Data Governance in Open Source AI

Enabling Responsible and Systemic Access

Alek Tarkowski, Open Future

in partnership with the Open Source Initiative

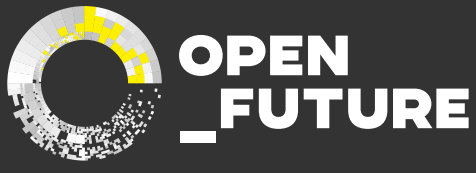


Table of Contents

Version 1.0 published January 22, 2025

Acknowledgments

This paper was co-designed with:

- Renata Avila, Open Knowledge Foundation
- Dr. Ignatius Ezeani, Lancaster University, UK, Nnamdi Azikiwe University, Nigeria, Masakhane NLP
- Ramya Chandrasekhar, CNRS Center for Internet and Society
- Maximilian Gantz, Mozilla Foundation
- Deshni Govender, GIZ FAIR Forward - AI for All
- Masayuki Hatta, Surugadai University
- Julie Hunter, LINAGORA
- Paul Keller, Open Future
- Stefano Maffulli, Open Source Initiative
- Ricardo Mirón, Digital Public Goods Alliance
- Kristina Podnar, Data & Trust Alliance
- Aviya Skowron, EleutherAI Institute
- Anna Tumadóttir, Creative Commons
- Joana Varon, Coding Rights
- Stefaan Verhulst, The GovLab, The Data Tank and New York University
- Thom Vaughan, Common Crawl
- Stefano Zacchiroli, Polytechnic Institute of Paris

Co-design process and workshop facilitated by Mer Joyce of Do Big Good.
Facilities for the workshop were kindly donated by LINAGORA.

ALFRED P. SLOAN FOUNDATION

This workshop was possible thanks to grant number 2024-22486 from the Alfred P. Sloan Foundation.

Executive Summary

As the Open Source Initiative convened its process to define Open Source AI, it became clear that organizations that care for open, fair and public-interest AI need to pay particular attention to and establish a shared position on data sharing and data governance.

Open Source Artificial Intelligence (AI) development presents an opportunity to democratize technological progress and reduce the concentration of power in the AI industry. However, its success depends heavily on the availability of high-quality, diverse datasets and robust data governance frameworks.

This white paper explores the intersection of data governance and Open Source AI, emphasizing the importance of responsible data sharing, community stewardship, and equitable practices in fostering innovation while protecting fundamental rights.

The broader challenges in data governance

Data is a critical resource for AI systems, yet its use is fraught with challenges. We face a paradox when it comes to data availability. On the one hand, data is abundant — best demonstrated by the fact that the entire open web's content is foundational to most generative models developed in recent years. On the other hand, it is scarce, as evidenced by those same models, for which access to proprietary, restricted data provides an advantage.

Publicly available datasets, such as those derived from web scraping, have historically supported AI advancements, but they also raise ethical concerns about privacy, consent, and data ownership. While vast amounts of data are accessible, much of it is proprietary, poorly curated, or unrepresentative of global diversity.

In this context, Open Source is the ideal way to create equitable and transparent AI systems. The Open Source Initiative (OSI) spearheaded efforts to understand openness for AI through the Open Source AI Definition (OSAID). However, the OSAID process revealed that more focus is needed on data governance, addressing the ethical and legal complexities of data sharing.

Critical challenges in data governance and AI development

1. **Data governance and ethical use:** Effective data governance must balance the need for open sharing with the protection of intellectual property, privacy, and community rights. Without such frameworks, there is a risk of exploitation, particularly in the Global South, where data extraction can reinforce systemic inequities.
2. **Openness standards and transparency:** The current definition of Open Source AI emphasize transparency, including clear documentation of data provenance, licensing, and removal of use restrictions. Yet, many models labeled "open" or "open source" even, lack full compliance with these principles, leading to confusion.
3. **Structural biases in data:** Many datasets used in AI development reflect biases based on language, geography, and socioeconomic status, resulting in AI systems that inadequately represent marginalized communities. This perpetuates global digital inequities and limits the inclusivity of AI solutions.
4. **Environmental sustainability:** The resource-intensive nature of AI development raises concerns about its environmental impact. Open data sharing initiatives can mitigate this by reducing redundant data collection and fostering more efficient AI training practices.
5. **Stakeholder representation:** The current AI ecosystem often prioritizes the needs of developers and large corporations over other stakeholders, such as data contributors, affected communities, and public-interest organizations. Bridging this gap requires inclusive governance models and collaborative approaches to data stewardship.

Strategies to Sustain Open Source AI

To address these challenges, the white paper identifies two key paradigm shifts:

1. **Adopting a data commons approach:** Moving beyond open data frameworks to broader data commons governance, which includes diverse forms of data sharing while protecting rights and ensuring equitable use. This approach acknowledges the varied nature of data, from fully open to restricted datasets, and promotes innovative licensing models, such as data trusts and cooperatives.
2. **Expanding the stakeholder universe:** Engaging a broader range of stakeholders in the AI lifecycle, including content stewards, data custodians, and impacted communities. By fostering partnerships between AI developers and these groups, new datasets can be responsibly created, curated, and shared.

Focus areas for action

The paper outlines six critical focus areas to advance data governance and Open Source AI:

1. **Data preparation and provenance:** Establishing robust standards for data collection, classification, anonymization, and metadata to ensure quality and traceability.
2. **Preference signaling and licensing:** Developing mechanisms like opt-out frameworks and social licenses to allow rights holders and communities to control data use.
3. **Data stewards and custodians:** Strengthening roles for data stewardship, including intermediary institutions that facilitate data sharing while ensuring ethical governance.
4. **Environmental sustainability:** Promoting practices that reduce the environmental impact of AI through shared datasets and efficient training methods.
5. **Reciprocity and compensation:** Implementing mechanisms that ensure value generated from shared data is equitably distributed, particularly to marginalized communities.
6. **Policy interventions:** Advocating for public policies that mandate data transparency, incentivize data sharing, and support the creation of open datasets.

Taken together, work in these various focus areas serves two goals. First, it serves the purpose of increased data sharing, by making various types of data easier to use, by increasing the quality of datasets and by ensuring that more data is available openly. Second, it protects knowledge commons by acknowledging a broad range of social aspects of data generation and associated legal frictions and deploying mechanisms other than licenses to offer adequate governance.

A path for better Open Source AI

Open Source AI has the potential to drive innovation, enhance transparency, and promote equity in the AI landscape. Achieving this vision requires a shift from quantity-driven data practices to a quality- and governance-focused approach. By adopting data commons frameworks, expanding stakeholder engagement, and addressing key governance challenges, the Open Source AI community can foster a more inclusive and sustainable AI ecosystem.

This white paper calls for collective action among developers, policymakers, and civil society organizations to establish shared standards and implement solutions that balance open sharing with responsible governance. Through these efforts, Open Source AI can deliver on its promise of serving the public good while respecting the rights and interests of all stakeholders.

Stefano Maffulli
Executive Director - Open Source Initiative

Introduction

For over two years, the Open Source Initiative (OSI) has convened a global, multi-stakeholder process to define Open Source AI, which resulted in the release of version 1.0 of the Open Source AI Definition (OSAID). The goal is to define those artificial intelligence (AI) systems that provide to various stakeholders (users, AI developers, practitioners, etc.) the equivalent of the freedoms that Open Source software offers. Throughout this process, it became clear that organizations that care for open, fair and public-interest AI need to pay particular attention to and establish a shared position on data sharing and data governance.

Data governance is understood in this paper as coordinated actions of different actors, using different instruments, methods and strategies that, taken together, create rules and norms for data. We use the term "governance" to cover these various means, most importantly including legal frameworks as well as standards or social norms. All these different factors impact the uses of data and determine what is permitted, required or prohibited by different actors. Various forms of data governance are crucial for ensuring that data is shared and that this is done fairly and responsibly.

