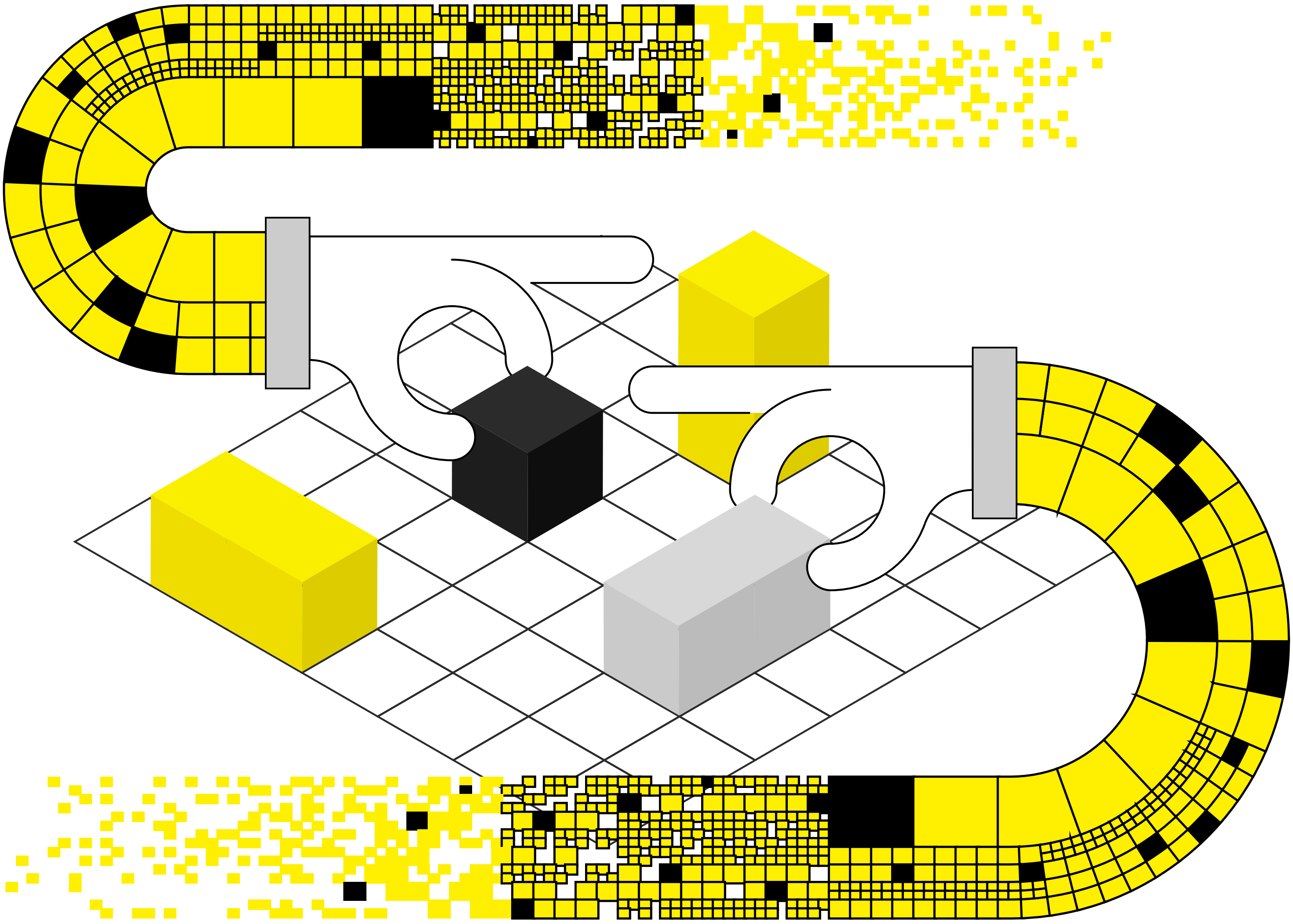
COMMONS-BASED DATA SET GOVERNANCE FOR AI



OPEN
_FUTURE

MARCH 2024

Alek Tarkowski, Zuzanna Warso

INTRODUCTION: THE CHALLENGE OF RESPONSIBLE AI TRAINING

This white paper proposes an approach to sharing data sets for AI training as a public good, governed as a commons. By adhering to six principles of commons-based governance, data sets can be managed in a way that generates public value while making shared resources resilient to extraction or capture by commercial interests.

Over the past five years, AI development has shifted from being largely a research discipline, rooted in the ethos of open source and open science. Today, it is mainly driven by the commercial efforts of a few companies that also reap the greatest benefits from shared knowledge and research – and build proprietary solutions on top of it. Open-source AI development and academic research play a role that's still important but diminishing.[1] And the value that these public interest actors generate is often captured. This trend is exacerbated by the immense resource hunger inherent in AI technology, including data, labor, environmental resources, and computing power.[2]

Today, we face a risk that AI development and deployment will lead to greater inequalities and further concentration of power within a select few companies and institutions. Existing platform monopolies are being strengthened through AI development and advantages in computing power, proprietary data, and customer bases. This unchecked concentration of power raises concerns about the social and economic impact of AI by marginalizing smaller companies and independent researchers and stifling competition.[3]

The term AI "democratization" is used to describe the goal of shifting this trend and making AI systems accessible to a broad range of individuals and organizations, and also of making positive outcomes spread more broadly. Democratizing AI can refer to AI development, use of AI, distribution of value generated by AI, and governance over AI systems.[4] The democratization of AI points to the possibility of reducing the concentration of power in the AI sector and addressing the negative social impacts of AI through measures that distribute power and

[1] Besiroglu, Tamay, Sage Andrus Bergerson, Amelia Michael, Lennart Heim, Xueyun Luo, and Neil Thompson. "The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?" arXiv, January 8, 2024. https://doi.org/10.48550/arXiv.2401.02452.

