https://doi.org/10.6092/issn.1825-1927/18877 ISSN 1825-1927

Hallucinating (or poorly fed) LLMs?

The problem of data accuracy

Elisabetta Stringhi*

Abstract.

Data scraping is crucial for large language models (LLMs) to gather substantial data for training. However, it raises concerns regarding accuracy. Web scraping systems lack filtering, leading to inaccurate and outdated information. Validating accuracy in large volumes is, however, technically demanding. Nevertheless, data accuracy is vital for output quality and user trust in LLMs. This presentation explores reconciling data scraping with accuracy, considering conflicting rights and interests at stake.

Keywords:

Data scraping; Large Language Models; LLMs; data accuracy; AI; Artificial Intelligence.

1 Introduction

Data scraping practices are the foundation of emerging technologies, especially large language models ("LLMs"), as they enable collection of large volumes of data and model training. While there is no agreed definition of LLMs, it is possible to state that a "large language model" is a subtype of Artificial Intelligence that has been trained on vast amounts of textual data to generate original content[1]. Thus, data scraping is critical to form databases at the core of LLMs-based applications. EU Member States took different approaches in regulating data scraping practices and protecting databases. Regardless of the different IP regulatory approaches, databases constituted for training purposes are tendentially protected or, anyways, kept confidential by developers and/or distributing business companies, as these assets are a determining factor for market competitiveness. Nevertheless, data scraping practices to train LLMs-based applications pose several techno-legal challenges for data protection, given the intrinsic properties of LLMs-based applications training. To mention but a few, data scraping is inherently conflicting with privacy-by-design and privacy-by-default principles, and data minimization. Besides, data scraping practices tendentially entail automated processing, and international data transfers. Above all, data accuracy is a striking critical issue of data scraping, as is further propagated in LLMs training and output generation. Compliance with data accuracy principle requires to keep processed personal data accurate and kept up to date, by means of reasonable measures. However, web scraping systems, such as bots, crawler, and/or spider software do not filter out inaccurate and out-of-date information. Nonetheless, it seems technically unfeasible to require validation of the datasets as accurate before training, given the large volume of datasets. Data accuracy is also key in securing output quality and building trust in LLMs. In fact, despite their potential, LLMs have been documented to produce seemingly credible, yet incorrect outputs in different fields, due to the quality of datasets. Therefore, guarantying accuracy and quality of data sets is a fundamental challenge, as LLMs-based applications are being distributed on the markets and made easily accessible to indefinite end users.

2 Methodology

The aim of this presentation is to critically investigate the feasibility of reconciling data scraping with data accuracy in LLMs. Paragraph 3 will provide a more comprehensive framework of the problem to the reader, from an EU IP law point of view. The

*University of Milan, Information Society Law Center; The article solely reflects the author's view and does not, in any way whatsoever, bind the Italian Data Protection Authority; ✉ elisabetta.stringhi@guest.unimi.it

[1] Bommasani, Rishi, et al., "On the opportunities and risks of foundation models", https://doi.org/10.48550/arXiv.2108.07258.

next section will instead address the major techno-legal challenges raised by LLMs training and deployment. Section 5 will analyse the issue of data accuracy, which stems from data scraping practices, with a focus on the conflicting fundamental rights and interests at stake. Finally, the paper will offer some practical indications, and future research lines.

The methodology incorporated a careful exploration of EU IP law and data protection law regulatory sources, including Directive 96/9/CE, EU Regulation 2016/679. IP law section 3.2. rests on the analysis of national laws implementing Directive 96/9/CE. The data protection paragraph is based also on the measures issued by competent Data protection authorities. To broaden the legal analysis, relevant doctrine has also been scrutinized.

3 Framing data scraping practices in the EU

3.1 Directive 96/9/EC and CJEU case law: 'sui generis' protection and data scraping

3.1.1 Scope of Directive 96/9/CE

Directive 96/9/EC[2] concerns the legal protection of databases in any form. Databases are meant as a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means. Within the CJEU case law, the notion of "database" is wide, unencumbered by considerations of a formal, technical or material nature[3]. The Directive is one of the most relevant frameworks in the EU on protecting databases from unauthorised acts. This instrument distinguishes between databases that are protectable under 'copyright', and those that are instead protected 'sui generis', to protect the 'sweat of the brows' of the maker. Practitioners invoke the sui generis protection to protect the financial, technical and human *substantial* investments made to acquire the set of data. In short, the benefit of said protection does not require any administrative formalities to be fulfilled or any prior contractual arrangement. However, proving the substantial investments may not be taken lightly, as the Court of Justice of the EU ("CJEU") has always strictly interpreted said requirement. As clarified by the CJEU, investment in the creation of a database may consist in the deployment of human, financial or technical resources but it must be substantial in quantitative or qualitative terms. The quantitative assessment refers to quantifiable resources and the qualitative assessment to efforts which may not be quantified, such as intellectual effort or energy, according to the 7th, 39th and 40th recitals of the preamble to the Directive[4].