Undertaking effective data governance in AI training has become a central issue among Open Source AI advocates due to the complex nature and diversification of shared data resources. To have effective governance requires navigating a matrix of rules and ethical principles to protect copyright, privacy and other data-related rights. Effective data governance must also consider the impact on fundamental rights, ensuring that equitable and inclusive standards are maintained within AI ecosystems and the oversight of such systems.

To address this challenge, the Open Source Initiative and Open Future organized a workshop, held on October 10 – 11, 2024 in Paris at LINAGORA's office, to address the combined challenge of ensuring data sharing and proper data governance. The workshop convened a small group of experts representing various stakeholder groups, various perspectives on data and AI, and various regions of the world. The point of reference for this workshop is an approach to data in AI systems that is proposed in the Open Source AI definition. Building on this basis, workshop participants considered what other actions can be taken to ensure that more data is being shared and that it is shared responsibly — for AI training and other uses. This entails both releasing more resources as Open Data and deploying other, commons-based sharing frameworks for those types of data and datasets that cannot be shared openly.

This white paper aims to explain to OSI constituencies and stakeholders the role of data governance and data sharing as they relate to Open Source AI development. It also offers specific approaches and strategies that can be undertaken both by Open Source AI developers and by other stakeholders to increase the amount of data that is properly governed and available for use in Open Source AI development. The ultimate goal is to support a future where data empowers various communities that create and own the data and not just corporations.

Throughout the paper, we understand an AI system, in line with the OECD definition adopted in the Open Source AI definition, as:

> *a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. Different AI systems vary in their levels of autonomy and adaptiveness after deployment.*[1]

Admittedly, this definition covers a broad range of systems, including generative AI and automated decision-making systems, foundation and small models, and AI systems trained on various types of data, for various purposes. In our conversations, we intentionally kept the scope broad, although at points the conversation focused on generative AI models and the data needed to train them.

NOTES

1 Open Source Initiative, "Open Source AI Definition." https://opensource.org/ai/open-source-ai-definition.

Data and training AI systems: the state of play

Data is both a key resource necessary to successfully develop AI systems and a significant policy topic, as efforts are undertaken globally to address the challenges of data sharing and data governance. The term data encompasses a broad range of resources, and different types of data are needed in various systems and in various stages of their development and use.[2]

Since at least the beginning of this century, training datasets have been developed alongside machine learning systems, including modern generative AI models. (And they have a much longer "prehistory," at least since 1973 when the "Lenna" standard test image began being used in image processing.) This is a history of collaboration and of sharing resources in the spirit of open science and Open Source development. Yet it is also a history of resource exploitation, disregard for existing norms and rules, and an inability to establish its own governance standards — as the story of the Lenna test image reminds us.

Equally important, everything available in digital form is now potential training data for AI systems. The canonical large language model is trained on the "raw internet" — content crawled and scraped from the publicly available web.[3] Web scraping for AI can be seen as another expression of the open web's optimistic ethos and one more proof of innovation enabled by the web. However, it can also be viewed as a new type of extraction and exploitation, at a massive scale, of global knowledge and cultural commons.[4] More generally, the use of data at a massive scale to train AI systems is seen by some as an expression of the opportunities for open data sharing, but seen by others as a new form of exploitation and misuse of the knowledge commons. The latter concerns lead to limited or non-availability of certain data sets and the emergence of a data winter, marked by decreased interest in seeing data as a resource that can be leveraged for the common good.[5]

Related to this are growing concerns that such use of data is a new form of digital colonialism.[6] This is because most data in digital form exists in developed nations and is not created in the Global South. This results in, inter alia, unequal control over and ownership of collected data, lack of relevance and representation in datasets, economic exploitation of Global South nations, loss of autonomy and inability to influence or control citizen rights, and a dependence on external technology produced in developed (global minority) nations. Generative AI models therefore have structural biases, which is especially problematic when these systems are deployed in Global South countries, displacing and misinforming local knowledge, culture and world views.

Today, we face a paradox when it comes to data availability. On the one hand, data is abundant — best demonstrated by the fact that the entire open web's content is foundational to most generative models developed in recent years. On the other hand, it is scarce, as evidenced by those same models, for which access to proprietary, restricted data provides an enormous market advantage. In addition, access to a seeming wealth of online data hides the fact that the web is not representative of the world's knowledge, languages and cultures. Gaps are structural and align with global divides and inequalities, especially between the Global North and the Global South.

The availability of data also does not necessarily translate into data quality, and purposeful AI development should focus on selecting the right data. Just because all of the web is available as a

source of AI training does not mean that the whole web should be necessarily crawled to obtain such training data, or that the effects will be positive. Data that is publicly available cannot be always legally scraped and used, largely due to intellectual property rights, various personal data protection laws and international human rights standards on the right to privacy.[7] Research also shows that scaling datasets leads to scaling of harmful content in these datasets as well.[8]

AI systems and openness

The "story arc" of open and closed AI systems is now a well-known narrative that begins with a state where many AI systems and components are open. Emblematically, the largest technology companies released key building blocks for AI, such as the PyTorch and TensorFlow libraries, as Open Source code. Similarly, OpenAI's first publicly released language model, GPT-2, was shared in 2019 on GitHub under a modified MIT license. Just three years later, in 2022, OpenAI made available its next model, GPT-3, only through a gated API and a user-facing chat interface. Yet, concurrently, open models were also being released under more liberal terms: GPT-NeoX, BLOOM and Stable Diffusion are just a few key examples.

Presently, two years later, the landscape is mixed. Most of the popular models, shared by big AI companies, are closed. On the other side of the spectrum, there is a growing ecosystem of models that aspire to meet a strong standard of openness, such as Olmo from the Allen Institute for AI, GPT-Neo from Eleuther AI, Bloom from the Big Science project, the open science model Aurora and multiple national initiatives, such as Swedish GPT-SW3, Singaporean Sea Lion, Polish Bielik or Spanish Ăguila.

In between, releases fall on a spectrum of approaches that are typically considered as not canonically open, but also not closed.[9] This is usually due either to the use of a non-standard license that is not fully compliant with existing standards (such as the Open Source Definition) or to the fact that not all key components are being openly shared. As a result, the term "Open Source" has been used to describe models with various levels of openness, many of which should more precisely be described as "open weight" models.[10]

Among the Big AI companies, attitudes towards openness vary. Some, like OpenAI or Anthropic, do not release any of their models openly. Others, like Meta, Mistral or Google, release some of their models. These models — for example, Llama, Mistral or Gemma — are typically shared as open weights models.[11]

The ecosystem of AI solutions released in various ways that broadly fit the category of open builds on components shared openly by various actors. Open weight releases shared by companies like Meta or Mistral have played a key role in scaling this ecosystem, with much innovation happening through the reuse of these technologies.[12]

These open alternatives are seen as one of the solutions to the growing challenge of concentrations of power in AI. While initially this challenge was not framed as one of the key AI risks, there is increasing awareness of this issue.[13,14,15] Seen from a Global South perspective, these concentrations are even more acute.[16]

There is currently no consensus that open AI systems will become viable alternatives to products offered by the AI incumbents. Specific ways in which various components are shared and made

transparent determine whether open solutions can indeed combat these concentrations of power.[17] It remains to be seen to what extent open AI technologies will be deployed in ways that are more self-sovereign and sustainable and whether they translate to greater market competition. The growing role of Llama as a foundation for an ecosystem that is not truly open makes some afraid of another "embrace, extend, extinguish" scenario.[18] Critics believe that the aim of these business strategies is mainly to bolster positions in the face of regulation and to entrench their dominance.[19]

There also remains uncertainty about the standard for open, or more precisely, Open Source models. The OSAID definition and other similar initiatives, such as the Linux Foundation's Model Openness Framework[20] and the Digital Public Goods Alliance's work on a standard for AI as a digital public good,[21] are addressing this issue.