[2] Widder, David Gray, Sarah West, and Meredith Whittaker. "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI." SSRN Scholarly Paper. Rochester, NY, August 17, 2023. https://papers.ssrn.com/abstract=4543807.

[3] Narechania, Tejas N., and Ganesh Sitaraman. "An Antimonopoly Approach to Governing Artificial Intelligence." SSRN Scholarly Paper. Rochester, NY, October 9, 2023. https://doi.org/10.2139/ssrn.4597080. Küsters, Anselm, and Matthias Kullas. "Competition in Generative Artificial Intelligence." Centrum für Europäische Politik (19 March 2024): https://www.cep.eu/fileadmin/user_upload/cep.eu/Studien/cepInput_Generative_Artificial_Intelligence/cepInput_Competition_in_Generative_Artificial_Intelligence.pdf

[4] Seger, Elizabeth, Aviv Ovadya, Divya Siddarth, Ben Garfinkel, and Allan Dafoe. "Democratising AI: Multiple Meanings, Goals, and Methods." In Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society, 715–22. AIES '23. New York, NY, USA: Association for Computing Machinery, 2023. https://doi.org/10.1145/3600211.3604693.

empowering those affected by its development and use.[5] Access to high-quality and legally- and ethically-usable data, available as a public good, is a necessary condition for democratizing AI.

Today, we face two seemingly opposing challenges with regard to the data needed and used for training AI.
On the one hand, there is a shortage of publicly available, high-quality data sets. This shortage adds to existing power asymmetries in AI developments that favor commercial companies that either have access to their own proprietary data or are able to purchase such data.[6] The lack of access to data limits both research and other public interest activities and market competition from smaller actors. Without proper data, it is impossible to build AI-driven solutions that combat global challenges related, for example, to climate catastrophe or health risks.

On the other hand, publicly available data (including resources shared openly) have been managed and used in ways that are sometimes extractive, harmful, or – in some cases – even illegal. Our case study on the use of openly shared photographs of people to train facial recognition technologies shows a history, going back at least a decade, of AI training data sets being created and used without proper governance or consideration of social impact.[7] Recent news about LAION, a data set built from publicly available web content that is the cornerstone of much of image model training, has shown that it has not been properly cleaned and includes links to illegal content. Data scraping practices, including the amalgamation of data from various sources, pose multiple problems, ranging from privacy issues[8] to concerns about the rights of content creators.[9] Researchers have been auditing data sets and uncovering cases where these have not been properly curated or governed. The issues include leakages between training data and test data, unintended biases and behaviors of the trained models, and models of lower quality than anticipated.[10] In light of this, there is a need to establish data set governance frameworks that reduce harm and promote the responsible use of data.

[5] John W. Murphy and Randon R. Taylor, "To Democratize or Not to Democratize AI? That Is the Question," AI and Ethics 15 June 2023): https://doi.org/10.1007/s43681-023-00313-5.

[6] Ciuriak, Dan. "The Economics of Data: Implications for the Data-driven Economy." CIGI, 2018. https://www.cigionline.org/articles/economics-data-implications-data-driveneconomy/.

[7] Tarkowski, Alek, and Zuzanna Warso. "AI Commons. Filling the Governance Vacuum Related to the Use of Information Commons for AI Training." Open Future, 2023. https://openfuture.eu/publication/ai-commons/.

[8] Training data sets for AI combine data from different sources and may contain personal data, such as names, phone numbers, etc. Moreover, research has shown that this identifiable information may be sometimes extracted from the model. See e.g.: Nicholas Carlini et al., "Extracting Training Data from Diffusion Models," arXiv (30 January 2023): https://doi.org/10.48550/arXiv.2301.13188.

[9] For an analysis of the copyright challenges, see e.g.: Andrés Guadamuz, "A Scanner Darkly: Copyright Infringement in Artificial Intelligence Inputs and Outputs," SSRN Scholarly Paper (Rochester, NY, 26 February 2023): https://doi.org/10.2139/ssrn.4371204.

[10] Shayne Longpre et al., "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI," arXiv (4 November 2023): http://arxiv.org/abs/2310.16787.

THE PURPOSE OF THIS WHITE PAPER

This paper is based on the premise that a commons-based approach has the potential to provide wider availability of high-quality, diverse data sets while ensuring that rights are protected and that data is used in a fair and responsible way. The true democratization of AI requires more than just ensuring as broad as possible access to as much data as possible. The normative principles that we propose will assist in ensuring that this data is not just accessible but also properly governed.

This approach builds on the insight that the dichotomy between innovation and responsible development, between market growth and proper governance, is a false one. Conditions for responsible development and use do not stifle the kind of innovation that serves society and the planet. Commons-based approaches provide ways of balancing public interest, economic growth, and respect for fundamental rights. In other words, they offer a governance framework that balances data sharing with rules for protecting the interests of data subjects and creators and concerns over sustainability.

This white paper outlines a set of principles for governing data sets as a commons. These principles, tailored for the governance of AI data sets, build on our previous work on Data Commons Primer. The purpose of defining these principles is two-fold:

- We propose these principles as input into policy debates on data and AI governance. A commons-based approach can be introduced through regulatory means, funding and procurement rules, statements of principles, or through data sharing frameworks.
- These principles can also serve as a blueprint for the design of data sets that are governed and shared as a commons. To this end, we also provide practical examples of how these principles are being brought to life. Projects like Big Science or Common Voice have demonstrated that commons-based data sets can be successfully built.

