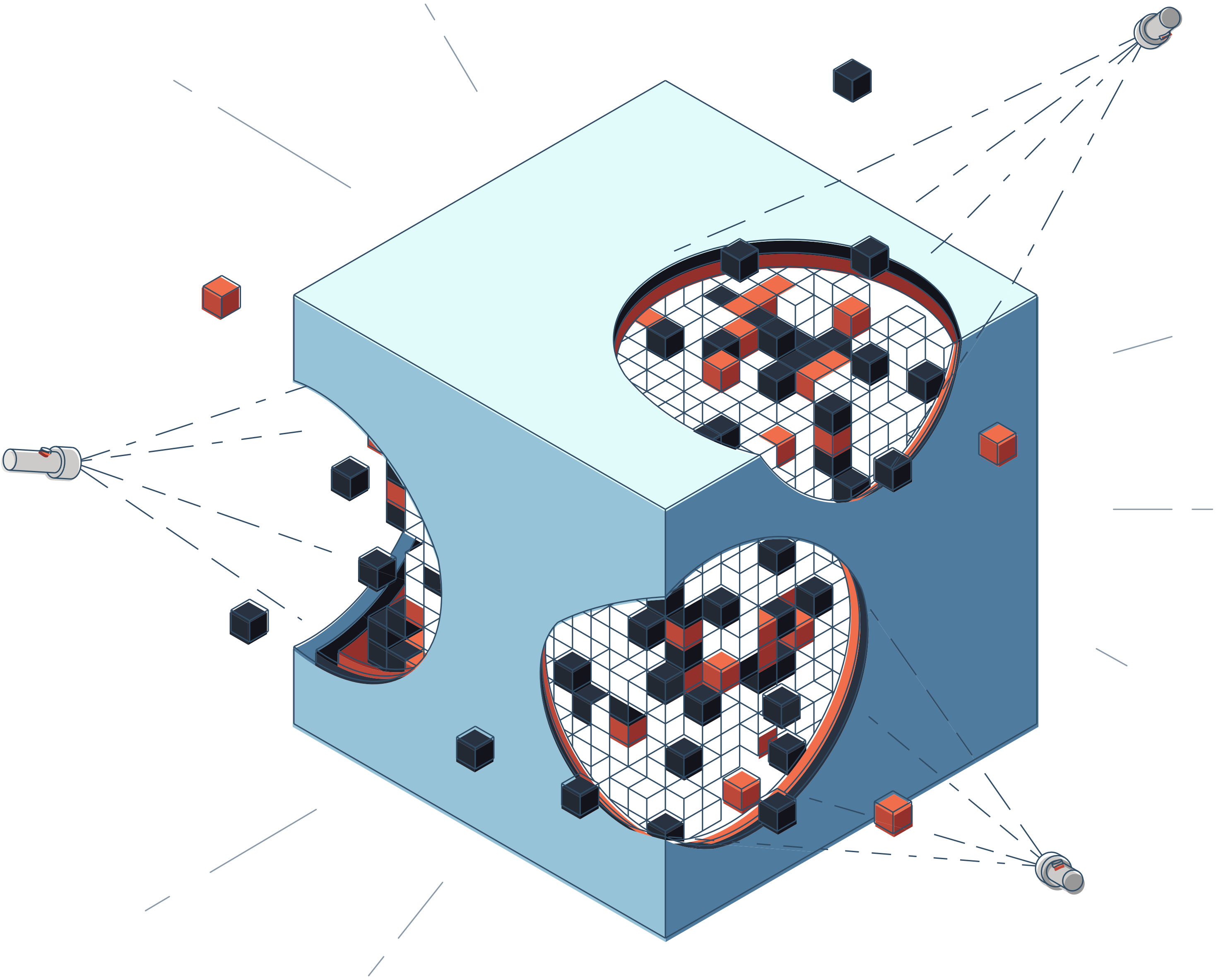
Sufficiently detailed ?

A proposal for implementing the AI Act's training data transparency requirement for GPAI



OPEN _FUTURE

moz://a

June 2024

1. A new data transparency requirement under the AI Act[1]

Under the European Union's new AI Act,[2] there is now[3] **a legal obligation to disclose information about content used to train general-purpose AI models** (GPAI). To facilitate transparency, providers of such models are expected to produce and publish **summaries of the training data**. These summaries must present an overview of the data sources and sets involved, including private and public databases, and include narrative explanations. They should be prepared according to **a template** provided by the AI Office.