The overall outlook is positive, with a growing ecosystem that continues to innovate and is working towards more openness and towards defining shared standards for openness in AI.

AI systems and data

The narrative surrounding AI training datasets and the diverse forms of data (both non-personal and personal), content and information comprising them is considerably more somber. While the development of AI models feels like an open-ended endeavor, AI development teams are assuming that data is a finite resource and we might soon reach a "peak human data" moment. Researchers from Epoch.ai argue that publicly available text data will become insufficient for training LLM sometime between 2026 – 2032, as models scale and require more training data.[22] Strictly speaking, this isn't entirely accurate — there remains a wealth of resources in the world that has not been harnessed for AI training. These resources are not easily accessible either because they have not been digitized, or because they are proprietary and not shared by their owners. As a result, training datasets are often homogenous and display bias concerning language, geography or ethnicity.[23]

There have been attempts to fill this gap with synthetic data and Microsoft's Phi models are the best example of such an approach.[24] At the same time, researchers are arguing about the risk of catastrophic "model collapse," as models train recurrently on their own content. And the very idea of "peak data," (and the need to continuously expand the size of training datasets) is a symptom of the inherent limitations of the current paradigm of building large-scale AI systems.

When it comes to data, it appears that AI developers are today not innovating but rather scavenging for what is already at hand. This perception is intensified by the fact that the utilization of humanity's digital resources for training AI models, on a global scale, is frequently seen as exploitative. This is due to extreme concentrations of power, with just a few actors, at a global scale, having the capacity to fully benefit from these resources — and also to enclose them, in yet another enclosure of the digital commons. At a global scale, improper or insufficient data governance leads to new forms of neo-colonialist data extraction.[25] This is exacerbated by practices of some actors, which have outsourced the labeling of AI training data to freelance workers in Global South countries, often on platforms that do not comply with basic fair work standards.[26]

The crawling and scraping of much of the public web by AI developers was perceived by many as a move that if not unprecedented — search engines have been crawling the web for decades — then certainly exacerbated the challenges and risks associated with the "platformization"[27] of the web. This negative view was further amplified by the fact that the training of models on web-scraped

data is often seen as not properly managed: conducted in ways that are perceived in some cases as outright unlawful and in others as at least morally questionable, not in line with research ethics, or unjust.[28]

Collections of content that are either in the Public Domain or openly licensed are another crucial resource that is being utilized and potentially exploited. A range of recent efforts to build such collections, intentionally designed for AI training, include the PD12M dataset of image-caption pairs or the Common Corpus text dataset. These intentional efforts are in stark opposition to an approach described by one of the participants of our convening as a "You Only Live Once (YOLO) data strategy," where as much data as possible is scraped and acquired in other ways, without caring for intellectual property, privacy and other rights. The relatively small volume of these open collections puts at a disadvantage actors who want to work with fully open and transparent datasets.

Just as is the case with web scraping, there is a perception that it's not merely the various collections, databases and repositories that are at risk of being exploited: the scale of use implies that the digital commons as a whole are being leveraged for AI training.[29] There is a persistent feeling that these uses might not adhere to either the letter of the law (as expressed through open licenses, or based on jurisdiction laws like fair use) or community norms.

Open Future's study of ways in which openly licensed photographs of people were repackaged and used as machine-vision datasets demonstrates that the issues are not new but have persisted for over a decade.[30] YFCC100M, the first such dataset, turned ten years old in 2024. The emergence of AI systems and the growing awareness of how they were trained on the digital commons are causing some to re-evaluate the value of open sharing.[31] Today's conversations about responsible licensing and preference signaling are indicative of this shift in norms around sharing. Efforts seem to focus less on ensuring openness and more on shoring up open resources against AI-related exploitation. A recent study by the Data Provenance Initiative shows a significant increase in entries that block AI crawlers in robots.txt files.[32] It is the most clear sign of this shift of attitude towards open sharing.

Moreover, there is a perception that AI developers have not been paying adequate attention to data governance. This encompasses various forms of disregard or a lax attitude toward licensing and intellectual property rules,[33] dataset quality and curation,[34] and transparency measures — to name just a few key concerns. A growing list of such issues, which surface as researchers investigate various datasets, suggests a pattern of disregard for data governance by researchers, engineers, developers and business people as they rush to build ever larger models and release new products. Admittedly, some of the Big AI companies are undertaking efforts to solve these problems. For example, the Data and Trust Alliance has been developing data provenance standards to address these issues.

NOTES

2 For a more detailed analysis of various types of data and governance requirements related to them, see: Open Data Policy Lab. "A Fourth Wave of Open Data? Exploring the Spectrum of Scenarios for Open Data and Generative AI." https://www.genai.opendatapolicylab.org/.

3 Typically, the term "web scraping" is used to describe data collection methods that bypass access rules, while "web crawling" denotes more responsible practices.

4 Pierce, David. "The text file that runs the internet." The Verge, 14 February 2024. https://www.theverge.com/24067997/robots-txt-ai-text-file-web-crawlers-spiders.

5 Verhulst, Stefaan G. "Are We Entering a 'Data Winter'?" Medium (blog), 23 January 2024. https://sverhulst.medium.com/are-we-entering-a-data-winter-f654eb8e8663.

6 Avila Pinto, Renata. "Digital sovereignty or digital colonialism?" Sur. https://sur.conectas.org/en/digital-sovereignty-or-digital-colonialism/.

7 See for instance privacy considerations for using social media data as a source of intelligence: Privacy International. "Social Media Intelligence." https://privacyinternational.org/explainer/55/social-media-intelligence.

8 Birhane, Abeba, Vinay Uday Prabhu, and Emmanuel Kahembwe. "Multimodal Datasets: Misogyny, Pornography, and Malignant Stereotypes." arXiv, 5 October 2021. https://doi.org/10.48550/arXiv.2110.01963.

9 Solaiman, Irene. "The Gradient of Generative AI Release: Methods and Considerations." arXiv, 5 Feb. 2023. https://doi.org/10.48550/arXiv.2302.04844.

10 Liesenfeld, Andreas, and Mark Dingemanse. "Rethinking Open Source Generative AI: Open-Washing and the EU AI Act." Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, 2024, pp. 1774–87. ACM Digital Library. https://doi.org/10.1145/3630106.3659005.

11 "On Open-Weights Foundation Models." Federal Trade Commission, 10 July 2024. https://www.ftc.gov/policy/advocacy-research/tech-at-ftc/2024/07/open-weights-foundation-models.

12 Osborne, Cailean, et al. "The AI Community Building the Future? A Quantitative Analysis of Development Activity on Hugging Face Hub." arXiv, 5 June 2024. https://doi.org/10.48550/arXiv.2405.13058.

13 "CMA AI strategic update." Competition & Markets Authority. https://www.gov.uk/government/publications/cma-ai-strategic-update/cma-ai-strategic-update.

14 Küsters, Anselm, and Matthias Kullas. "Competition in Generative Artificial Intelligence." Centrum fur Europaische Politik. https://www.cep.eu/fileadmin/user_upload/cep.eu/Studien/cepInput_Generative_Artificial_Intelligence/cepInput_Competition_in_Generative_Artificial_Intelligence.pdf.

15 Vipra, Jai, and Sarah Myers West. "Computational Power and AI." AI Now Institute. https://ainowinstitute.org/publication/policy/compute-and-ai.

16 Gimpel, Lea, et al. "Democratizing AI for the Public Good: Key Concepts and Recommendations." T20 Policy Brief. https://www.t20brasil.org/media/documentos/arquivos/TF05_ST_05_Democratizing_AI_fo66d5d70141505.pdf.

17 Tiwari, Udbhav. "The Openness Imperative: Charting a Path for Public AI." AI Now Institute. https://ainowinstitute.org/publication/ix-the-openness-imperative-charting-a-path-for-public-ai.