AI DEVELOPMENT AND THE PARADOX OF OPEN

Commons-based approaches should be seen as a continuation of the vision for knowledge sharing, best expressed by the motto of Wikimedia: of a world where every single human being can freely share in the sum of all knowledge. It builds on the frameworks and governance mechanisms of open sharing by adding mechanisms that make it more resilient, equitable, and responsible.

Democratization and openness go hand in hand – democratizing AI means making AI technologies and resources available to a broader audience, ensuring transparency and accountability, facilitating collaboration, and allowing the free exchange of knowledge and data. However, beyond making AI accessible to a wide range of individuals and organizations, true democratization should also involve dismantling mechanisms inherent to AI development that perpetuate inequalities. The commons-based vision of data governance assumes that traditionally understood openness will not in isolation dismantle existing power imbalances in technology governance.

This is because of the Paradox of Open: open sharing of resources can both challenge and enable the concentration of power.[11] The last twenty years have shown that, without safeguards, openness can disproportionately serve the interests of parties with more economic power and resources. In the context of AI development, this is particularly true for companies that did best in the previous wave of digital innovation and enjoy significant economies of scale, to the detriment of parties that contribute to the commons but lack such power.

Incorporating data into AI training data sets and collections introduces an additional layer to the already complex socio-economic conditions of digital commons production and (re)use.
We have pointed out elsewhere that, in a situation where no sustainable and inclusive models for the development and maintenance of the digital commons can be developed, only the most privileged people or companies can afford to participate and reap benefits of the commons. In addition to limiting creativity and competition, this can lead to the erosion of the trust and collaboration necessary to sustain the digital commons, which would be devastating to shared resources in the long run.

Sustainable and just practices are, therefore, necessary to preserve the integrity and usefulness of data sets as they evolve.

The principles that we propose adhere to the spirit of sharing while allowing the development of a spectrum of approaches. In some cases, Open Data frameworks and other approaches (sometimes described as “open access commons”) remain the best way to create public value. Yet, in others, stronger forms of commons governance, gated access to resources, and additional mechanisms for ensuring responsible and equitable use should be introduced.

[11] Keller, Paul, and Alek Tarkowski. “The Paradox of Open.” Open Future, 2020. https://paradox.openfuture.eu.

A SPECTRUM OF APPROACHES TO DATA SET SHARING

When we talk about data and data sets in the context of AI training, it is important to keep in mind that data is not a homogeneous concept. Data comes in many different forms and is governed by different, sometimes overlapping, legal frameworks (which also vary widely internationally, although they are generally harmonized within the EU).

In the context of generative AI models, the term "training data" is most often used to refer to the data used to teach models how to classify data, recognize patterns, and make other decisions. Such collections often consist of copyrighted works and other protected subject matter. Other types of training data may consist of personal data (subject to personal data protection laws) and non-personal data such as industrial data, anonymized data, statistical and administrative data, and similar.

The principles proposed in this white paper can be applied to different types of data sets and a broad range of AI applications based on the use of machine learning technologies. At the same time, it's important to understand that there is a broad diversity of data sets and that there is, therefore, no "one size fits all" approach to data sharing.

Copyright and privacy or personal data rights are the two most important factors that determine the manner how a data set can be shared – how "open" it can be. There is a spectrum of openness of data sets running from content that is not subject to copyright, and does not include personal data, to highly sensitive data, such as health records. Commons-based governance offers a spectrum of approaches that are suited for the four typical scenarios for data sharing:

	OPEN	**GATED**
PERSONAL DATA	Common Voice	UK Biobank
NON-PERSONAL DATA	Wikimedia	HathiTrust

Four typical scenarios for data sharing, with examples.

- **Open sharing of personal data.** For example, Common Voice, a collection of datasets of voice recordings.
- **Open sharing of non-personal data.** For example, Wikipedia and its sister projects.
- **Gated sharing of personal data.** For example, UK Biobank, which makes biomedical data available for research to improve public health.
- **Gated sharing of non-personal data.** For example, HathiTrust, which makes library collections available for non-consumptive, research uses.

In each of these cases, the general principle holds true: that there is a need for high quality training data, and that commons-based approaches allow such data to be shared in a responsible manner. And in each case, the general principles proposed in this white paper need to be translated into specific governance mechanisms.

Similarly, the term AI is being used today to describe a broad range of technologies and their applications. In particular, it covers both so-called generative AI models (also described as general purpose or foundation models) and specialist systems that don't require such models.

While the principles are meant to be generally applicable, many of our examples relate to the development of generative AI models – and to related training data sets that consist of cultural or scientific works. This is due to the fact that these types of data sets and their uses trigger some of the key questions concerning the use of the commons for AI training.

WHY WE NEED COMMONS-BASED DATA SET GOVERNANCE FOR AI

Commons-based data set governance offers a means to address the challenges identified in the previous section. In the remainder of this paper, we propose principles and mechanisms based on the idea of the commons. This approach seeks an alternative to proprietary ownership of data and, in this way, hopes to address the problems of concentrations of power, extractive practices, and unequal access to resources.

Typically, resources that are managed as a commons are collectively or communally stewarded and governed. Treating data and data sets for AI training as commons – that is, as resources designed and managed in a collective or communal way, with established rules for access, sharing, and use – provides a framework for balancing the sharing of data on the one hand and the need to protect the rights and interests of the creators and subjects of the data on the other hand. It is also an approach that is meant to ensure a more sustainable approach to managing the data set.

The notion of a "commons-based data set" can be interpreted in two ways. One is that it is a data set derived from the Digital Commons, i.e., digital resources shared in the public interest, governed democratically, and collectively overseen. This is the case with Wikipedia – which in recent years became a key data source for AI training – or open-source software, which is a key component of many AI systems. The other meaning implies adherence to commons principles in managing the data set itself, even if the underlying data sources are not inherently part of the Digital Commons. This is the case with the Dolma data set created by the Allen Institute for AI. Although not all resources in the data set are openly available, the data set itself adheres to the principles of commons-based governance.