The AI Act's preamble states that the purpose of these "sufficiently detailed summaries" is to **facilitate the exercise and enforcement of rights under Union law by parties with a legitimate interest**. The legitimate interest may relate to the protection of copyright, which is explicitly mentioned in the recital 107. However, the range of legitimate interests of parties interested in increased transparency of data used in the development of GPAI goes **beyond copyright issues**.

The purpose of this paper is twofold. First, it clarifies the **categories of rights and legitimate interests that justify access to information about training data**. Second, it provides **a blueprint for the forthcoming template for the "sufficiently detailed summary."** The blueprint seeks to strike a balance between serving these interests in **a meaningful way** while respecting the rights of all parties concerned, including the privacy of data subjects, and taking account of the need to protect trade secrets and confidential business information.

[1] This paper was written by Zuzanna Warso (Open Future), lead author, and Maximilian Gahntz (Mozilla Foundation) and Paul Keller (Open Future), contributing authors. The authors would like to thank Yacine Jernite, Shayne Longpre, Abeba Birhane, Stefan Baack, Kris Shrishak, Aviya Skowron, Alex Hanna, Mark Dingemanse, Frederike Kaltheuner, Claire Pershan, Michał "rysiek" Woźniak, Natali Helberger, Joao Quintais, Toni Lorente for their valuable feedback. This paper and accompanying blueprint build on previous work on AI training data documentation, including: Timnit Gebru et al., "Datasheets for Datasets" (arXiv, 1 December 2021), https://arxiv.org/abs/1803.09010; Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson, "Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI" (arXiv, 3 April 2022), https://arxiv.org/abs/2204.01075; Emily M. Bender and Batya Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," Transactions of the Association for Computational Linguistics, (December 2018), 6: 587–604, https://aclanthology.org/Q18-1041/; Shayne Longpre et al., "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI" (arXiv, 4 November 2023), https://arxiv.org/abs/2310.16787.

[2] Text of the AI Act as approved by the Council: European Parliament and Council of the European Union, "Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," https://data.consilium.europa.eu/doc/document/PE-24-2024-INIT/en/pdf.

[3] The AI act was signed into law on 13 June 2024. It will enter into force 20 days after publication in the Official Journal of the European Union which is expected to happen some time in July. The GPAI rules will take effect within 12 months thereafter. As a result, providers of GPAI models will be required to publish data summaries starting in mid-2025.

Artificial Intelligence Act

Article 53 (1) d: Providers of general purpose AI models shall [...] (d) draw up and make publicly available a sufficiently detailed summary of the content used for training of the general-purpose AI model, according to a template provided by the AI Office.

Recital 107: In order to increase transparency on the data that is used in the pre-training and training of general purpose AI models, including text and data protected by copyright law, it is adequate that providers of such models **draw up and make publicly available a sufficiently detailed summary of the content used for training the general purpose model**. While taking into due account the need to protect trade secrets and confidential business information, this summary should be **generally comprehensive in its scope instead of technically detailed** to facilitate parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law, for example by listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used. It is appropriate for the **AI Office to provide a template** for the summary, which should be simple, and effective, and allow the provider to provide the required summary in narrative form.

2. “Legitimate interest”

Access to AI training data and transparency about data sources and data sets are essential for improving accountability in AI development. For various reasons, including privacy, data protection, and copyright concerns, not all training data can be openly shared. However, these same reasons necessitate training data transparency. In particular, trade secrets should not serve as a blanket justification for not disclosing information about the content used to train GPAI in a situation where there are legitimate reasons to make information about the training data public.

The High-Level Expert Group on AI set up by the European Commission in June 2018 has recognized transparency and traceability of data as crucial components of achieving “Trustworthy AI.”[4] Data transparency is a prerequisite for improving large model interpretability, enabling benchmarking and auditability, and facilitating reproducibility in AI research and development. Recital 107 of the AI Act made this principle actionable for general-purpose AI models. It has operationalized it by stating that the publicly available "sufficiently detailed summary" is specifically intended to enable "parties with legitimate interests, including copyright holders, to exercise and enforce their rights under Union law."