3.1.2 Mandatory rights of lawful users

The Database Directive establishes mandatory rights for lawful users of databases that are protected by the sui generis right under the Database Directive thereof. Specifically, the rightsholder of sui generis rights in databases may prevent extraction and/or re-utilisation of the whole or of a substantial part, evaluated qualitatively and/or quantitatively, of the contents of that database. The rightsholder may not prevent a lawful user from extracting and/or re-utilising insubstantial parts of its contents, for any purposes whatsoever[5]. A lawful user of a database which is made available to the public in any manner may not cause prejudice to the holder of a copyright or related right in respect of the works or subject matter contained in the database[6]. The CJEU case law has interpreted the term "re-utilization" as referring to *any form* of making available to the public all or a substantial part of the contents of a database by the distribution of copies, by renting, by on-line or *other forms of transmission*[7]. Evidently, said broad interpretation of the notion of 're-utilization' poses several legal challenges to the deployment of LLMs, which i) have been trained on vast databases containing data systematically collected through software and ii) may make available to the public if not all, at least substantial parts of it.

[2] Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

[3] Great Chamber, Case C 444/02, *Fixtures Marketing*, 9 November 2004, paragraphs 20 to 32.

[4] Great Chamber, Case C 444/02, *Fixtures Market*ing, 9 November 2004, paragraph 38.

[5] Article 8 of the Database Directive.

[6] Second Chamber, Case C-30/14, *Ryanair Ltd. v. PR Aviation PV*, 15 January 2015, paragraphs 8-9.

[7] see Great Chamber, Case C 444/02, *Fixtures Marketing*, 9 November 2004; Great Chamber, Case C-203/02, *The British Horseracing Board Ltd. and Others v. William Hill Organization Ltd.*, 9 November 2004, ECLI:EU:C:2004:695; Fifth Chamber, Case C-202/12, *Innoweb BV v Wegener ICT Media BV and Wegener Mediaventions BV*, 19 December 2013, ECLI:EU:C:2013:850.

3.2 Implementing Directive 96/9/EC in Member States

Under Directive 96/9/EC, Member States were required to protect databases by copyright as intellectual creations, and by introducing the so-called ‘database right’. 1st January 1998 was the deadline for implementing the Directive in domestic laws, however only few States (Germany, Sweden, the United Kingdom and Austria) respected it. Instead, most EU States transposed it between 1998 and 2000. During this transposition process, EU Member States undertook harmonized, yet different regulatory approached in implementing Directive 96/9/EC. It is worth noting that, in addition to remedies already provided under their legislation for infringements of copyright or other rights, Member States were further encouraged to introduce appropriate remedies against unauthorized extraction and/or re-utilization of the contents of a database[8]. Depending on how largely or strictly some key-notions of the Directive (such as the ones of ‘extraction’ and ‘re-utilization’) have been interpreted in the Member States, ‘lawful’ extraction or re-utilization operations on database may vary significantly.

3.2.1 Italy

Article 1 of Italian Law No. 633 of 22 April 1941[9], article establishes that the law protects the authorship over creative works, including databases, defined as “collections of works, data or of other elements independent from each other, systematically or methodically disposed, and individually accessible by electronic systems or other ways. The protection of databases may not be extended to their content and is without prejudice to any right that exists on them”. Instead, the German Copyright Act[10] does not include any definition of database notion in general, but instead it defines three categories of databases.

3.2.2 Germany

In Germany, repeated and/or systematic extraction/re-use of insubstantial parts was under the scrutiny of the sui generis right infringement, when the sum of all these extractions and/or re-utilisations was above the threshold of what has been perceived as substantial part of the database. However, The German Federal Court held in landmark judgement “*Zweite Zahnarztmeinung II*”[11], that the repeated and/or systematic extraction/re-use of insubstantial parts infringes the sui generis right also in cases, where the sum of all these extractions and/or re-utilisations was below the threshold, but it would have been above if these acts continue in the future.