The commons-based governance model is not a silver bullet that will solve all the challenges posed by the use of increasingly large training data sets, but it is suited for three goals. They provide ways of stewarding access and ensuring a proper level of openness and conditions for sharing data. They create the opportunity for collective governance of the data set and for ensuring democratic control. And, finally, they enable public value to be generated.[12]

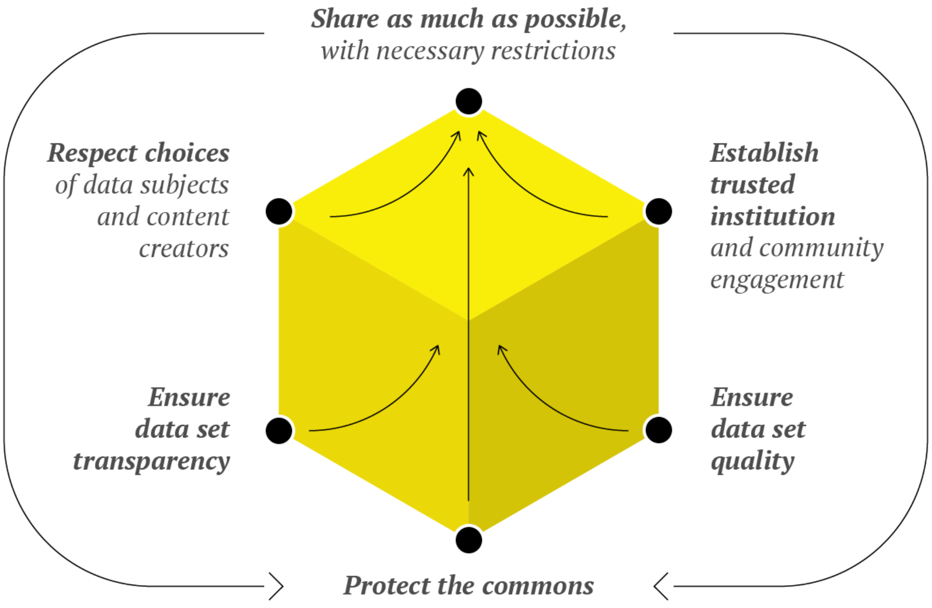
[12] We provide a more detailed description of the advantages of commons-based governance, and the goals that it helps achieve, in our Data Commons Primer: https://openfuture.eu/publication/data-commons-primer/.

SIX PRINCIPLES FOR COMMONS-BASED DATA SET GOVERNANCE

Commons-based data set governance is a spectrum of approaches aimed at facilitating access and use of data. It builds on over 15 years of work on Open Data sharing frameworks – which remain an important type of data governance also in the space of AI development. At the same time, it acknowledges a broader range of approaches to data governance, suitable in those cases where Open Data frameworks are not fit for purpose.

The growing adoption of the FAIR data principles, on which our framework builds, is an example of how more fine-grained approaches to data governance are being developed. Thinking in terms of a spectrum of approaches moves the data governance debate beyond a binary choice between data that is open and closed.

The Open Data Policy Lab, organized by the GovLab, argued in 2021 for a "Third Wave of Open Data" that prioritizes responsible use and data rights – and acknowledged a lack of governance frameworks that serve this goal.[13] This set of principles offers a blueprint for frameworks that can fill this gap.


Share as much as possible, *with necessary restrictions*
Respect choices *of data subjects and content creators*
Establish trusted institution *and community engagement*
Ensure data set transparency
Ensure data set quality
Protect the commons

[13] Young, Andrew, Andrew J. Zahuranec, and Stefaan Verhulst. "The Third Wave of Open Data." The GovLab, 2020. https://blog.thegovlab.org/the-third-wave-of-open-data.

Taken together, these principles define guidelines for sharing data and data sets, for ensuring that they are protected and managed responsibly, for ensuring transparency and respecting choices of data subjects, for ensuring data quality, and finally for proper institutional and participatory governance.

This is an overview of the six principles, which are explained in more detail in the following sections of this white paper.

1. **Share as much as possible while maintaining necessary restrictions**: The first principle emphasizes the importance of defining essential restrictions on openness to protect various interests. It is critical to assess what aspects would benefit from openness and what challenges might arise. The governance of a data set requires a balance between open access and necessary restrictions, often implemented through open licensing and other access control measures.
2. **Be transparent about the data and provide documentation**: Transparency about data sets, including thorough documentation of data sources and creation processes, is essential for informed discussions about AI and accountability in AI development. Standardized approaches to transparency, such as datasheets and data nutrition labels, facilitate information sharing and collaborative efforts to improve data sets. Without data transparency, it is difficult to monitor if shared resources are used in AI training.
3. **Respect the choices of data subjects and content creators**: Data governance should respect the decisions of individuals who contribute data or creative works. Legal frameworks and voluntary measures help to ensure that the decisions of data subjects and content creators are respected, striking a balance between openness and individual agency.
4. **Protect the commons**: Commons-based data set governance recognizes data sets as collectively owned and managed resources that serve both the community and the public interest. Mechanisms to protect the commons include consideration of working conditions, fair compensation, and mechanisms to ensure that value generated by the commons is returned.
5. **Ensure data set quality**: Data set quality is critical to maintaining them as reliable and inclusive resources. Attention to data set quality includes addressing bias, ensuring data sets are free of discriminatory elements, distinguishing between human-generated and synthetic content, and creating purpose-built data sets to mitigate risks associated with publicly available data.
6. **Establish trusted institutions and ensure community engagement**: Trusted institutions perform stewardship functions in the governance of commons-based data sets, ensuring proper management, access control, and fair treatment of contributors. Community engagement is essential for participatory governance, with identification of relevant communities critical for democratic oversight and decision-making processes.