Individuals and groups may need access to information about GPAI training data for a variety of legitimate reasons. This section outlines the rights and legitimate interests that transparency of GPAI training data would serve.

[4] High Level Expert Group on Artificial Intelligence, “Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment,” 2020, https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

The protection of copyright

The transparency provision in Article 53(1)d stems from amendments proposed by the European Parliament. These amendments were introduced in response to concerns raised by various organizations representing authors and other rightholders within the cultural and creative industries, as many GPAI models, particularly generative AI models, are trained on the results of human creativity, much of which is protected by copyright.[5] Transparency of the training data is necessary to allow creators to determine whether their works have been included in the GPAI model training data.

Under Article 4(3) of the 2019 Directive on Copyright in the Digital Single Market (CDSM), authors and other rightholders have the right to opt out of the use of their works for text and data mining (TDM). The AI Act clarifies that the training of generative AI models is a form of text and data mining. Against this backdrop, Article 53(1)c of the AI Act requires “providers of general-purpose AI models” to “put in place a policy to comply with Union copyright law, and in particular to identify and comply with, including through state of the art technologies, a reservation of rights expressed pursuant to Article 4(3) of Directive (EU) 2019/790.” In the context of copyright, the transparency provision in Article 53(1)d is thus intended to **allow authors and other rightholders to verify that providers of generative AI models are complying with the two conditions for lawful text and data mining contained in the TDM exception in Article 4 of the CDSM Directive**. First, that the works used as training data have been lawfully accessible (Article 4(1)) and second, that they have not been opted out (Article 4(3)).

Privacy and data protection rights

In addition to copyrighted material, the content used for training, validating, or testing general-purpose AI models could include personal data. According to European legislation on personal data protection, individuals (the data subjects) have several pertinent rights in this scenario. These rights include the ability to access their personal data, correct any inaccuracies in their personal data, and request the deletion of their personal data. Asserting these rights can be difficult in the case of AI training data.

Some GPAI model providers claimed they were unable to specify what personal data was contained in the training data.[6] However, a recent report from the European Data Protection Board states that technical impossibility cannot be invoked to justify non-compliance with

[5] Providers of popular generative models have asserted that “it would be impossible to train today’s leading AI models without using copyrighted materials”. See e.g., OpenAI, Submission to the UK House of Lord, Communications and Digital Select Committee inquiry: Large language models, 2024, https://committees.parliament.uk/writtenevidence/126981/pdf/.

[6] See reports on complaints to data protection authorities in Germany and Poland about unlawful processing of personal data: Natasha Lomas, "ChatGPT’s 'hallucination' Problem Hit with Another Privacy Complaint in EU," TechCrunch, 29 April 2024, https://techcrunch.com/2024/04/28/chatgpt-gdpr-complaint-noyb/; Natasha Lomas, "Poland opens privacy probe of ChatGPT following GDPR complaint," TechCrunch, September 21, 2023, https://techcrunch.com/2023/09/21/poland-chatgpt-gdpr-complaint-probe.

the requirements of the GDPR.[7] Moreover, researchers have built search indexes over large training data collections, enabling searches for names, addresses, phone numbers, and other information, demonstrating that data protection rights can be enforceable.[8] Transparency of GPAI training data is therefore necessary to **allow data subjects, first, to determine whether model providers are processing personal data and, second, to exercise their rights under personal data protection laws**.

In the context of content used to train general-purpose AI models and personal data protection, it is important to differentiate between model inputs and outputs. Providers might contend that privacy and data protection are safeguarded as long as models do not produce outputs that contain personal data. Yet, the possibility of models memorizing – and subsequently reproducing – training data, including personal (or copyright-protected) data, has been well-established[9] and needs to be accounted for in any assessment of risks to people's privacy and data protection rights.

Further, a perspective focused on model outputs oversimplifies the data protection and privacy risks tied to AI systems. As emphasized, for example, in the case law of the European Court of Human Rights, the mere storage of data relating to an individual's private life constitutes an interference within the meaning of Article 8 of the European Convention on Human Rights, which guarantees the right to respect for private and family life, home, and correspondence.[10] Although companies are not directly bound by the European Convention on Human Rights, states have an obligation to protect individuals from abuses by companies. Moreover, businesses have the responsibility to account for the impact of their operations on human rights.