18 "Embrace, Extend, and Extinguish." Wikipedia, 30 Dec. 2024. https://en.wikipedia.org/w/index.php?title=Embrace,_extend,_and_extinguish&oldid=1266186280.

19 Widder, David Gray, et al. "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI." Social Science Research Network. https://doi.org/10.2139/ssrn.4543807.

20 White, Matt, et al. "The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence." arXiv, 18 Oct. 2024. arXiv.org. https://doi.org/10.48550/arXiv.2403.13784.

NOTES

21 "Core Considerations for Exploring AI Systems as Digital Public Goods," Digital Public Goods Alliance. https://www.digitalpublicgoods.net/AI-CoP-Discussion-Paper.pdf.

22 Villalobos, Pablo. "Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data." Epoch AI, 6 June 2024. https://epoch.ai/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

23 See, for example, https://arxiv.org/abs/2402.02680v2 on geographic bias, https://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf on race and gender bias in face recognition datasets, and https://dl.acm.org/doi/pdf/10.1145/3442188.3445922 on linguistic bias.

24 Gunasekar, Suriya, et al. "Textbooks Are All You Need." arXiv:2306.11644. arXiv, 2 Oct. 2023. https://doi.org/10.48550/arXiv.2306.11644.

25 Muldoon, James, and Boxi A. Wu. "Artificial Intelligence in the Colonial Matrix of Power." Philos. Technol. 36, 80 (2023). https://doi.org/10.1007/s13347-023-00687-8.

26 OII | Fairwork AI report: Improving employment conditions for invisible internet workers https://www.oii.ox.ac.uk/news-events/fairwork-ai-report-improving-employment-conditions-for-invisible-internet-workers/.

27 For a definition of platformization, see: Poell, Thomas, et al. "Platformisation." Internet Policy Review, vol. 8, no. 4, Nov. 2019. https://policyreview.info/concepts/platformisation.

28 Buschek, Christo, and Jer Thorp. "Models All The Way Down." Knowing Machines. https://knowingmachines.org/models-all-the-way.

29 Keller, Paul. "AI, the Commons, and the Limits of Copyright." Open Future. https://openfuture.eu/blog/ai-the-commons-and-the-limits-of-copyright.

30 Tarkowski, Alek, and Zuzanna Warso. "AI Commons. Filling the governance vacuum on the use of information commons for AI training." Open Future. https://openfuture.eu/publication/ai-commons/.

31 Hong, Shannon Y., and Alek Tarkowski. "Alignment Assembly on AI and the Commons. Outcomes and learnings." Open Future. https://openfuture.eu/publication/alignment-assembly-on-ai-and-the-commons-outcomes-and-learnings/.

32 Longpre, Shayne, et al. "Consent in Crisis: The Rapid Decline of the AI Data Commons." arXiv, 24 July 2024. arXiv.org. https://doi.org/10.48550/arXiv.2407.14933.

33 Longpre, Shayne, et al. "A Large-Scale Audit of Dataset Licensing and Attribution in AI." Nature Machine Intelligence, vol. 6, no. 8, Aug. 2024, pp. 975–87. https://doi.org/10.1038/s42256-024-00878-8.

34 See for example: Crawford, Kate, and Trevor Paglen. "Excavating AI: the politics of images in machine learning training sets." PhilPapers. https://philpapers.org/rec/CRAEAT-5; and Tarkowski, Alek, and Zuzanna Warso. "AI Commons. Filling the governance vacuum on the use of information commons for AI training." Open Future. https://openfuture.eu/publication/ai-commons/.

The challenge: the openness of datasets and Open Source AI development

Over the last two years, several initiatives have been aimed at establishing a standard for openness of AI systems. These include the Linux Foundation's Model Openness Framework (MOF), the Digital Public Goods Alliance's standard for AI as a digital public good, and the Mozilla's Convening on openness in Artificial Intelligence. All of these initiatives had to start with conceptualizing what is an AI system and what are its identifying components — so that a standard of openness for these various components could be defined. All these frameworks list data as one of the key components of AI models or AI systems. As a result, considerations of openness of AI models must address the gnarly question of the openness of data: to what extent and in what way does openness of the training data (or lack thereof) determine the openness of the overall AI system?

There is still a lack of agreement as to the correct approach, although consensus is emerging. First, there is broad agreement that some form of transparency of data is a precondition for openness of the overall system: visibility into the data's source and provenance (e.g., where is the data coming from, how was it generated, when and by whom), rights (who owns the intellectual property to determine the use of and monetization of the data) and restrictions (e.g., what are the applicable data privacy restrictions, including purpose of use, localization and geographic processing requirements). The OSAID requires that the data information be shared. Similarly, the MOF framework of the Data Card is a basic component required in all systems, and the Mozilla framework considers data documentation a key attribute of an open model.

Secondly, the abovementioned issue regarding the openness of datasets remains unresolved. Since data is considered a key component of AI models, it would seem that openness of data is an obvious requirement for the overall model/system to be considered open. Supporters of this requirement point to the clear-cut definitions of open data as a reference point. They also argue that the availability of training data (and not just information about the data) is necessary for auditing, verifying and replicating open models.

The counterargument to this is based on considerations of factors that limit openness of data, or make it impossible: compliance with regulations (taking into account varied laws across jurisdictions), consent — in particular concerning personal data, legal and ethical considerations, or the worry of communities that openness without restrictions might lead to privatization of their data by third parties.

As a result, draft AI openness frameworks consider openness of data as an optional, aspirational goal. The MOF includes open datasets in the most complete class of AI systems, called "open science" — but acknowledges that "Open Model" and "Open Tooling" classes can meet the standard without making datasets openly available.

Critics of such graded approaches, which make full openness of data desirable but optional, argue that it runs the risk of supporting open washing — in the sense that under this definition models criticized until now as misusing the term "Open Source" will fit the definition. For example, Google's Gemma — often described as an "open model," as no training data is shared, would potentially fall under the definition.

This debate puts aside the fact that some commercial models – Llama, most prominently – that are presented as open also do not meet other requirements. For example, their licenses introduce limitations that make them non-compliant with open licensing standards, and they fail to share other system components, such as training code.

The importance of data transparency and access to data, even if the latter is contested as a requirement for Open Source AI systems, signals the need for not just more data to be shared, but also for better data governance.

Such data governance also needs to navigate the risk that open data generated by and for communities could be opportunistically exploited by powerful third parties. Hence, many community driven AI data collections end up in a dilemma: Protecting openness and respecting the data rights of marginalized communities will limit the general ability to grow a global pool of Open Source AI. Paradoxically, by trying to avoid the freeriding of some, everyone might end up with less Open Source AI which can be used by anyone, including vulnerable and marginalized communities.[35]

Addressing this issue requires taking into account the complexity of data as a resource. At its most basic, various data governance models are needed for four different types of data:

- **Open data:** data that is freely accessible, usable and shareable without restrictions, typically under an open license or in the Public Domain[36] (for example, OpenStreetMap data);
- **Public data:** data that is accessible to anyone without authentication or special permissions (for example, Common Crawl data). Note that this data can degrade as web content becomes unavailable;
- **Obtainable data:** data that can be obtained or acquired through specific actions, such as licensing deals, subscriptions or permissions (for example, ImageNet data);
- **Unshareable non-public data:** data that is confidential or protected by privacy laws, agreements or proprietary rights and cannot be legally shared or publicly distributed.

NOTES

35 "Open-Source AI Data Sharing: yes! Data Colonialism: no!" Open for Good Alliance. https://medium.com/@openforgood/open-source-ai-data-sharing-yes-data-colonialism-no-3062a922de03.

36 For a more detailed definition, see: "What is Open Data?" Open Data Handbook. https://opendatahandbook.org/guide/en/what-is-open-data/.