3.2.3 France

France transposed the Directive into the French Intellectual Property Code[12] by the law of the 1st of July 1998. Contrarily to what is provided, for instance, in Italy, the French law does not define what extraction and reutilisation mean. However, the French judges mostly rely on CJEU case-law. Said two rights are conferred to the database maker under article 342-2 of the French Intellectual Property Code and it foresees that any action that exceeds the conditions of normal utilization of the database is prohibited. For instance, this was held in a case where a competitor website obtained via a search engine information from a job search website[13]. According to article 342-3 of the French Code, it is not contrary to extract insubstantial parts, if they are lawfully accessed. Nevertheless, the French legislator went beyond the exceptions foreseen in the Database Directive by introducing an exception for datamining.

3.2.4 Gaps and inconsistencies of the Database Directiv

The analysis of these few examples from national IP frameworks shows some gaps and inconsistencies of the Directive, which give rise to not but few legal uncertainties across EU States.

[8] See recital 57, Directive 96/9/EC of the European Parliament and of the Council of 11 March 1996 on the legal protection of databases.

[9] Italian Law No. 633 of 22 April 1941, namely Italian regulation on copyright.

[10] Copyright Act of 9 September 1965 (Federal Law Gazette I, p. 1273), as last amended by Article 25 of the Act of 23 June 2021 (Federal Law Gazette I, p. 1858).

[11] *Zweite Zahnarztmeinung II*. BGH, Urteil v. 01.12.2010, Az. I ZR 196/08.

[12] Loi n. 92-597 du 1er juillet 1992 relative au Code de la propriété intellectuelle.

[13] TGI Paris, 3e ch., 5 Sept. 2001, *Expertises* 2001, 391.

The European Commission conducted an assessment in 2005, to determine the suitability of the Database Directive[14]. The assessment concluded that the Directive had failed to achieve its intended purpose of promoting investment in database production across the European Union. In 2018, the European Commission conducted a second evaluation, which largely reached the same conclusion[15]. It also noted that the Directive might not be equipped to address the emerging challenges posed by AI and big data. However, the Commission chose to maintain the Directive in its current form due to the absence of a clear consensus regarding potential changes or abolition.

The second evaluation received positive reception from many individuals, considering the growing awareness of the various types and applications of data that could drive the future digital economy. Nevertheless, the evaluation raises several questions. For instance, distinguishing between data "creation" and "obtaining" has become increasingly challenging due to the widespread use of automated data gathering, such as data scraping. Consequently, it is likely that the protection of database rights will be reconsidered in the foreseeable future. Like the EU Commission considered, the Database Directive framework might benefit from a legislative update, in order to keep pace with the current technological evolution – and the deployment of LLMs.

Regardless of the different IP regulatory approaches, it is important to bear in mind that databases created to train LLMs are protected or, anyways, kept confidential by business companies, as these assets are a determining factor for market competitiveness.

4 Data scraping as the foundation for LLMs: techno-legal challenges

Data scraping practices pose several techno-legal challenges from an intellectual property and copyright law perspective, without mentioning the critical data protection issues. Fundamentally, said challenges are due to the intrinsic properties of LLMs-based applications training.

Data scraping involves the use of robotic applications deployed to automate tasks, which include scanning, coping and extracting the information published on websites and webpages, and subsequently storing and indexing the extracted information. It is also known that training of LLMs relies on different datasets.

For instance, Open AI used at least five distinct datasets to train ChatGPT model[16]: (1) Common Crawl[17]; (2) WebTex2, text of webpages from all outbound Reddit links from posts with 3+ upvotes; (3) Books1; (4) Books2; and (5) Wikipedia. While certain datasets are static, other ones are dynamic, meaning that are ongoing and continuously updated with further information present online. This is the case of "Common Crawl", which consists in a massive collection of web pages and websites derived from large-scale web scraping and contains petabytes of data collected over twelve years, including raw webpage data, metadata extracts, and text extracts from all types of websites. To give a picture, the Common Crawl dataset constitutes nearly a trillion words.

The next sections will respectively address the IP and data protection critical issues. Paragraph 4.1. will offer the reader a more comprehensive understanding of the technical and legal challenges raised by data scraping, while paragraph 4.2. will focus on data protection aspects. It should also be noted that IP challenges are very often connected with data protection concerns. Take as an example the exploitation of image rights to generate a new audiovisual output, which may involve the processing of personal data.