In light of these considerations, even if a GPAI model does not produce personal data, the mere inclusion of such data in the training set by the GPAI model provider poses a data protection and privacy issue. Thus, transparency about the content used to train general-purpose AI models and whether it includes personal data is more than a step toward

[7] See paragraph 7 of the European Data Protection Board, "Report of the work undertaken by the ChatGPT Taskforce," 2024, https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf.

[8] See, e.g., Aleksandra Piktus et al., 'GAIA Search: Hugging Face and Pyserini Interoperability for NLP Training Data Exploration' (arXiv, 2 June 2023), http://arxiv.org/abs/2306.0148; Yanai Elazar et al., 'What's In My Big Data?' (arXiv, 5 March 2024), http://arxiv.org/abs/2310.20707.

[9] See, for example, Stella Biderman et al., "Emergent and Predictable Memorization in Large Language Models," 37th Conference on Neural Information Processing Systems (2023), https://proceedings.neurips.cc/paper_files/paper/2023/file/59404fb89d6194641c69ae99ecdf8f6d-Paper-Conference.pdf; Valentin Hartmann et al., "SoK: Memorization in General-Purpose Large Language Models" (arXiv, 24 October 2023), https://arxiv.org/abs/2310.18362; Carlini et al., "Extracting Training Data from Large Language Models" (arXiv, 15 June 2021), https://arxiv.org/abs/2012.07805; "The New York Times Company v. Microsoft Corporation," U.S. District Court Southern District of New York, 1:23-cv-11195, (27 December 2023), https://nytco-assets.nytimes.com/2023/12/NYT_Complaint_Dec2023.pdf.

[10] European Court of Human Rights. 2008. S. and Marper v. The United Kingdom. Application nos. 30562/04 and 30566/04. Judgment (Strasbourg). The ECtHR has dealt with cases related to data protection under Article 8 of the European Convention on Human Rights. One of the key principles established by the ECtHR is that operations involving personal data, such as collection, storage and use, may infringe the right to respect for private life.

ensuring that the outputs would not contain such information; it is also a prerequisite to understanding the potential risks associated with the inputs.

The right to science and academic freedom

The Charter of Fundamental Rights protects the freedom of the arts and sciences, which encompasses the freedom to conduct research, pursue academic activities, and engage in scientific inquiry without undue interference. The right to science does not require private parties, such as AI providers, to share AI training data. However, it does create an obligation for states to establish an environment that allows the scientific community to exercise its scientific freedom. This includes conducting scientific reviews, applying scientific scrutiny, replicating experiments, and validating scientific results.

Researchers often face obstacles in carrying out such investigations. A notable example is the evaluation and red teaming of generative AI systems. Despite the critical role of independent evaluation in identifying the risks posed by these systems, the terms of service and enforcement strategies used by prominent AI companies can discourage good-faith safety evaluations. This has led some researchers to fear that conducting such research or publishing their findings could result in account suspensions or legal reprisals. Although some companies offer researcher access programs, these are limited in scope and often seen as an inadequate substitute for independent research access due to limited community representation, inadequate funding, and lack of independence from corporate incentives.[11] **Because of the direct link between training data and model behavior,**[12] **the lack of access to and information about training data has created challenges in understanding the various forms of risk and harm associated with the use of AI.**[13] In addition, replication and validation of experiments and their results are critical to robust science. However, without well-documented training data, experiments cannot be replicated, and the validity of claims cannot be verified.

The current situation highlights the need for a more robust and enabled ecosystem to study and investigate AI systems and critical components used to train them, such as data, and underscores the importance of policies that allow researchers the freedom to conduct scientific research. These policies must include a requirement that AI providers be transparent about the data used to train models. Information about training data at the level of detail recommended in the template blueprint is necessary to operationalize this requirement, as it will **allow researchers to critically evaluate the implications and limitations of AI development**, identify potential biases or discriminatory patterns in the data, and reduce the risk of harm to individuals and society by encouraging provider accountability.

[11] See, e.g., Shayne Longpre et al., "A Safe Harbor for AI Evaluation and Red Teaming”, Knight First Amendment Institute, March 5, 2024, https://knightcolumbia.org/blog/a-safe-harbor-for-ai-evaluation-and-red-teaming; and the call for creating a “Safe Harbout for Independent AI Evaluation: https://sites.mit.edu/ai-safe-harbor/.

