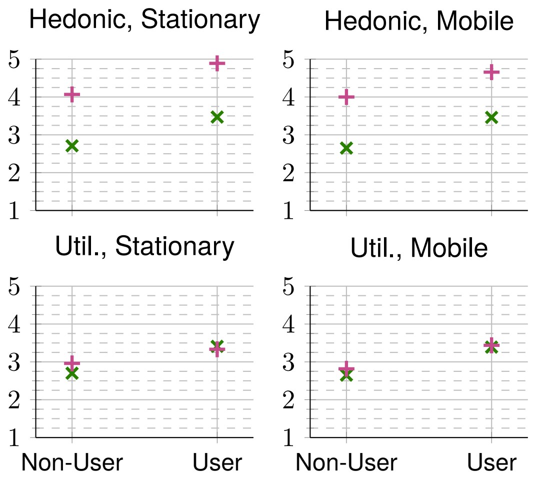

Figure 3: Mean Difference between hedonic and utilitaristic benefits and usage of stationary and mobile VUIs for German (×) and Finnish (+) participants (H3, H4).

For Hypothesis 5 and 6, as both variables are interval-scaled, meaning they are not dichotomous as in usage (1) non-usage (0), a Spearman correlation was conducted in this case. Regarding Hypotheses 5 on the relationship between perceived hedonic and utilitaristic benefits and trust in VUIs, there is a strong statistical effect for the German population no matter whether stationary or mobile devices are used (Hedonic. $r_s = .414$, $p < 0.001$ Utilitaristic: $r_s = .355$, $p < 0.001$). Also regarding the Finnish sample, there is a strong correlation between perceived hedonic and utilitaristic benefits and trust in VUIs (Hedonic: $r_s = .416$, $p = 0.002$ Utilitaristic: $r_s = .326$, $p < 0.013$). Thus, for both populations, we can state that high hedonic and utilitarian quality perception is associated with high trust.

Regarding Hypothesis 6 on the relationship between perceived hedonic and utilitaristic benefits and privacy concerns in VUIs, only for the German population a clear correlation can be found, regardless of a stationary or mobile device type (Hedonic: $r_s = -.314$, $p < 0.001$ Utilitaristic: $r_s = -.267$, $p < 0.001$). For the Finnish population, privacy concerns, just like in the hypotheses above regarding utilitarian quality, do not play a role here (Hedonic: $r_s = -.132$, $p = 0.191$ Utilitaristic: $r_s = -.144$, $p = 0.171$).

In Hypothesis 7, we assume that individuals with lower technology commitment use voice assistants to a lesser extent. Significant differences in the German population and the Finish population were observed regarding technology commitment and technology affinity. While a significant correlation was confirmed for the German questionnaire in terms of all three scales and both stationary and mobile usage (technology acceptance: $r_{s-stat.} = .278$, $r_{s-mob.} = .334$, $p < 0.001$, technology competence belief: $r_{s-stat.} = .115$, $r_{s-mob.} = .217$, $p < 0.001$, and technology control beliefs: $r_{s-stat.} = -.132$, $r_{s-mob.} = .129$, $p < 0.001$), for the questionnaire used in the Finnish study regarding technology affinity, no correlation was found, neither for the stationary ($r_s = -.036$, $p = 0.407$) nor for the mobile usage ($r_s = -.069$, $p = 0.324$).

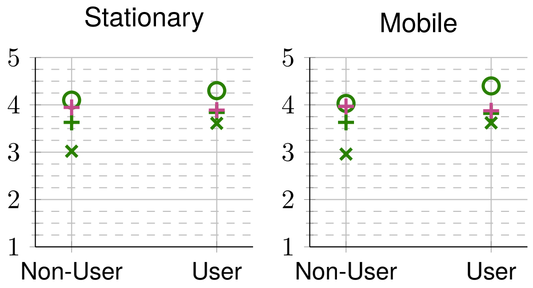

Figure 4: Mean Difference between technology commitment/affinity and usage of stationary and mobile VUIs for German (technology acceptance ×, technology competence belief ○, and technology control beliefs +) and the technology affinity (+) of the Finnish participants (H7).

Ragarding Hypothesis 8, Statistical correlations can be found between the Big Five dimensions of Neuroticism, Conscientiousness, Openness, and VUI usage. Significant differences in the German population and the Finnish population between stationary usage and the openness to experience scale (German: $r_{pb} = -.106$, $p = 0.005$, Finnish: $r_{pb} = .266$, $p = 0.037$). This observation aligns with theoretical expectations, as previous research suggests cultural variations in attitudes towards technology adoption and openness to new experiences (Bouwman et al., 2007). For other personality dimensions as well as for mobile usage, no significant correlations were found.

5. Discussion

The analysis of the data reveals interesting correlations between hedonistic and utilitarian benefits, trust, and privacy concerns regarding the use of voice assistants. Both in Germany and Finland, higher trust correlates with higher perceived quality of hedonistic and utilitarian benefits. In other words, the higher the trust in the technology, the higher the perceived quality of hedonistic and utilitarian benefits, and vice versa. In the German context, privacy concerns do not play a role in the perception of hedonistic and utilitarian benefits. This might be due to the fact that data privacy does not hold the same societal significance in Finland as it does in Germany, or it could indicate a higher level of awareness among Finnish participants. It should be noted that the current findings are based on a relatively small sample size, which may lead to potential underestimation of effects due to its limited scale.

Another interesting aspect is the relationship between technical knowledge and the use of voice assistants. While there is a clear correlation between technology commitment and voice assistant usage in Germany, there is no such correlation for technology affinity in the Finnish context. Further research should analyze whether this difference is due to openness to new technology or the level of technological education.