4.1 IP Law Challenges

4.1.1 Unauthorised use of protected work and contractual breache

To train LLMs, protected pieces of work are extracted, copied and reused. GPT-3 model was trained on systematically scraped 300 billion words online, including books, articles, websites, posts and social media content, without any authorization. Therefore, data scraping may also violate contractual terms and conditions published on proprietary websites to regulate the provision of digital services. It is also worth wondering whether certain contractual limitations, although publicly accessible on the exploited websites, may be effectively enforced, e.g. by giving rise to a contractual liability lawsuit. This would highly depend on the contractual law applicable in the country of residence of the website's owner.

[14] DG Internal Market and Services Working Paper, First evaluation of Directive 96/9/EC on the legal protection of databases, 12 December 2005, Brussels.

[15] Commission Staff Working Document, Evaluation of Directive 96/9/EC on the legal protection of databases, SWD (2018) 147 final, 25 April 2018.

[16] Radford, Alec, et al. "Language Models are Unsupervised Multitask Learners," *Computer Science* (2019).

[17] Available at: https://commoncrawl.org/.

4.1.2 The economic impact of data scraping

Several commenters pointed out how the distribution to the public of LLMs may ultimately deprive websites of their traffic, hence impoverishing authors and creators and, eventually, lead to an unprecedented change in the web as we know it[18].

4.1.3 Implementing counter-measures against data scraping

For the described reasons, many websites and platforms are implementing contractual and technological counter-measures, in order to limit data scraping practices. On this regard, the cases of Twitter and Reddit are noteworthy. These social media platforms adopted technological limitations to prevent systematic and large-scale data scraping, including API changes and visualisation limitations. According to a first declaration of its CEO, Twitter unverified user accounts will only be able to see 600 posts per day, whereas the limit is 300 for newly created accounts. Verified accounts will be allowed to read only 6.000 posts per day. Afterwards, the CEO stated that said limits would soon change to 8000 tweets for verified accounts, 800 for unverified accounts, and 400 for new unverified accounts. Regarding Reddit, it is worth considering that the dataset "WebTex2" is owned by Open AI and built on every webpage linked to Reddit in all posts that received at least 3 "Karma" votes. Said Reddit posts tendentially include outbound links to popular websites, containing copyright work or materials protected by IP rights, and also personal data. According to Reddit's CEO, "The Reddit corpus of data is really valuable. But we don't need to give all of that value to some of the largest companies in the world for free"[19]. Therefore, Reddit changes APIs and updated the pricing policies, despite the silent protest of third-party app developers.

4.2 Data protection challenges

4.2.1 Different data processing activities in LLMs functioning

Looking at the data protection challenges, personal data may be processed in different ways within the LLMs training or deployment context. This is important to clarify, as the debate revolving around personal data processing made via LLMs has not always been clear on this matter. For instance, personal data may have been part of the data set used to train the model. Considering the search plug-ins embedded in certain models, personal data may also be searched. Personal data may also be provided to the system by the end user, for instance as part of a prompt or other kind of input.

4.2.2 Lack of a clear legal basis for data scraping and LLMs training

A major concern regards the absence of a legal basis to process personal data for the purposes of training LLMs. Personal data are extracted from publicly available databases or websites, without the data subject's express authorisation and a legitimate basis to process said personal data. This concern has been first raised by the Italian Data Protection Authority, in its notorious measure against Open AI LLC[20]. Subsequently, most EU DPAs raised similar concerns. Interestingly, the class action against Open AI filed on 28th June 2023[21] underlines this as well, due to the economisation of personal data processing. To give a picture, the class action highlights that a single internet user's information can be valued anywhere from $15 to $40 an individual's online identity can be sold for $1,200 on the dark web. Within these evaluations, the class action stresses that "*Defendants' misappropriation of every piece of data available on the internet, and with it, millions of internet users' personal information without consent, thus represents theft of a value*"[22].

[18] There is a growing body of literature on this regard. See, among others: Agrawal, Ajay, et al. "ChatGPT and how AI disrupts industries." *Harvard business review,* (2022); Åström, Josef, Wiebke Reim, and Vinit Parida. "Value creation and value capture for AI business model innovation: a three-phase process framework." *Review of Managerial Science* 16.7 (2022): 2111-2133; Goldstein, Josh A., et al. "Generative language models and automated influence operations: Emerging threats and potential mitigations." *arXiv preprint arXiv:2301.04246* (2023).

[19] Gintaras Raauskas, "Redditors on Strike but Company Wants OpenAI to Pay Up for Scraping", *CYBERNEWS,* https://cybernews.com/news/reddit-strike-api-openai-scraping.

[20] Garante Italiano per la protezione dei dati personali, Decision of 30 March 2023, no. 112.