[12] See, e.g., Shayne Longpre et al., “A Pretrainer’s Guide to Training Data: Measuring the Effects of Data Age, Domain Coverage, Quality, & Toxicity” (arXiv, 13 November 2023), http://arxiv.org/abs/2305.13169. Abeba Birhane et al., “Into the LAIONs Den: Investigating Hate in Multimodal Datasets” (arXiv, 6 November 2023), http://arxiv.org/abs/2311.03449.

[13] See, e.g., Shayne Longpre et al., “A Safe Harbor for AI Evaluation and Red Teaming” (arXiv, 7 March 2024), http://arxiv.org/abs/2403.04893.

The prohibition of discrimination and respect for cultural, religious, and linguistic diversity

Transparency about the content used to train models is key for investigating and mitigating bias in all artificial intelligence systems. **Transparency in training data is not an end in itself but a necessary condition for preventing AI systems from being used in ways or contexts that may result in discriminatory outcomes**. In the development of GPAI, transparency in training data is also critical. Because of the broad applicability of GPAI models in different sectors, any biases in the training data can be mirrored across multiple downstream applications. This is not a localized problem affecting a single group or context. Instead, it has the potential to affect a wide range of people in a variety of settings, amplifying the effects of any inherent biases. More robust and comprehensive documentation of the data used to train an upstream GPAI model will also enable downstream providers – such as companies integrating a GPAI model into their products and services or public institutions – to conduct more targeted testing of such a model, including for discriminatory biases and other risks, and subsequently put in place more targeted and effective mitigations and other countermeasures.

Having **information about the training data is further critical for those who have been discriminated against or harmed otherwise in a process in which a GPAI model was used and wish to lodge a complaint or seek redress**. For example, more transparency may help determine whether the discrimination was the result of negligence or omission in the use of an AI system and hold those responsible accountable. The AI Liability Directive proposal[14] recognizes the unique characteristics of AI, including complexity, autonomy, and opacity, which make it difficult or prohibitively expensive for victims to identify the responsible party and prove the requirements for a successful liability claim. Transparency of GPAI training data is a step toward addressing this challenge by empowering those affected by the use of the models.

Furthermore, the EU Charter of Fundamental Rights emphasizes the importance of respecting cultural, religious, and linguistic diversity. This principle is important in the context of GPAI models because it emphasizes the need for the models to be inclusive and representative of various cultural, religious, and linguistic backgrounds. The training data for GPAI models should reflect human diversity, including linguistic and regional diversity. **Knowledge about the training data is critical in determining whether AI systems favor one group over another and whether they can be expected to interact appropriately with users from various backgrounds while providing the same level of service.**

Fair competition

Under the AIA, providers of general-purpose AI systems are obligated to prepare and share information on data sets and training methodologies with the AI Office upon request. While

[14] See: European Commission, "Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)," 2022, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496.

this is a significant step toward ensuring accountability and transparency in AI systems, it is important to note this information is not intended to be shared publicly.

However, secrecy regarding training data can become a significant competitive advantage. This poses a challenge to other stakeholders in the AI ecosystem, including smaller and medium-sized AI providers and startups. Some companies have gone as far as admitting that they withhold information about their training data, among other things, to secure a competitive advantage.[15] **A lack of transparency around training data can also enable anti-competitive practices by the industry's incumbents** — for instance, it can help them secure preferential licensing agreements with rightholders (e.g., publishers) and can obstruct smaller companies from identifying potential sources of (and partners to share) high-quality training data.

Access to sufficiently detailed information about training data and its sources can thus help to level the playing field and **prevent larger or more established companies from gaining an unfair advantage without revealing confidential information and trade secrets**. Through a lightweight intervention such as mandating robust transparency measures around training data, the Commission can therefore help spur more competition in the AI industry. This approach would reflect the EU's commitment to ensuring a fair and level playing field for businesses, which is an integral part of the Treaties (Art. 101-109 TFEU).