Principle 1. Share as much as possible while maintaining necessary restrictions

This first principle, while preserving the spirit of sharing, emphasizes the importance of defining the essential restrictions on openness, taking into account the interests that must be protected. In governing any data set, it is essential to map benefits and aspects that are improved by making the data set open and available, and the challenges that might arise. The choice of specific mechanisms that either increase or restrict access and use of the data set should be based on this assessment. Potential issues arising from the unrestricted sharing of online resources include privacy violations, concerns about content creators' rights, and the fair treatment of contributors involved at various stages of data/content curation, as well as the sustainable and just maintenance of digital commons as such.

By recognizing and addressing the tension between the sharing of resources and the individual and collective rights related to these resources, a commons-based approach intends to preserve the status of resources as public goods while mitigating negative consequences that might arise from sharing. Building on this first principle, data and data set governance extend beyond the dichotomy of either making it openly accessible or fully restricting access.

In the field of generative AI, copyright is the central legal mechanism for regulating data access and use during AI training. A commons-based data set can be based on three broad categories of content: content that is shared under a free or open license, content in the Public Domain, or content that can be used under an exception or limitation to copyright.

Voluntary adoption of open licensing stands out as a key method for facilitating data availability. Copyright exceptions and limitations provide users with additional ways to use data. Data sets consisting of works in the Public Domain can be made available for AI training by anyone. By relying on these copyright-related mechanisms, data sets can be shared to the extent that this is possible within the boundaries of law. Additional measures – covered by the principles that follow – introduce necessary restrictions and ensure that the sharing is done fairly.

To this end, new mechanisms are being devised, with the aim of combining open licensing with more fine-grained ways of managing access. These include gated access mechanisms that enable registering requests for access to data sets, monitoring their uses, or assessing their impact. In order to access a gated model, the user needs to contact the data set owner and provide credentials. Access restrictions are often used to limit the users of a data set to a narrow, closed group. Gated access is compatible with open licensing and other sharing frameworks. It provides more fine-grained permission methods without excluding any users upfront.

Data sharing frameworks that need to take into account state regulation and differences between jurisdictions are a source of legal complexity. Data sets that include publicly available copyrighted works need to conform with copyright exceptions in the given jurisdiction. A variety of approaches to regulating AI training (and, more generally, machine learning, or text and data mining) through copyright law add a factor of complexity for data set governance. Under EU copyright law, as it relates to data mining and AI training, all lawfully accessible copyrighted works can be used for AI training without the explicit permission of their rights holders. For AI training in the context of academic research, this rule is absolute. In all other contexts, this rule applies unless the rights holders have made a (machine-readable) rights reservation.

Other jurisdictions, like Japan or Korea, also have strong exceptions for text and data mining. Yet, just like in Europe, there are ongoing policy debates on how these apply to AI training. And in some parts of the world, the situation is even less clear. In the United States, the fair use status of training AI works without the explicit permission of the rights holders has not been confirmed by the courts.[14]

EXAMPLES:

Gated access to Hugging Face data sets

Hugging Face is a platform for collaborative AI development that enables sharing of AI system components, including data sets, as well as collaboration on their development. Gated access to both data sets and to models is a key permission interface enabled on the platform. The Allen Institute for AI uses the gated access mechanism for sharing its Dolma data set on Hugging Face.

Findata and the Kapseli secure operating environment

Findata is the Finnish authority that hosts and makes available health data for secondary uses (for example, research). Users apply for data permits to access the data, which is then made available in aggregated form through Kapseli, a secure operating environment. Kapseli operationalizes the first principle of sharing as much data as possible while respecting the decisions of the data subjects. The system complies with the principle of minimization of personal data – i.e., to ensure that the processing of personal data is limited to the information that is directly relevant and necessary to achieve a specific purpose.

[14] Sag, Matthew. “Copyright Safety for Generative AI.” SSRN Scholarly Paper. Rochester, NY, December 3, 2023. https://doi.org/10.2139/ssrn.4438593.

Principle 2. Be transparent about the data and provide documentation

Transparency regarding data sets – which involves thoroughly documenting data sources and providing details about how the data was created, collected, refined, and annotated – contributes to informed discussions about AI and promotes accountability in AI development. It makes AI's capabilities and limitations easier to understand and helps stakeholders identify the biases of a model.

Today, major commercial models are released not only without publicly sharing the training data sets, but even without basic information about data provenance, data set composition, and characteristics. This applies both to models that are closed – such as Google's Gemini models – and to models that are openly shared, such as Meta's Llama models. An audit of 1800 data sets shows that there are diminishing efforts among AI developers to document the training data sets.[15]

While the degree of openness of a data set depends on the characteristics of the data set, the principle of data set transparency should be applied universally. In this sense, it is a condition of meaningful openness of AI, as it is a necessary measure allowing the explainability of AI models, especially in the pre-modelling stage. Transparency fosters collaborative efforts for the improvement and refinement of data sets.[16]

A range of tools are currently being developed to offer standardized approaches to transparency and accountability. In 2021, the "Datasheets for datasets"[17] paper proposed a standardized approach to data set transparency. Datasheets – such as the standardized data set cards required when sharing data sets on Hugging Face – encourage the machine learning community to prioritize transparency and accountability and facilitate the exchange of information between data set creators and data set users. Other initiatives are developing more advanced forms of transparency, through multi-disciplinary efforts to audit and trace the lineage of data sets systematically.[18]

Data set transparency is also needed to monitor whether and how digital resources shared in the public interest (the Digital Commons) are used in the development of AI models. More

[15] Longpre, Shayne, Robert Mahari, Anthony Chen, Naana Obeng-Marnu, Damien Sileo, William Brannon, Niklas Muennighoff, et al. "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI." arXiv, November 4, 2023. http://arxiv.org/abs/2310.16787.