[21] *PM v. OpenAI LP*, N.D. Cal., Class action complaint, case 3:23-cv-03199, document no. 1, Filed on 28th June 2023.

[22] *PM v. OpenAI LP*, N.D. Cal., Class action complaint, case 3:23-cv-03199, document no. 1, Filed on 28th June 2023, paragraph no. 3.

4.2.3 Privacy by design and by default principles and data minimization

Data scraping practices have been deemed inherently conflicting with privacy-by-design, privacy-by-default, and data minimization principles. Pursuant to Article 25 of EU Regulation 2016/679 [23], the data controller shall implement appropriate technical and organisational measures to implement data-protection principles, such as data minimisation, in an effective manner and integrate the necessary safeguards into the processing, as well as ensure that, by default, only personal data which are necessary for each specific purpose of the processing are processed. The latter obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons. Article 5 of EU Regulation 2016/679 defines data minimization as a guarantee that the personal data are accurate, relevant and limited to the extent necessary with regard to the purposes for which they are being processed. According to Recital 59 of the Regulation[24], personal data must only be processed if the purpose of the processing may not reasonably be fulfilled by other means. In short, minimisation requires optimising the processing, by limiting the extent of data categories, the level of detail or precision of data, the granularity of the collection and the number of impacted data subjects. During the training phase, the trainer of the LLM model who acts in the quality of data controller shall balance the need of the data to train the model in regarding the risk for the rights and freedoms of the data subjects. Nevertheless, the use of bots, crawlers and other scraping software or automated systems to extract data, including personal data, inherently conflicts with the aforementioned principles. This has also been emphasised in the data protection class lawsuit against OpenAI LLC. In particular, the lawsuit addressed the generation of inaccurate personal data as (un)desired output, by mentioning the case of US law professor Jonathan Turley, who ChatGPT falsely accused of sexually harassing one of his students, even providing a "source" for the purported crime via a news article that it invented[25]. Take another practical, real-life case concerning the processing of health data. Using the online tool "Have I Been Trained," a data subject discovered that her private medical file, including photographs taken of her body as part of clinical documentation when she was undergoing treatment for a rare genetic condition were publicly available online and memorised in the Common Crawl archive[26]. Moreover, LLM-based plug-ins or APIs allow the data controllers to collect and track information, including personal data, from other software applications embedding their LLM, such as, for example, Stripe, Microsoft Teams, Bing, Zillow, etc. to train the same model.

5 The problem of data accuracy

5.1 Defining accuracy in the EU framework

The inherent conflict of data scraping practices with privacy-by-design, privacy-by-default, and data minimization principles gives rise to further challenges, above all the problem of accuracy. Accuracy is included under Article 5(1)(d) of EU Regulation 2016/679 as a key principle providing the basis for the protection of personal data. Personal data shall be accurate and, where necessary, kept up to date; every reasonable step must be taken to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased, or rectified without delay. The wording of Article 5 of EC Convention 108[27] inspired Article 6 of the former Data Protection Directive, which virtually replicated its provisions while adding certain complements, and which in turn has served as a basis for Article 5 EU Regulation 2016/679. Article 5 of Convention 108 contains the same principles, including accuracy of data. Recital n. 39 of EU Regulation 2016/679[28] clarifies that data controllers shall take every reasonable step to ensure that inaccurate personal data are rectified or deleted. Furthermore, pursuant to Recital 71, all personal data, whether directly collected or inferred, must respect the accuracy principle, by means of appropriate mathematical or statistical procedures for the profiling. Therefore, it is mandatory to prove and to document that the procedures used for the inference of the information on a data subject are accurate and, therefore, stable and predictable.

[23] 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).

[24] See Recital 59 of Regulation (EU) 2016/679.

[25] *PM v. OpenAI LP*, N.D. Cal., Class action complaint, case 3:23-cv-03199, document no. 1, filed on 28 June 2023, paragraph no. 216.

[26] *PM v. OpenAI LP*, N.D. Cal., Class action complaint, case 3:23-cv-03199, document no. 1, filed on 28 June 2023, paragraph no. 216.

[27] Convention for the Protection of Individuals with regard to Automatic Processing of Personal Data, ETS No. 108, 28th January 1981, Strasbourg.

[28] See recital 39 of Regulation (EU) 2016/679.