Other legitimate interests that require transparency of training data

From the standpoint of consumer protection, the transparency of training data also holds significant importance. In the absence of transparency in AI training data, consumers lack insight into the possible biases, errors, or distortions that the AI systems they use might carry forward. Transparency of training data would help them understand the limitations and potential inaccuracies of AI systems, thereby empowering them to make informed decisions. In this context, consumer organizations have a critical role to play. They can educate consumers about their rights and hold AI providers accountable. **Transparency of AI systems would empower consumers and consumer organizations to assert their rights under consumer protection laws and to confront AI development practices that may be unfair or misleading.**

Transparency of training data is also important from a child rights perspective. Children are increasingly interacting with AI systems, whether through educational software, digital toys, or online platforms. These interactions can have a significant impact on their development, learning, and well-being. However, without transparency into the training data used by these AI systems, it is difficult to assess whether the systems are safe, appropriate, and beneficial for children. For example, **an AI system trained on data containing inappropriate, harmful, or toxic content could expose children to such content, violating their right to safety and protection**.

[15] See, e.g., "GPT-4 Technical Report," OpenAI, 2023, https://cdn.openai.com/papers/gpt-4-system-card.pdf.

3. Scope and granularity of the summary

Recital 107 provides some background and more detailed guidance and instruction on the new transparency requirement. According to the recital, the summary should cover data used “in the pre-training and training of general purpose AI models.” To adequately reflect the intent of the article, the transparency requirement should be interpreted to include data used **in all stages of model training — from pre-training to fine-tuning**.

The summary must encompass various types of data, including but not limited to text and data protected by copyright law. Providers must ensure that the summary is **comprehensive in scope** to enable stakeholders with legitimate interests, such as copyright holders or data subjects, to exercise their rights under Union law effectively. While the summary should not be "overly technical" in that the degree of complexity obstructs transparency to both experts and laypeople, it should contain **sufficient technical detail to provide meaningful insights** for all relevant stakeholders.

The summary should include **a clear listing of the primary data collections or sets utilized** in training, such as large private or public databases, along with **narrative explanations of other data sources used**. For the purposes of the summary and to avoid confusion, the term "**data source**" should refer to the origin of the data set, which can include a variety of sources such as web data, data obtained through commercial arrangements, user data collected from system interactions, and data created specifically for the system by data workers. On the other hand, "**dataset**" (or, in the language of the recital, “data collection”) should refer to the processed and filtered data points extracted from these data sources and represented in a consistent format. This includes data points used at various stages of the development process to train models.

In the Annex, we provide a blueprint for the forthcoming "sufficiently detailed summary" template, designed to serve the interests and rights outlined in the preceding sections.

4. Conclusion

Transparency around data used to train AI can serve a variety of different functions: it can strengthen the ability of people and organizations to exercise their rights, it can enable independent research and scrutiny of one of the key inputs in the AI development process, and it can enhance accountability across the AI industry. At the same time, it has become very clear that **opacity around training data is strategically used to shield companies developing GPAI from scrutiny and competition at the expense of both rightholders and other parties.**

The sufficiently detailed summary of data used to train GPAI stipulated in the AI Act thus provides an important mechanism to enhance transparency in this respect. The AI Act is further clear in stating that this summary should protect other parties' legitimate interests, and – as outlined above – there is a range of parties whose interests are affected in this context. But for the summary to be effective in practice, the information provided by GPAI developers needs to be both **meaningful and comprehensive**. It must further be useful to both rightholders and technical experts. This is the standard to which the template to be provided by the European Commission should be held.

The blueprint for the template outlined in this brief, developed in collaboration with experts from various sectors and disciplines, sets out what an effective summary and meaningful documentation of training data should look like. It can also serve as input to discussions on this issue and as a baseline for the Commission's implementation work in developing the template.

Acknowledgments

Authored by Zuzanna Warso (Open Future), lead author, and Maximilian Gahntz (Mozilla Foundation) and Paul Keller (Open Future), contributing authors.

The authors would like to thank Yacine Jernite, Shayne Longpre, Abeba Birhane, Stefan Baack, Kris Shrishak, Aviya Skowron, Alex Hanna, Mark Dingemanse, Frederike Kaltheuner, Claire Pershan, Michał “rysiek” Woźniak, Natali Helberger, Joao Quintais, Toni Lorente for their valuable feedback.