[16] For an overview of why transparency is important, see e.g.: Rishi Bommasani et al., "The Foundation Model Transparency Index," arXiv (19 October 2023): http://arxiv.org/abs/2310.12941. The approach taken by the authors to assess the transparency of AI models has received some valuable critique, see: Nathan Lambert, SE Gyges, Stella Biderman, Aviya Skowron, "How the Foundation Model Transparency Index Distorts Transparency", EleutherAI (26 October 2023): https://blog.eleuther.ai/fmti-critique/.

[17] Timnit Gebru et al., "Datasheets for Datasets," arXiv (1 December 2021): http://arxiv.org/abs/1803.09010.

[18] Shayne Longpre et al., "The Data Provenance Initiative: A Large Scale Audit of Dataset Licensing & Attribution in AI," arXiv (4 November 2023): http://arxiv.org/abs/2310.16787.

generally, data metrics frameworks provide a means of tracking how data is being used. This is important both for monitoring and oversight, but also for understanding better the field of AI development and the significance of various data sets.

Transparency information itself can be considered a commons, as it is crucial for data set providers to share this information with all entities in the AI development lifecycle. This information should be shared in standardized, machine readable formats. The Data Provenance Initiative is an effort to aggregate transparency data.

EXAMPLES

Datasheet for the Pile
The Pile is a massive (825 GiB) text corpus. It was created by EleutherAI for large-scale language modeling efforts. EleutherAI published an extensive datasheet accompanying the data set. The document is intended to inform people interested in using the Pile for natural language processing. It includes information about the data sets in the stack, the motivation for creating the data set, how the data set has been used, what other tasks it could be used for, who funded the creation of the data set, and so on. The Pile data set has defined a standard of full transparency of training data.

The Data Nutrition Label
The Data Nutrition Project took inspiration from nutritional labels for food as it researched ways of ensuring deeper transparency of data sets. The "nutrition labels" describe the data set's "key ingredients," such as meta-data and populations, statistical distributions, or missing data. The project offers a tool for providing standardized labels based on a modular transparency framework that can be adjusted to the specifics of a given data set.

AI Act transparency obligations
The European AI Act, adopted in early 2024, introduces data transparency obligations. Providers of general-purpose AI models – for example, large language models – are obliged to provide information on data used for training. In addition, for systems considered high-risk, their developers need to provide more detailed datasheets.

Principle 3. Respect the decisions of data subjects and content creators

Respecting data subjects' and content creators' decisions should be a tenet of data set governance. This principle embodies the idea that people who contribute information or whose data is collected and used should have a say in whether their data or creative works become a part of a data set and part of the AI training environment.

In some jurisdictions, there are legal frameworks in place that address this issue. For example, the European Union has adopted laws that address personal data protection or that give owners of works the ability to opt out of text and data mining. While these rules provide some degree of legal certainty, their efficacy depends on their implementation and on the availability of technical standards. Commons-based data sets need to implement these rules in order to comply with the law in jurisdictions where such rules exist.

More generally, European data protection rules spell out that personal data can only be used with a proper legal basis. Article 6 of the General Data Protection Regulation sets out what these potential legal bases are, namely: consent, contract, legal obligation, vital interests, public task, or legitimate interests. For jurisdictions where such laws are not in place, voluntary measures introduced as part of data set governance offer a means of respecting the decisions of data subjects and content creators.

The principle is not absolute, and there exist circumstances where the requests of the data subject or rightsholder may not be the sole determining factor in whether the content can be processed. One notable example of when the use of data does not require consent pertains to exemptions for using data for research purposes.

The challenge is to strike a balance between the principle of openness and the respect owed to those who contribute potentially valuable data sources for AI. Ideally, the need to strike this balance should not be viewed as a binary choice but rather as a more fine-grained approach that acknowledges the multifaceted nature of user agency in determining the fate of the data and content to which they contribute. For example, someone might grant permission for their data to be used in one specific context but not in another. Therefore, besides legal and technical standards, additional guidelines (e.g., community standards) might be essential to protect individual or group agency. These rules should act as checkpoints, ensuring that the use of data in AI training adheres to the principles of consent and personal autonomy.

EXAMPLES

CARE principles

The CARE principles (collective benefit, authority to control, responsibility, ethics) are principles for Indigenous data governance. They were adopted by the Global Indigenous Data Alliance and emphasise the need to assert greater control over the application and use of Indigenous data and Indigenous knowledge for collective benefit. The purpose of these principles is to complement and balance the FAIR principles of data sharing that focus on the reusability of data.

Privacy Pledge

The Digital Data Commons Privacy Pledge is a set of standardized commitments that safeguard a data subject's privacy rights. The pledge was designed as part of DECODE, a European project aimed at creating tools that balance individual control over personal data and data sharing. These include a promise to respect privacy, data deletion rights, and limitations of purpose, among others. These Pledges can be used by entities using personal data in order to protect personal data and empower data subjects to share data.

Principle 4. Protect the commons

While the previous principle addressed the challenges of protecting individual decisions and user rights, there is also a need to protect the commons as a whole. Commons-based data set governance recognizes that a data set is a collectively owned and managed good that serves the needs of both the community that manages it and the broader public interest. This principle addresses the need to develop and use the commons in a sustainable way. It is also a principle that recognizes the relational19 and collective20 nature of data.

In the current landscape, there is a notable absence of safeguards ensuring that the use of open or publicly available data in AI applications is sustainable. On the contrary, extraction of value from the commons, without giving back to the commons and ensuring its sustainability, is a prevailing trend. Our study of early face recognition training data sets, built from openly licensed photographs of people, showed that there is at least a decade-long history of practices related to AI training that can be seen as a form of free-riding, often also with disregard for rules and norms set by providers of the data sets.