Problem definition

The development of Open Source AI systems, which can compete with large proprietary solutions that dominate the market and thus address the risk of adverse concentrations of power, is limited today by insufficient availability of data that can be used to train those systems. Key causes for this state of things include a lack of incentives for data sharing, a lack of public investment and a lack of mechanisms for controlling how data is reused.

Of the four types of data outlined in this white paper, the convening that we organized focused on conditions for making available and sharing open and public data. Less attention was paid specifically to obtainable and unshareable data, although proposed data governance solutions can be adapted to these types of data as well. The overall assumption was that open and public data are the only ones available as shared resources compared to other types.

At the same time, data sharing efforts should not be seen simply as focused on releasing as much data as possible, as openly as possible. Instead, proper data preparation, data governance frameworks and stewardship functions should be the starting point for any AI training dataset. Open sharing or other forms of making datasets obtainable are then treated as results of proper data governance.

Greater care for data governance and acknowledgment that it should be the foundation of AI governance[37] will help alleviate concerns over value extraction and exploitation related to data use. This is important in today's zeitgeist, where a shift in attitudes toward sharing data and content openly is occurring. The empirical data from the "Consent in crisis" study shows that in the last few years, for the first time, a significant closure of web content is visible in robots.txt files, as owners of various public resources introduce limitations in reaction to the constant growth in scale of web crawling and web scraping.[38] While the study examined only a specific form of sharing — that of web content, facilitated by robots.txt permissions — it should be seen as symbolic and representative of a broader trend.[39]

To address these concerns, open sharing of various resources needs to go hand in hand with greater attention placed on data quality, data governance, responsible sharing and respect and protection for various rights in data. In the context of Open Source AI development, this requires a shift from concern over the volume of data that is available for AI training to the quality of the data and specific governance mechanisms that ensure that data is shared in ways that are equitable, sustainable and protected from value extraction.

In other words, there is a need to shift from a perspective that focuses on building Open Source models and treats data as a means towards that end. Instead, sharing data should be treated as a goal in itself, with a different set of incentives and challenges to consider. From this perspective, preservation of the knowledge commons and acknowledgment of various social aspects of data generation become key issues that the Open Source AI development ecosystem needs to pay attention to.

Finally, there is a third need to look beyond open licensing and consider other mechanisms — often not copyright-based — to address various legal frictions that hinder data sharing and to deal with considerations of social impact. Work on opt-out frameworks and preference signaling for AI training are the best examples of this trend, but the range of issues and possible solutions is much broader.

NOTES

37 Verhulst, Stefaan G., and Friederike Schüür. "Interwoven Realms: Data Governance as the Bedrock for AI Governance." Data & Policy Blog, 20 Nov. 2023. https://medium.com/data-policy/interwoven-realms-data-governance-as-the-bedrock-for-ai-governance-ffd56a6a4543.

38 Longpre, Shayne, et al. "Consent in Crisis: The Rapid Decline of the AI Data Commons." arXiv, 24 July 2024. arXiv.org. https://doi.org/10.48550/arXiv.2407.14933.

39 Tarkowski, Alek, and Zuzanna Warso. "Shifting Tides. The open movement at a turning point." Open Future. https://openfuture.eu/publication/shifting-tides/.

A paradigm shift is needed

Open Source AI developers, in order to address the data-related challenges in AI development, need to treat the definitions and standards of open AI systems as foundations on which further strategies can be built. This requires two, related paradigm shifts:

- **Adopting a "data commons" approach:** We propose shifting from a focus on just the openness of data to data commons and more varied forms of data governance. This also entails licensing innovation, that aims to introduce novel mechanisms while preserving core open functions of the licenses. This shift is needed to address the complexity of data that potentially can be used in AI training datasets. For some types of data, a basic open-sharing framework falls short of preventing data exploitation. This paradigm necessitates more robust commons-based governance models. The shift would result in an acknowledgment of a gradient of data sharing approaches, where open data is the optimum on one side of the spectrum, with other data sharing approaches — suited for cases where open sharing is not desirable or attainable — on the other side. The shape of these solutions is currently under development, with key questions concerning the extent to which sharing can be meaningful while being gated.
- **Expanding the stakeholder universe:** We propose shifting from solely meeting AI development needs to a broader view of data sharing that serves the needs and objectives of a broader set of stakeholders. A good example of a limited perspective taken by AI developers comes from the space of language model development, where developers tend to treat everything as data that can be turned into tokens and packaged into datasets. The stated goal of many advocates of Open Source AI development is to obtain access to as much data as possible, shared openly or permissively. The shift is needed to take into consideration the needs and goals of various other stakeholders who are rights holders in creative or research works, stewards of various collections, or administrators of repositories. Such expansion and bridging of perspectives is necessary to successfully share new sources of data.

First paradigm shift: from beyond open data to data commons

The concept of data commons and the various data governance models stemming from it have been increasingly prominent in debates about data governance. Definitions of data commons (and digital commons more broadly) focus on collaborative, democratic and participatory approaches to data governance.[40, 41, 42] The idea of data commons, treated more broadly, encompasses any approach that opposes the concentration of power and value extraction.

In this report, the term data commons is used to describe a broad range of approaches to data sharing, from Open Data to more limited forms of sharing, aimed at providing greater control over data and protecting various rights in data, while ensuring public interest reuse. Using this concept is a way of acknowledging that a spectrum of approaches to data access and sharing are needed to make the best use, in the public interest, of data that is in turn open, public, obtainable and unshareable.

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 fairly and responsibly. This is often done by making sure that anyone who uses a commons-based type of AI resource is also obliged to give back any improved or augmented version of the AI resource to the local and global community. In addition, it offers a set of normative principles that guarantee that data is not just accessible but also adequately governed and that it can be a shared resource sustained by many for many. Data commons in relation to AI have been explored by various organizations and initiatives, including Open Data Policy Lab,[43] the Collective Intelligence Project,[44] CNRS,[45] Open Future[46] and Coding Rights[47] — to name just some of them. The use of Open Data in the AI development context should be seen as a specific case of a data commons approach — recently explored by organizations like the Open Data Policy Lab, Open Data Charter,[48] Mozilla Foundation,[49] and the Digital Public Goods Alliance.[50]

Commons-based approaches balance 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.

Data commons approaches often share data in ways that are more limited, granular or conditional than the open data approach. While this initially might seem like a limitation — an approach that waters down data-sharing requirements and thus the usefulness of the data — it, in fact, ensures that more types of data can be shared. Data commons approaches adhere to the spirit of sharing while providing stronger forms of commons-based governance that ensure responsible and equitable use of data. Access with certain conditions (e.g. reciprocity) to data is a basic component of commons-based approaches. Other mechanisms are meant to protect from the exploitation of the commons: use of the common pool, without contributing back to its sustainability.

It is important to note that open data and data commons are not opposed but are part of the same spectrum. This was acknowledged in 2021 by GovLab's Open Data Policy Lab, as it argued for a third wave of open data that prioritized responsible use and data rights.[51] Copyright and privacy or personal data rights are the two most important factors determining how a data set can be shared — how "open" it can be. There is a spectrum of openness of data sets ranging from content that is not subject to copyright and does not include personal data to highly sensitive data. For this reason, it is no longer sufficient to say: can it be made open? The question instead becomes: how can as much data as possible be shared with necessary restrictions?

While restrictions were traditionally seen by designers of open frameworks as limitations to the power of open sharing, they are something different from a data commons perspective: rules that are necessary to balance the value of openness with considerations of data rights, fairness and equity. At its most basic, a data commons framing also means greater care for how data is curated and made useful — avoiding a negative scenario where troves of data are available but unusable or of very low quality.