6. Conclusion

The present study offers insights into the relationships between trust, privacy concerns, hedonistic and utilitarian benefits, as well as technical knowledge and the use of voice assistants. The results indicate that higher trust in the technology is associated with a higher perceived quality of hedonistic and utilitarian benefits in both countries. However, privacy concerns do not seem to be relevant to the perception of benefits in Finland, unlike in Germany.

Another interesting finding is the difference in the relationship between technical knowledge and the use of voice assistants between the two countries. While technology commitment is associated with higher usage in Germany, there is no such correlation for technology affinity in Finland. This suggests potential cultural differences or differences in the level of technological education that should be further investigated in future studies.

These findings can contribute to optimizing the development of voice assistant technologies and developing targeted measures to promote their acceptance, both in Germany and Finland.

7. Acknowledgments

This research has been funded by the Volkswagen Foundation in the project AnonymPrevent (AI-based Improvement of Anonymity for Remote Assessment, Treatment, and Prevention against Child Sexual Abuse) and by the BMBF in the Project Medinym (KI-basierte Anonymisierung personenbezogener Patientendaten In Klinischen Text- und Sprachdatenbeständen).

8. Bibliographical References

Bitkom. 2022. Umfrage zu den gründen der ablehnung von sprachassistenten in deutschland im jahr 2022.

H. Bouwman, C. Carlsson, F. J. Molina-Castillo, and P. Walden. 2007. Barriers and drivers in the adoption of current and future mobile services in finland. *Telematics and Informatics*, 24(2):145–160.

T. M. Brill, L. Munoz, and R. J. Miller. 2019. Siri, alexa, and other digital assistants: a study of customer satisfaction with artificial intelligence applications. *Journal of Marketing Management*, 35(15-16):1401–1436.

D. Dhagarra, M. Goswami, and G. Kumar. 2020. Impact of trust and privacy concerns on technology acceptance in healthcare: An indian perspective. *International Journal of Medical Informatics*, 141:104164.

European Commission. 2022. *Digital Economy and Society Index (DESI) 2022. Thematic chapters*.

T. Franke, C. Attig, and D. Wessel. 2019. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ati) scale. *International Journal of Human–Computer Interaction*, 35(6):456–467.

M. Haase, J. Krüger, and I. Siegert. 2023. *User Perspective on Anonymity in Voice Assistants*. Springer International Publishing, Cham.

B. Kinsella. 2020. Nearly 90 Million U.S. Adults Have Smart Speakers, Adoption Now Exceeds One-Third of Consumers. voicebot.ai. [Online; posted 28-Apr-2020].

S. Kleinberg. 2018. 5 ways voice assistance is shaping consumer behavior. think with Google. [Online; posted Jan-2018].

J. Krüger and I. Siegert. 2020. das ist schon gruselig so dieses Belauschtwerden - subjektives Erleben von Interaktionen mit Sprachassistenzsystemen zum Zwecke der Individualisierung . In *Sprachassistenten - Anwendungen, Implikationen, Entwicklungen : ITG-Workshop : Magdeburg, 3. März, 2020*, page 29.

D. J. Leiner. 2019. Sosci survey (version 3.1.06). Available at https://www.soscisurvey.de.

A. Leschanowsky, B. Popp, and N. Peters. 2023. Privacy strategies for conversational ai and their influence on users' perceptions and decision-making. In *Proceedings of the 2023 European Symposium on Usable Security*, EuroUSEC '23, page 296–311, New York, NY, USA. Association for Computing Machinery.

G. McLean and K. Osei-Frimpong. 2019. Hey alexa … examine the variables influencing the use of artificial intelligent in-home voice assistants. *Computers in Human Behavior*, 99:28–37.

F. J. Neyer, J. Felber, and C. Gebhardt. 2012. Entwicklung und Validierung einer Kurzskala zur Erfassung von Technikbereitschaft. *Diagnostica*, 58(2):87–99.

C. Olson and K. Kemery. 2019. Voice report: From answers to action: customer adoption of voice technology and digital assistants.

J. Osborne. 2016. Why 100 million monthly cortana users on windows 10 is a big deal. TechRadar. [Online; posted 20-July-2016].

V. Pitardi and H. R. Marriott. 2021. Alexa, she's not human but… unveiling the drivers of consumers' trust in voice–based artificial intelligence. *Psychology & Marketing*, 38(4):626–642.

B. Rammstedt and O. P. John. 2005. Kurzversion des Big Five Inventory (BFI-K). *Diagnostica*, 51(4):195–206.

Y. Sinha, J. Hintz, M. Busch, T. Polzehl, M. Haase, A. Wendemuth, and I. Siegert. 2022. Why eli roth should not use tts-systems for anonymization. In *2nd Symposium on Security and Privacy in Speech Communication*, pages 17–22, Incheon, Korea. ISCA.

Splendid Research GmbH. Januar 2019. Studie: Digitale sprachassistenten und smart speaker. [Online; posted 2021].

M. Vimalkumar, Sujeet Kumar Sharma, Jang Bahadur Singh, and Yogesh K. Dwivedi. 2021. 'okay google, what about my privacy?': User's privacy perceptions and acceptance of voice based digital assistants. *Computers in Human Behavior*, 120:106763.

C. Wienrich, C. Reitelbach, and A. Carolus. 2021. The trustworthiness of voice assistants in the context of healthcare investigating the effect of perceived expertise on the trustworthiness of voice assistants, providers, data receivers, and automatic speech recognition. *Frontiers in Computer Science*, 3.

Author Index