Advocating for openness and the various commons that comprise the Digital Commons requires a good understanding of the working conditions of those who are expected to contribute to the commons. This awareness is necessary to develop mechanisms and tools that encourage contributions without resorting to unfair practices or co-optation. The commons often depend on the contributions of free labor and volunteers. Sustainability of the data set, therefore, means, first of all, the need to consider the socioeconomic conditions surrounding its creation and maintenance. Appropriate support is needed for creators and stewards of a commons. Otherwise, participation in the governance of a commons becomes a luxury available only to those with the means and spare time.

Just as important are considerations of the working conditions of people hired to do work related to AI training. In recent years, data labeling work that is crucial to the development of commercial models has been outsourced to workers who have been poorly paid and not offered any work guarantees – in stark contrast to the conditions offered to the AI development teams. Commons-based approaches need to offer fair working conditions and wages to workers. While commons-based efforts often do not rely on paid labor, when they do, a standard for ensuring fair working conditions is a necessary part of this approach.

Moreover, there is a need to create mechanisms that ensure that value generated thanks to the commons is partially given back. Companies that use open or publicly available resources to create proprietary data sets should contribute back to the commons. Share Alike clauses

[19] Salomé Viljoen, "A Relational Theory of Data Governance," Yale Law Journal 131, no. 2 (November 2021): http://dx.doi.org/10.2139/ssrn.3727562.

[20] Singh, P.J. and Gurumurthy, A., "Economic Governance of Data: Balancing individualist-property approaches with a community rights framework," IT for Change: https://doi.org/10.2139/ssrn.3873141.

included in open licenses and other copyleft mechanisms have been used to establish some forms of reciprocity when using shared resources. These ensured that resources built with the use of the Digital Commons would also be shared. Unique ways in which data is used to train AI models mean that these clauses might not be valid further down the AI development lifecycle, and new mechanisms need to be devised. Furthermore, we need to look at other forms of reciprocity that go beyond copyright considerations and deal with fairness of labor or financial contributions to the commons. The right tools to operationalize this principle must strike a balance between openness and fairness.

Finally, the data sets used for AI training are part of a broader AI development ecosystem that has significant environmental impacts. In particular, generative, general-purpose AI systems require a significant amount of energy to complete the training phase. As of yet, there are no standardized, widely adopted approaches to report the energy consumption and carbon emissions of AI. However, there are growing efforts to address the environmental impact of AI through the development of reporting frameworks and guidelines.[21]

EXAMPLES

Wikimedia Enterprise

Wikimedia Enterprise is a paid service targeted at commercial users that provides access to the encyclopedia's content through an Enterprise API. This new commercial API is targeted at the biggest commercial platforms that, for years, have relied on Wikipedia as a key source of raw knowledge. Access fees provide a new, diversified revenue source for Wikimedia projects. And while Wikimedia Enterprise was not created specifically with AI training in mind, it is relevant also for these efforts.

FairWork

Fair Work evaluates commercial platforms based on five principles of fair work that address payments, working conditions, contracts, management, and representation of workers. These have been established collaboratively, with the participation of platform workers. In 2023, FairWork published a new set of principles targeted at AI companies and released its first rating of Sama, a data annotation company.

Mozilla Common Voice

Common Voice was a project joint project by Mozilla and the FAIR Forward initiative to develop data sets that could be used to build voice technology in low-resource languages: Kinyarwanda, Kiswahili, and Luganda. A data donation platform was established together with three language communities that brought together commercial, non-commercial, and government actors. Altogether, volunteers participating in the project provided almost 4,000 hours of audio recordings. These were then used to build machine learning models.

[21] Warso, Zuzanna. "Addressing AI Energy Consumption: Why the EU Must Embrace Ecodesign for Software | TechPolicy.Press." Tech Policy Press, September 13, 2023. https://techpolicy.press/addressing-ai-energy-consumption-why-the-eu-must-embrace-ecodesign-for-software.

Principle 5. Ensure the quality of the data set

This principle, building on the previous principle of protecting the commons, seeks to ensure the quality of the data sets. In other words, to manage the data set so that it remains a generative resource. Maintaining quality is critical to ensure that the commons continue to be a reliable and inclusive resource for AI development.

A consistent commitment to quality is an important aspect of protecting the commons. This entails taking a systematic approach to addressing biases in data sets and ensuring that they are free of discriminatory or illegal elements. Today, even major data sets – especially those created by scraping the public web – are sometimes created without proper attention given to the quality of the data.

Ensuring data set quality also means dealing with the specific risk of "pollution" associated with the outputs of generative AI and synthetic data. Research shows that using such outputs may compromise the quality and integrity of models built on their basis.22 In order to mitigate this risk, data set quality mechanisms need to distinguish between human-made and synthetic content and data.

Finally, better quality can also be achieved by creating purpose-built data sets aimed at solving specific problems. Today, many of the language models are trained with data sets built through web scraping, with all the inherent limitations and risks. Purpose-built data sets help address concerns over risks related to open or publicly available data. This approach can go hand in hand with data sets built through opt-ins and data donations.

Mitigating this risk involves setting up community rules, monitoring, and possibly some form of regulation that prohibits introducing AI-generated content into the pool of commons.

EXAMPLES

AI Act labeling rules

The European AI Act, adopted in early 2024, mandates the labeling of the outputs of generative AI models. This is meant to help combat "deepfakes" and other forms of disinformation created with the use of AI models. Labeling generative AI outputs will also help in managing data sets and ensuring their quality, by making it easier to distinguish synthetic content.