Ultimately, a commons approach needs to find a way to balance a fair sharing of obligations and value for and between the following groups[52]:

1. Communities and organizations involved in the collection of the AI data;
2. Communities and organizations who aggregate, curate and vet such data;
3. Institutions that nurture and use such data;
4. People who are represented in such data;
5. Entities that use data for AI training (and might or might not contribute back);
6. The general public (who expects to benefit from better AI tools.)

From the perspective of standards for the openness of AI models, adopting a data commons perspective means that data will be the model component for which standards for sharing will be most challenging to define. It is also worth noting that, while various data commons frameworks have been proposed — via notions like data trusts, data cooperatives and others — few of these frameworks have been successfully deployed in real life and even fewer have scaled successfully. Examples of such an approach include the Development Data Partnership, Industry Data for Society Partnership and the UN Biodiversity Lab.

This means that data commons — to be successful — requires something that the Open Source movement has done successfully: defining a set of relatively simple, standardized mechanisms that can be deployed to further commons-based data governance.

Considerations of a spectrum of data commons approaches should also not detract us from supporting and maintaining open-sharing frameworks deployed in the last two decades. Similarly, there is only a fine line running between gated forms of sharing and closed or proprietary modes of data ownership. To give one example, opt-out mechanisms are today considered a new, consent-based mechanism that can go hand in hand with open sharing frameworks. Looking beyond individual consent, there are also efforts to govern collective rights in data, for example through a social license.[53] In each case, additional governance mechanisms run the risk of reducing the data commons through a shift to permission-based, licensed uses, contrary to the traditional ethos of Open Source or open data. Opt-outs, especially if occurring at scale, can also introduce bias into datasets that are meant to be representative of certain populations or types of content.

Second paradigm shift: a stakeholder universe beyond AI developers and dataset creators

Framing the issue as one of sharing training datasets for AI systems is just one way of considering data governance and sharing. The development of AI systems offers a crucial perspective for thinking about data, as it provides a strong case for creating value using publicly available data. In the past, lack of reuse has been a constant worry of advocates for data sharing and other related forms of openness. There was a risk that, on average, investments in open resources would not result in any meaningful use. Today, any publicly available data is potentially useful as a resource for AI training and research.

Yet, as was signaled earlier, the use of data can quickly turn from beneficial to lacking purpose or even exploitative. Ways in which LLM training datasets have been created in the past suggest that AI developers tend to treat all resources through the lens of them being potential training data and judge their value based solely on whether they can easily be turned into high-quality tokens. This is brought to an extreme in closed data sets used to train the latest generations of LLMs, with their complexity obfuscated and reduced to a single value: billions or trillions of tokens. It is as if AI developers treated dataset creation as a primarily technical process focused on successfully manipulating vast data sets so that model training architectures can be ingested, disregarding in the process various factors that they consider externalities.

Research into AI training datasets shows that in various cases creators have been ignoring such aspects as the cultural context, norms around content use, or even various rights in data, from copyright to privacy.[54] Eryk Selvaggio summarizes critical research on LAION datasets by stating that there is "ongoing negligence of the AI data pipeline."[55] The issue is exacerbated by the fact that datasets like LAION often have foundational character and are reused as core building blocks for a great number of models.

At fault here are mainly the producers of the dominant, commercial AI models, who are increasingly criticized for their data harvesting and dataset curation practices, as well as exploitative labor practices related to data work. Open Source AI developers, in turn, have often attempted to set higher standards of data governance. The BigScience project, which resulted in BLOOM LLM, is an excellent example of an AI development initiative that aimed to build and curate datasets responsibly and inclusively. More recent examples include the Common Corpus, PD12M and Dolma.

Datasets built out of web-scraped content constitute a particularly important category due to their broad reuse by producers of both open and closed models. Proper governance and quality control of these datasets is therefore essential. A study of Common Crawl, a source of web-crawled data, by Stefan Baack shows that problems are caused not by the dataset itself, but by ways in which it is being used by AI builders (for example, to create datasets out of samples of this source).[56]

To increase the pool of data, new data sources, collections and datasets need to be made available — and this requires the collaboration of AI developers with those stakeholders who own, steward or have access to such data. Bridging the gap between AI builders and content stewards is necessary not just to unlock new training data; these collaborations can establish better data governance practices, building on the experience of stewards of various types of resources.

This paradigm shift will require acknowledging that data is more than just a resource for AI model training used at a massive scale. AI developers need to acknowledge that data is being, or can be, shared for other purposes and that making datasets available for AI training can also have other benefits which are positive externalities of AI development. They also need to acknowledge that there are competing considerations, such as personal data protection or copyright or generating value to a certain community, that can legitimately result in data not being available as input for training AI models. They also need to understand that while various types of content have an obvious use for data training, other stakeholders might be questioning these uses. Finally, while for AI developers, the creation of models with the use of training data is a goal in itself, other stakeholders will want to see these models serve their specific needs and goals.

There is also a more general need to map all the stakeholders involved in the life cycle of an AI system, including those impacted by the system and the authorities that might need to audit such

systems. In this life cycle, those contributing to the training datasets later often become users of the AI systems or are impacted by them. Taking this into account, mechanisms that introduce reciprocity and make the life-cycle circular and regenerative, need to be part of data and AI governance. For example, the Open for Good Alliance has proposed the following measures that would help avoid unfair exploitation of communities and stakeholders in the Global South: introduction of "share alike" approaches; sustainable access to compute; support for data curation; capacity building; and support for the development of quantized, low-resource models.[57] Reciprocity, as a general principle for data and AI governance, ensures the sustainability of the data commons, which are being used — and in many cases exploited — in AI development.[58]

This shift can go hand in hand with and build upon the current AI development trend related to smaller generative models, such as the previously mentioned GPT-SW3, Bielik or Sea Lion. While there is still a lack of clarity on what constitutes a small model as opposed to a large model, there is clearly a new category of models that is qualitatively different from the large commercial models. The distinction is related not just to their smaller size. These alternatives use various AI training architectures and model fine-tuning techniques that allow more energy-efficient models to be derived from existing models and small models to compete successfully with larger ones. For example, the Masakhane organization has released InkubaLM-0.4B, a small model for five African languages that are considered "low resource."

Yet this trend is not just about effectiveness coupled with smaller size. Large commercial models are built as monolithic, general-purpose technologies that purportedly can be used globally. Yet research shows that they embed biases, knowledge and cultural gaps that make them colonial: reducing the diversity of cultures, knowledge and contexts. An emerging ecosystem of smaller models signals the possibility of technologies tuned towards local contexts and needs and also more efficient and sustainable to deploy. In such a scenario, data can be made available not for a massive, monolithic technology, but for a local one.

This paradigm shift should begin by establishing stronger ties with organizations and experts working in other fields of openness, such as Open Data, Open Science, Open Access or Open Culture. These entities have decades of experience making various types of content and data available. They are also facing the same challenges of open sharing, which have been identified in the Open Source AI development space. Collaboration with these entities and networks would allow for shared exploration of uses of AI systems. For example, a growing number of initiatives are looking at AI in science from a combined perspective of open science and Open Source AI. Even speaking about "AI for science" offers an important shift in narrative from a general-purpose system that might as well be purposeless to a specialized use that can easily be understood as serving the public interest and, thus, less exploitative.

Such collaborations would allow for:

- **Commons-Based Standards:** Development of new frameworks and standards for open and commons-based sharing of resources;
- **Training Resources:** New resources to be made available using these frameworks and standards for AI training and other purposes; and
- **Use cases:** Use of these resources as data for the training of AI models that would be purposefully built to address various public interest issues.

NOTES

40 Dulong de Rosnay, Melanie, and Felix Stalder. "Digital commons." Internet Policy Review, 2020, Concepts of the digital society, 9 (4). https://policyreview.info/concepts/digital-commons.

41 Krewer, Jan, and Zuzanna Warso. "Digital Commons as Providers of Public Digital Infrastructure." Open Future. https://openfuture.pubpub.org/pub/digital-commons-public-digital-infra/.