5.2 The challenge of complying with accuracy

When constituting the dataset for training LLMs and training LLMs, the data controller shall still respect said principle, by ensuring that the data that is being processed, generated and linked to the data subject fulfil the accuracy requirements. Accuracy of inferred data may be affected by certain factors. Above all, the presence of mistakes, errors, biases that impact on the training and validation of datasets. Such biases may be inherent, such as bad quality of the data, absent data or selective sampling. They also could be errors of representation and measurement due to the way the dataset is shaped. The model may also evolve, due to the use made by a group of data subjects whose social characteristics may introduce new feedback biases. Also, the implementation of the LM system may affect data accuracy. The Spanish data protection authority[29] suggests, as possible corrective measures, to implement sanitation and traceability metrics and techniques, in order to guarantee the quality of training datasets. Furthermore, by following the classification of the Spanish DPA between "hard" and "soft" data, the controller should carefully assess the accuracy problems that may arise from using or giving greater weight to "soft" data as a source of information.

5.3 Accuracy and particular categories of personal data

Accuracy becomes particularly critical when biometric personal data are processed, including, without limitation, facial recognition, vocal recognition, fingerprint recognition, etc. Within such context, it is pivotal to take into due consideration the problem of false positives, false negatives and other critical issues that might arise whilst implementing LLMs systems. This is especially true when the special category of personal data processed might negatively affect vulnerable data subjects, namely, individuals with disabilities, neurodivergence, or belonging to racial, religious or gender minorities.

5.4 Testing LLMs for adequate application contexts

Finetuning LLMs is crucial to ensure that the resulting services meet the set standards and requirements for a specific envisaged application and use. For instance, it is known that LLMs may not be applied to decision-making contexts. LLMs could rather being employed for "creative" purposes, e.g. creating a new text, poem, piece of advertising, document, or answer questions prior to a decision-making process – activities that shall nevertheless be always supervised by humans, especially in sensitive contexts (i.e. healthcare, justice, etc.). The case of a US lawyer using ChatGPT LM model to create a legal document submitted to a court for litigation purposes [30] effectively demonstrated this concept. Moreover, validating the processing performed by LLMs must be conducted in the same conditions reflecting the real context where the processing itself is expected to be take place. The validation shall be reviewed periodically by the data controller, in order to consider the potential changes and evolutions of the socio-technical context and the processing via LM, also due to technological releases. When LLMs develop and improve, especially when they learn from their interactions with individual users (data subjects) and interactions with other people's data (third data subjects), it's important to periodically evaluate and re-assess the model itself. This re-assessment is essential to ensure that the model's understanding and responses align with ethical[31] and legal standards. By evaluating how the LLMs evolve through these interactions, organizations can maintain the model's accuracy, fairness, and privacy considerations over time.

6 Final remarks: possible reconciliations and solutions

Undeniably, LLMs represent a unique innovation opportunity for many businesses and society as a whole. However, LLMs need to be subjected to appropriate safeguards, especially with regard to the examined accuracy problem. The accuracy principle is highlighted as one of the fundamental data protection principle in EU Regulation "GDPR". Hence, compliance with said Regulation will remain key in securing said goal, as well as in verifying the validity of processing.

[29] Agencia Espanola Proteccion Datos, RGPD compliance of processings that embed Artificial Intelligence: An introduction, February 2020. This document is a translation from the original in Spanish, "Adecuación al RGPD de tratamientos que incorporan Inteligencia Artificial". See also Bommasani, Rishi et al. "On the opportunities and risks of foundation models", https://doi.org/10.48550/arXiv.2108.07258., p. 30.

[30] Armstrong, Kathryne. "ChatGPT: US lawyer admits using AI for case research", *BBC*, 27 May 2023. However, for a different opinion on how foundational model may enhance the level of accuracy in the legal domain, see Nunes Vieira, Lucas, et al. "Understanding the societal impacts of machine translation: a critical review of the literature on medical and legal use cases". *Information, Communication & Society* (2020), 1–18.

[31] On this regard, the EU is at the forefront of ethical standard drafting. Among numerous projects, see Draft Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence, European Commission, December 2018, Brussels. See also Artificial Intelligence for Europe. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions, European Commission, April 2018, Brussels.

Auditing may be a powerful tool, which should not be overvalued, nevertheless. In this sense, auditing procedures should be standardised and adequately performed to assess that, as suggested by AI researcher Timnit Hebru[32], there is a documented process of analysis, development and/or implementation of LLMs, showing respect of accuracy. Furthermore, said process should also evidence the existence or absence of personal data or profiling[33] without a human intervention, as well as the analysis of the efficiency of the anonymisation and the pseudonymisation methods. Said documentation should also show the legal grounds for the processing, and the legitimate interest assessment, if the chosen legal ground is legitimate interest[34]. Furthermore, auditing should also verify the information and the effectiveness of the implemented transparency mechanisms, as well as eventual performance of PIAs. Above all, audits shall focus on the compliance with the application of data protection measures by default and by design, in the training process of LLMs, and the analysis of the accuracy, fidelity, quality and biases of the data used or gathered for the development or the operation of the LLMs component, as well as the data sanitation methods used with regard to the data.