Content provenance and labeling

The Coalition for Content Provenance and Authenticity is an example of an effort to introduce voluntary labeling for content generated with generative AI systems. The Coalition was created by industry actors to agree on a technical standard for certifying the provenance of a work and

[22] Ilia Shumailov et al., "The Curse of Recursion: Training on Generated Data Makes Models Forget," arXiv (31 May 2023): https://doi.org/10.48550/arXiv.2305.17493.

its status as created by humans or AI-generated. Labeling initiatives are also undertaken by communities of creators. DeviantArt, one of the largest online art communities, has introduced in mid-2023 a requirement for users to label content created with generative AI tools.

Language Model Evaluation harness

Language Model Evaluation Harness is a set of over 60 freely available, standardized benchmarks managed by the EleutherAI nonprofit. Benchmarks are important tools for evaluating AI technologies, such as large language models. Evaluation frameworks that are shared and standardized are today a cornerstone of research on large language models (LLMs), as they provide a basis for comparing various models.

Principle 6. Establish a trusted institution and ensure community engagement

Commons-based data sets might require a trusted institution to fulfill various stewardship functions. The governance of data sets for AI training is often made more complex due to layered sharing frameworks in comparison to other data sources – such as, for example, Open Government Data or Open Access publications that adhere to more straightforward and well-established sharing models.

In recent years, various data intermediary models have been proposed in debates on data governance, including data trusts, data cooperatives, or data unions.[23] There are ongoing efforts to bring these intermediation frameworks to life. In some jurisdictions, these intermediaries have been recognized by law – for example, the European Data Governance Act defines and regulates data intermediaries. In each case, the intermediating institution is tasked with managing collective aspects of data governance.

There is a need to apply the concept of a trusted institution more broadly to data set sharing. Such institutions should manage access and use permissions, ensure participatory governance of the data set, and monitor uses. With regard to licensed content, the institution should enforce licensing conditions, resolve compatibility issues, and ensure the validity of the licensing information. The role of the institution is also to provide necessary resources (infrastructure, computing power, funding, human capacity) to properly design, build, and manage the data set. Finally, the institution should ensure that the commons are protected, both by providing fair treatment of the contributors and by negotiating with AI developers that use the data set.

The concept of a commons-based data set entails that the data sets are not solely managed by a trusted institution but also by a community in some capacity. This can be a community of contributors to the data set, or a community that holds rights – both individual and collective – to the data. In some cases, this community can be precisely defined. In others, the term should be understood more loosely as a group of people with a stake in the data set and how it is used. Identification of the relevant community is important for ensuring participatory, collective governance. The trusted institution should seek to ensure the participation of community members in decisions about the data set. In some cases, the institution works with the community to ensure both fair treatment of contributors and sustainability of their collective work.

The need for democratic control and participatory governance is closely related to the need to define a community that has a relationship with the data set. Forms of democratic control will vary, also depending on what groups of people have a relationship with the data set at different stages of the lifecycle.

[23] Duncan, Jamie. "Data Protection beyond Data Rights: Governing Data Production through Collective Intermediaries." *Internet Policy Review* 12, no. 3 (September 5, 2023). https://policyreview.info/articles/analysis/data-protection-beyond-data-rights.

EXAMPLES

Te Hiku Media

Te Hiku Media is a Māori nonprofit that is building a set of natural language processing tools to protect and revitalize the Māori language. The initiative has collected data donations by gathering voice recordings and used them to build a language model that can be used in educational apps. The initiative ensures that their technological solutions are sovereign and that the linguistic commons of their language and culture are protected. To this end, they have also designed a bespoke license that combines open sharing with communal control.

BLOOM language model

The BigScience Large Open-science Open-access Multilingual (BLOOM) language model was not only one of the first open-source LLMs, but also an example of collaborative development. The BigScience project that produced BLOOM was a one-year collaboration of over 1,000 experts. The project produced the ROOTS Corpus, a data set of 1.6TB of text covering 49 languages. The work was done through the collaborative efforts of hundreds of researchers, and issues of ethics, harm reduction, and governance were brought to the forefront. The ability to successfully create a massive, multilingual corpus was largely due to the collaborative nature of the effort, which brought together researchers with expertise in different languages. And in turn, the lack of contributors fluent in some languages was the main limitation to providing even greater linguistic diversity to the ROOTS data set and, thus, to the BLOOM model.

Wikimedia

Wikipedia, together with its sister projects, is the biggest source of knowledge, shared globally as a commons. The online encyclopedia is managed by the Wikimedia Foundation, which plays the role of the trusted institution. At the same time, Wikipedia is created and maintained thanks to the voluntary efforts of a global community of editors, administrators, and organizers. Community norms, codes of conduct, and collectively developed procedures and rules define the relationship between the foundation and the community of editors. In recent years, Wikipedia has proved to be one of the most important data sources for training language models.

ABOUT OPEN FUTURE

Open Future is a European think tank that develops new approaches to an open internet that maximize societal benefits of shared data, knowledge and culture.

Dr. Alek Tarkowski is the Strategy Director at Open Future. He holds a PhD in sociology from the Polish Academy of Science. He has over 15 years of experience with public interest advocacy, movement building, and research into the intersection of society, culture, and digital technologies.

Dr Zuzanna Warso is the Research Director at Open Future. She has over ten years of experience with human rights research and advocacy. In her work, she focuses on the intersection of science, technology, human rights, and ethics. She holds a Ph.D. in International Law.

ACKNOWLEDGMENTS

The authors would like to thank Susan Aaronson, Lila Bailey, Ramya Chandrasekhar, Paul Keller, Jan Krewer, Katja Mayer, Aditya Singh, Aviya Skowron, Derek Slater, Stefaan Verhulst, Adam Zable for valuable feedback and contributions to the report.