42 Tarkowski, Alek, and Zuzanna Warso. "Commons-Based Data Set Governance for AI." Open Future. https://openfuture.eu/publication/commons-based-data-set-governance-for-ai.

43 Chavetz, Hannah, et al. "Tackling 4 'Common' Problems: Accelerating the Next Generation of Data Commons at a Time Of Artificial Intelligence." Data & Policy Blog. https://medium.com/data-policy/tackling-4-common-problems-accelerating-the-next-generation-of-data-commons-at-a-time-of-a84e2c9a1602.

44 Huang, Saffron, and Divya Siddarth. "Generative AI and the Digital Commons." https://www.cip.org/research/generative-ai-digital-commons.

45 Benhamou, Yaniv, and Melanie Dulong de Rosnay. "Open Data Commons Licenses (ODCL): Licensing Personal and Non Personal Data Supporting the Commons and Privacy." Social Science Research Network. https://doi.org/10.2139/ssrn.4662511.

46 Tarkowski, Alek, and Zuzanna Warso. "Commons-Based Data Set Governance for AI." Open Future. https://openfuture.eu/publication/commons-based-data-set-governance-for-ai.

47 Varon, Joana, Sasha Costanza-Chock, and Timnit Gebru. "Fostering a Federated AI Commons ecosystem." T20 Policy Briefing. https://codingrights.org/docs/Federated_AI_Commons_ecosystem_T20Policybriefing.pdf.

48 "Using Open Data to Improve AI Initiatives." Open Data Charter. https://medium.com/opendatacharter/using-open-data-to-improve-ai-initiatives-ca9e37f22f3d.

49 Basdevant, Adrien, et al. "Towards a Framework for Openness in Foundation Models." Mozilla Foundation. https://assets.mofoprod.net/network/documents/Towards_a_Framework_for_Openness_in_Foundation_Models.pdf.

50 Nordhaug, Liv Marte. "The Role of Open Data in AI Systems as Digital Public Goods - Digital Public Goods Alliance." Digital Public Goods Alliance. https://www.digitalpublicgoods.net/blog/the-role-of-open-data-in-ai-systems-as-digital-public-goods/.

51 Verhulst, Stefaan, G., et al. "The Emergence of a Third Wave of Open Data." Open Data Policy Lab. https://opendatapolicylab.org/images/odpl/third-wave-of-opendata.pdf.

52 "Open-Source AI Data Sharing: yes! Data Colonialism: no!" Open for Good Alliance. https://medium.com/@openforgood/open-source-ai-data-sharing-yes-data-colonialism-no-3062a922de03.

53 Verhulst, Stefaan G., Laura Sandor and Julia Stamm. "The Urgent Need to Reimagine Data Consent." Stanford Social Innovation Review. https://ssir.org/articles/entry/the_urgent_need_to_reimagine_data_consent.

54 See for example these critical studies of datasets: Buschek, Christo, and Jer Thorp. "Models All The Way Down." Accessed 20 December 2024. https://knowingmachines.org/models-all-the-way; Birhane, Abeba, Vinay Prabhu, Sang Han, and Vishnu Naresh Boddeti. "On Hate Scaling Laws For Data-Swamps." arXiv, 28 June 2023. https://doi.org/10.48550/arXiv.2306.13141; and Birhane, Abeba, Sepehr Dehdashtian, Vinay Uday Prabhu, and Vishnu Boddeti. "The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models." arXiv, 7 May 2024. https://doi.org/10.48550/arXiv.2405.04623; Paullada, Amandalynne, et al. "Data and its (dis) contents: A survey of dataset development and use in machine learning research." Patterns, Volume 2, Issue 11, 100336. https://www.cell.com/patterns/fulltext/S2666-3899(21)00184-7.

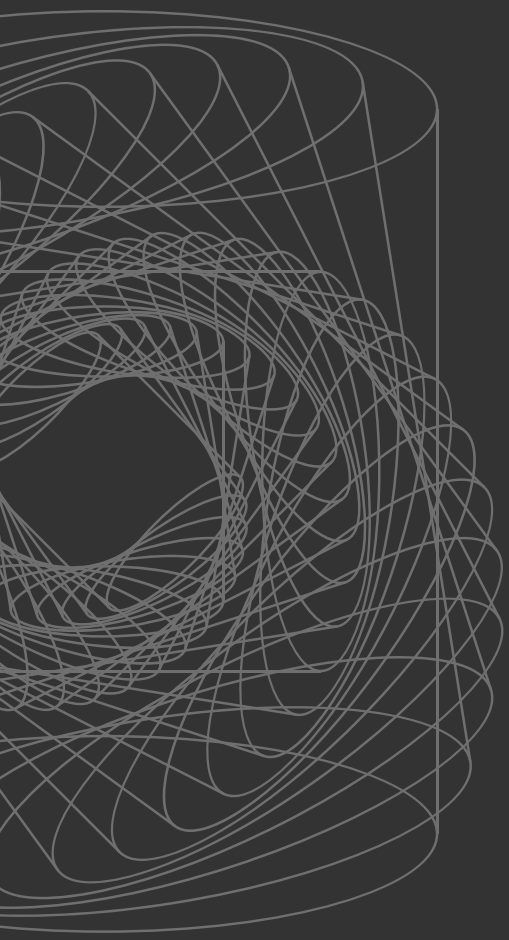
NOTES

55 Salvaggio, Eryk. “LAION-5B, Stable Diffusion 1.5, and the Original Sin of Generative AI | TechPolicy.Press.” Tech Policy Press, 2 January 2024. https://techpolicy.press/laion5b-stable-diffusion-and-the-original-sin-of-generative-ai.

56 Baack, Stefaan. “Training Data for the Price of a Sandwich.” Mozilla, 6 February 2024. https://foundation.mozilla.org/en/research/library/generative-ai-training-data/common-crawl/.

57 Open for Good Alliance. “Open-Source AI Data Sharing: yes! Data Colonialism: no!” Medium. https://medium.com/@openforgood/open-source-ai-data-sharing-yes-data-colonialism-no-3062a922de03.

58 Keller, Paul. “AI, the Commons, and the Limits of Copyright.” Open Future. https://openfuture.eu/blog/ai-the-commons-and-the-limits-of-copyright.



Searching for solutions

To address the challenges outlined above, the Open Source AI community should:

- **Scale the Data Commons Approach:** ensure that the data commons grows and that more data is shared, both as open data and through other data-sharing mechanisms meeting the definition of a commons (including new licensing regimes); and
- **Expand the Stakeholder Universe:** ensure that commons-based governance mechanisms are used to share both existing and new data sources.

To deal with the dual challenge of making data available for AI training and protecting it from exploitation, the Open Source AI community needs to move beyond the issue of open data requirements and actively support efforts to expand, sustain and protect data commons.

This requires acknowledging that various categories of data exist. For some, Open Data will be the right approach, one that builds on decades of experience with open data sharing. For others, new data sharing mechanisms need to be used. These have been intensively explored and conceptualized — yet there are today few practical implementations in the space of AI training. And the requirements for governing data as a commons often feel complex — especially in comparison to the relatively simple frameworks for open sharing. Thus, there is a need to seek within the broad space of commons-based data governance mechanisms solutions that are relatively simple, standardized and replicable at scale — through which data commons can be expanded.

To address the dual challenge of the lack of data and exploitation of existing data, three categories of data sources need to be dealt with:

1. **Current Sources of Shared Data:** Data that is already being shared: for these sources, there is a need to ensure that they are properly governed, and commons-based mechanisms can ensure that this is done fairly and responsibly. This in particular concerns datasets built on top of web crawl data, which continue to be core resources for training AI models;
2. **Future Sources of Shared Data:** for these sources, making them available requires removing existing barriers to sharing, which in turn often requires alignment with the needs and interests of individuals and entities owning the rights to these resources or stewarding them. There is also a case to be made for new digitization efforts and for a renewed commitment to data sharing from public institutions; and **Closed and Proprietary Data:** for these sources of data, used in AI training and owned by the dominant AI companies, there is a need to ensure — at a minimum — that data provenance and data information standards are met. There is also the need to revisit proposals for mandating access to some types of privately owned data.