The process of data scraping plays a crucial role in gathering substantial data for the training of large language models (LLMs). However, this method raises significant concerns, particularly regarding the accuracy of the gathered information. Web scraping systems often lack effective filtering mechanisms, leading to the inclusion of inaccurate and outdated data. Furthermore, biases present in the training datasets, which are created through data scraping, pose significant challenges in ensuring unbiased outputs from these models. One of the major hurdles in this scenario is the validation of accuracy and identification of biases within large volumes of scraped data. Technically, this validation process is demanding. Despite the challenges, ensuring data accuracy is vital for maintaining the quality of the output generated by LLMs and for establishing trust among users.

This paper explored the delicate balance between conflicting rights and interests, attempting to address the tension between the need for extensive data and the requirement for precision in training models. By examining these issues, it sought to shed light on potential solutions and strategies for improving the accuracy of large language models while respecting privacy, fairness, and regulatory guidelines.

Future research efforts should focus on standardizing auditing procedures to assess effective compliance with data accuracy and substantial fairness in output generation. As a matter of fact, accuracy is not a new critical issue in the domain of new technologies. However, the alarm in the LLM field is represented by the scale and magnitude of output generation. Several known techniques, such as "AIAs", have been studied in regard with AI systems. Future research lines should focus on adapting said procedures to LLMs training and testing, by adopting adequate metrics. Such metrics should take into account the logic to be implemented, so that it does not originate inaccuracies by design, in order to use mature test models and implement checks to detect design errors. Synthetic data may also be a viable solution to avoid or reduce potential harms deriving from inaccurate generalist datasets. Anyway, future investigations should above all address the source of issue, namely how said LLMs' datasets are acquired and formed, by implementing documentation frameworks aimed at motivating data selection and collection processes[35].To quote AI researchers and authors Birhane and Prabhu,

> "*Feeding AI systems on the world's beauty, ugliness, and cruelty, but expecting it to reflect only the beauty is a fantasy.*"[36].

7 Bibliography

Agencia Espanola Proteccion Datos, RGPD compliance of processings that embed Artificial Intelligence: An introduction, February 2020.

[32] Bender, Emily M., et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Conference on Fairness, Accountability, and Transparency* (FAccT '21), March 3–10, 2021, Virtual Event, Canada, ACM, New York, NY, USA, 2021, pp. 14.

[33] See Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679. Article 29 Working Party. Adopted on 03 October 2017, last updated and adopted on 06 February 2018.

[34] See Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, Article 29 Working Party, April 2014.

[35] On this regard, refer to the proposals of Bender, Emily M. and Batya Friedman, "Data statements for natural language processing: Toward mitigating system bias and enabling better science," *Transactions of the Association for Computational Linguistics* 6 (2018): 587–604; Gebru, Timnit, et al. "Datasheets for Datasets," 2020; and Mitchell, Margaret, et al. "Model cards for model reporting," *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

[36] Ruha, Benjamin. "Race After Technology: Abolitionist Tools for the New Jim Code". Polity Press, Cambridge, 2019. 1541.

A. Agrawal, et al., *ChatGPT and how AI disrupts industries*, in “Harvard business review”, 2022.

Artificial Intelligence for Europe. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions -, European Commission, April 2018, Brussels.

J. Aström, et al., *Value creation and value capture for AI business model innovation: three-phase process framework*, in “RMS”, 16:2111–2133, 2022.

C. Basta, et al., *Evaluating the Underlying Gender Bias in Contextualized Word Embeddings*, in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* 33–39, 2019.

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

E. M. Bender, et al., *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* in *Conference on Fairness, Accountability, and Transparency (FAccT ’21)*, March 3–10, 2021, Canada, ACM, New York, NY, USA, 2021.

S. Bird, *Social Mobile Technologies for Reconnecting Indigenous and Immigrant Communities*, in *People.Policy.Place Seminar*, Northern Institute, Charles Darwin University, Darwin, Australia (2016).

S. L. Blodgett, et al., *Language (Technology) is Power: A Critical Survey of ‘Bias’ in NLP*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Online (2020) 5454–5476.