Annex 1: Blueprint of the template for the summary of content used to train general-purpose AI models (Article 53(1)d AIA)

1. General information

1.1. **Size:** The total size of the training data (e.g., the total number of works or tokens, including information about the tokenization process for better comparability), the total number of data sets and their size in GB;

1.2. **Ethical review:** Has an ethical review process (e.g., by an institutional review board or an external review board) been conducted as part of the following steps related to the training data:

- data collection,
- data filtering,
- further steps of data processing (yes/no)

If yes, information on who carried out the ethical review, the outcome, the main objectives guiding the review process, and the criteria used to evaluate the outcome.

2. Data sources and data sets:

2.1. Information on the **data collection**: information on the sources from which data used in all stages of training, from pre-training to fine-tuning, was obtained – i.e., whether it was:

- scraped from the internet. If so, information about the crawling methodology used to obtain the data, including, for example, the seed and link selection criteria, as well a weighted list of the top 5 percent or 100 000 domains by data modality (e.g., text, images, video).
- collected from public repositories. If so, the names of those public repositories and steps taken to convert the data archive into a training data set.
- sourced from proprietary databases. If so, information about the source and the owner of the database should be provided.
- acquired or licensed from third parties. If so, information about the third-party source and the license, including whether the licensing arrangement is exclusive or not, should be provided.
- generated by users of products or services offered by the provider. If so, by which products or services.
- generated by the provider. If so, by what methods.
- or obtained through other means. If so, details of the means should be provided.

2.2. For each data source, the **date range** of the training data. This should include **data cutoffs** for online data and, if applicable, other data sources such as archival data.

2.3. Information on mechanisms and policies that have been implemented to ensure respect for **opt-outs** under Article 4 of the CDSM.

2.4. Information on **the legal basis** for data collection and processing, including, if applicable, the legal basis for the processing of personal data under the GDPR.

2.5. Information on **anonymization techniques** that were implemented. When personal data is not anonymized, a justification for not anonymizing and information on how long the data is kept before being deleted should be provided.

2.6. Information on **intermediaries** or entities involved in the acquisition, sharing, or transfer of the data, including information on data licensing agreements or permissions.

2.7. **List of data sets** that were used in training the model. The list must include information about **what percentage** of the total training data each data set represents. For each data set:

- Data set identifier: the data set's name, if applicable, along with a link to the data set, as well as the name and URL for the data collection of which the data set is a part.
- Owners or curators: groups or teams that own or assemble the data set, if applicable, authors associated with the dataset.
- Data set domain: humans (e.g., data set about people), objects (e.g., data set about places or objects), etc.
- Type/modality: text, image, audio, video, etc. If multimodal, what combination (e.g., image-text pairs).
- If language data, what language/languages.
- Information about the purpose for which the data set was used, explaining why this is the right data set for a particular purpose.

3. Data diversity

3.1. Proportions of the data of relevant categories and characteristics (such as linguistic or regional diversity) included in the training data.

3.2. Information on the steps taken to ensure diversity and representativeness of training data across relevant categories (e.g., demographics, languages).

4. Data processing

4.1. Information on the methodology and processes used for annotation and labeling (e.g., crowdsourcing, supervised learning, etc.).

4.2. If crowdsourcing was used, information about the recruiting criteria of the annotators, the specific task they were given, and how the quality of their work was evaluated to understand the fitness of the approach to the purpose of the AI model development.

4.3. Explanation of measures to ensure reliability and accuracy of annotations and labeling.

4.4. Description of preprocessing steps applied to the different types of training data (text, image, speech, etc.), such as filtering, data cleaning, tokenization, feature extraction, detoxifying, etc.

This should include information on the:

- filtering processes and methods used for **inclusion** (i.e., metrics or cosine similarities used to decide the cut-off point for what content makes it into a data set), with a sufficient level of detail to understand the motivations behind using these methods and how they have been employed; and
- filtering processes and methods used for **removing** outliers, anomalies, etc., with a sufficient level of detail to understand the motivations behind using these methods and how they have been employed.

4.5. Explanation of how preprocessing may have affected the characteristics of the data used for training, e.g., how it might have affected feature representation, class imbalance, etc.

4.6. Data Sampling: Information on sampling methods used to select training data (e.g., random sampling, systemic sampling, stratified sampling, under- or over-sampling, or other techniques).