R. Bommasani, et al. *On the opportunities and risks of foundation models*, https://doi.org/10.48550/arXiv.2108.07258.

L. Breitfeller, et al., *Finding Microaggressions in the Wild: A Case for Locating Elusive Phenomena in Social Media Posts*, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China (2019) 1664–1674.

R. Brewer, and A. M. Piper, *‘Tell It Like It Really Is’: A Case of Online Content Creation and Sharing Among Older Adult Bloggers*, in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016) 5529–5542.

Commission Staff Working Document, Evaluation of Directive 96/9/EC on the legal protection of databases, SWD (2018) 147 final, 25 April 2018.

DG Internal Market and Services Working Paper, First evaluation of Directive 96/9/EC on the legal protection of databases, 12 December 2005, Brussels.

Draft Ethics Guidelines for Trustworthy AI, High-Level Expert Group on Artificial Intelligence, European Commission, December 2018, Brussels.

E. Fast, et al., *Shirtless and Dangerous: Quantifying Linguistic Signals of Gender Bias in an Online Fiction Writing Community*, in *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 10 (2016).

D. Fišer, et al. (Eds.), *Proceedings of the 2nd Workshop on Abusive Language Online,* (ALW2) Association for Computational Linguistics, Brussels, Belgium, 2018.

S. T. Fiske, *Prejudices in Cultural Contexts: Shared Stereotypes (Gender, Age) Versus Variable Stereotypes (Race, Ethnicity, Religion)*, in “Perspectives on Psychological Science” 12, no. 5 (2017): 791–799.

T. Gebru, et al., *Datasheets for Datasets*, 2020. *Communications of the ACM* 64.12 (2021): 86-92.

J.A. Goldstein, et al., *Generative language models and automated influence operations: emerging threats and potential mitigations*, arXiv preprint arXiv:2301.04246 (2023).

D. K. Kanbach, et al., *The GenAI is out of the bottle: generative artificial intelligence from a business model innovation perspective*, in “Review of Managerial Science”, 13 September 2023.

K. Kurita, et al., *Measuring Bias in Contextualized Word Representations*, in *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (2019) 166–172.

M. Martindale, and M. Carpuat, *Fluency Over Adequacy: A Pilot Study in Measuring User Trust in Imperfect MT,* in *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas* (Volume 1: Research Track), Association for Machine Translation in the Americas, Boston, MA, 2018, pp. 13–25.

K. Mc Guffie, and Alex Newhouse, *The Radicalization Risks of GPT-3 and Advanced Neural Language Models*, in "Computers and Society", (2020).

J. Mendelsohn, et al., *A Framework for the Computational Linguistic Analysis of Dehumanization*, in "Frontiers in Artificial Intelligence", 3 (2020): 55.

M. Mitchell, et al., *Model cards for model reporting*, *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

R. C. Moore, and W. Lewis, *Intelligent Selection of Language Model Training Data*, in *Proceedings of the ACL 2010 Conference Short Papers*, Association for Computational Linguistics, Uppsala, Sweden (2010) 220–224.

J. Pratik, et al. *The State and Fate of Linguistic Diversity and Inclusion in the NLP World*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, (2020) 6282–6293.

A. Radford, et al., *Language Models are Unsupervised Multitask Learners*, in "Computer Science" (2019).

S. T. Roberts, et al. (Eds.), *Proceedings of the Third Workshop on Abusive Language Online*, Association for Computational Linguistics, Florence, Italy, 2019.

D. Rodriguez Maffioli, *Copyright in Generative AI Training: Balancing Fair Use through Standardization and Transparency,* DOI:10.13140/RG.2.2.18478.48961, 2023.

B. Ruha, *Race After Technology: Abolitionist Tools for the New Jim Code,* Polity Press, Cambridge, 2019. 1541.

E. Seo Jo, and T. Gebru, *Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning*, in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020) 306–316.

E. Sheng, et al., *The Woman Worked as a Babysitter: On Biases in Language Generation*, in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Association for Computational Linguistics, Hong Kong, China (2019) 3407–3412.

M. Tulio Ribeiro, et al., *Beyond accuracy: Behavioral testing of NLP models with CheckList*, arXiv preprint arXiv:2005.04118 (2020).

S. Umoja Noble, *Algorithms of Oppression: How Search Engines Reinforce Racism*, NYU Press, 2018.

M. Young, et al., *Toward Inclusive Tech Policy Design: A Method for Underrepresented Voices to Strengthen Tech Policy Documents*, in "Ethics and Information Technology" (2019): 1–15.