There is a long list of specific issues that are transversal to this categorization and that need to be addressed for existing data to be shared more responsibly, for new data sources to be made available and for proprietary data to be made more accessible. Among these, key issues include:

- **Regulation and Legal issues:** A patchwork of varied laws in different jurisdictions creates a significant obstacle to sharing data, from a global perspective. At the same time, favorable legislation can support data sharing efforts, while also setting boundaries for data that should not be made available;
- **Building global and regional alliances** that work on concrete AI data commons and share lessons learned on working governance-approaches, especially those that have global reach and include Global South communities and the public interest;
- **Licensing Frameworks:** while data commons approaches go beyond open licensing, these frameworks still remain core data sharing mechanisms. There is a need to both maintain existing licensing frameworks and to innovate on novel licensing approaches; and
- **Dataset Design:** There is a continuing need to innovate on dataset design, with a focus on data governance mechanisms that ensure data quality, protect rights in data rights and ensure responsible sharing.

Six focus areas for data and Open Source AI

Below, six focus areas for improving data sharing and data governance are presented, together with examples of specific initiatives and solutions.

Focus area: data preparation

Most data use cases begin with data preparation. When done properly, it becomes a solution to many problems and concerns related to datasets and their governance. There are cases of training datasets being released without the necessary preparation, necessitating issues to be fixed retroactively — decreasing trust in these datasets and in some cases leading to harm. These can be avoided by proper data preparation up front.

Data preparation, combined with data provenance standards, creates a signal — or a reflection point — for developers, so that they understand at what point they need to consider various data governance issues, before they develop and deploy a technology based on this data. These signals can carry information not just about regulatory and compliance issues like privacy or copyright, but also about community norms. For example, where healthcare data is used for machine learning models, data preparation and provenance can signal to developers to verify patient consent for secondary data usage or ensure that patient information is anonymized in line with privacy laws. But provenance, especially standards, can also indicate if the data was collected with specific community engagement principles, such as including diverse demographic representations, so that developers ensure their models are inclusive and ethically aligned with healthcare community principles.

Proper data preparation includes procedures and frameworks for:

- Provenance: these form a basis of trust by providing visibility into where the data came from, how it was created, the potential risks and how to mitigate them, and appropriate and inappropriate uses. Proper provenance information provides a framework for other issues of focus listed below, such as licensing or consent signals;
- Anonymization: crucial for datasets with Personally Identifiable Information (PII) or Protected Health Information (PHI) and allowing reuse of data without abuse of data subjects' rights;
- Classification, annotation and tagging: processes that increase the quality of data; and
- Standardization: Enables interoperability of datasets, supports transparency and makes auditing easier.

Proper data collection is related to data preparation and is a measure that ensures that data is representative and equitable. Combined with provenance and classification, this allows for structural gaps in the dataset to be identified. Data collection also should include proper assertion of rights and identification of preference signals.

Examples

- Data nutrition labels
- Data and Transparency Alliance's data provenance metadata standard
- Data Provenance Initiative

Focus area: preference signaling and licensing

There is a need to develop signals that provide broader information than just licensing, which today is captured in terms of service, contracts or public licenses. This also means looking beyond copyright-related signals around data usage. In many cases, this is the issue of transferring mechanisms that are common in data sharing agreements made between two parties into mechanisms that can support open data sharing of some form.

In recent years, frameworks for preference signaling have become the focus of attention in debates about AI training data and are seen as being as relevant as those for the assertion of rights, such as open licensing frameworks. Preference signals are understood as mechanisms that allow rights holders to establish more fine-grained terms under which the data can (or cannot be) used for AI training.

Robots.txt files are the canonical example of preference signaling. These are based on community norms and at the same time are a standardized framework that functions — and is successfully enforced — across the open web. Today, there is a need to develop a more fine-grained vocabulary that distinguishes various types of web crawling and web scraping activity, not all of it related to generative AI training.

Key development challenges for preference signaling concern infrastructure and enforcement. There is a need to develop mechanisms for signaling preferences across various layers of the internet stack, between various modalities and for various types of data. Related to this, there is currently a lack of clarity on how these signals could be enforced, as they are not based on copyright licensing mechanisms that offer some means of enforceability. Speaking more generally, there is a need to translate legal and licensing mechanisms to the technical and operational level.

The development of Open Source AI offers the possibility of building technology that is more contextualized, developed closer to local communities and fitting their needs. Therefore, the discussion also considered ways in which preferences can be expressed by communities, rather than individual creators. For example, in the case of linguistic data — especially for "rare languages" collected from local communities — there is a need to ensure that the collection and use of such data will be equitable with direct and immediate or near-immediate benefits to the community. While open licensing frameworks are not suited for that, they can be combined with additional mechanisms that express community norms and preferences. Innovation in licensing also includes exploration of social licenses[59] and frameworks that combine open licensing and data trust approaches.[60]

Examples

- Our Knowledge, Our Way, guidelines created by the Australian Commonwealth Scientific and Industrial Research Organisation (CSIRO)
- Te Hiku Media's Kaitiakitanga license and other examples of indigenous licensing frameworks
- Do Not Train Registry, created by Spawning.AI
- Creative Commons work on preference signalling

Focus area: data stewards and custodians

The growing complexity of data sharing frameworks and the growing need to establish collaborative, collective approaches can be addressed by introducing data stewardship functions. Data stewardship is a function typically defined in private companies or public institutions, held by units that are empowered to initiate, facilitate and coordinate data collaboration and sharing.[61]

Traditionally, open sharing is seen as an approach that requires only a minimal stewardship function, related to license enforcement and technical operations. Today, the limits of this approach are visible, and data stewardship is needed to prepare the data, maintain the data and reduce various frictions related to data sharing while protecting the various rights and interests in data. While some datasets are criticized for not being responsibly governed, those with a stewardship function — like Common Crawl — set a higher standard for data governance. Also, many collections that either are used for AI training or could become such sources are being stewarded by various public interest institutions: libraries, heritage institutions, research bodies, etc.

Data exchanges are specific forms of data stewardship that serve as trusted intermediaries between data owners and data users. These can ensure sound governance, compensation or solution-focused use of data.

As data stewardship approaches are explored, it would be beneficial to identify common, replicable stewardship frameworks that could be introduced. This would allow, on one hand, for best practices to scale and, on the other, would increase certainty on the side of data users. Such replicable frameworks should be based on existing standards (for example, data provenance metadata standard or opt-out vocabulary standard) and aim to collectively develop further standards that are needed.

One specific aspect of data stewardship is license enforcement. There is a need to explore how collective institutions, similar to collective management organizations, could bring claims for violations to ensure more trust that license conditions will be followed.

Examples

- Open Data Commons licensing framework, developed by Open Knowledge Foundation
- MIDATA, a health data cooperative
- Open Humans
- Common Crawl's work on stewarding a shared repository of web crawl data
- Software Heritage Foundation's universal software archive

Focus area: environmental sustainability

The environmental impact of AI systems was identified as a major challenge to Open Source AI, yet few approaches to addressing these issues were discussed. Overall, the sharing of data is seen as contributing to lessening this impact. For example, public databases of web-crawled data help reduce the amount of excessive web crawling and web scraping that is taking place. The need to address these concerns can be a strong incentive for data sharing — for example, data sharing could be part of commercial environmental, social and corporate governance (ESG) strategies.[62]

Care for environmental sustainability also means developing more transparent information about the sustainability of various datasets.

Examples

- ESG DataBank
- Sensor Observation Service
- SensorThings API

Focus area: reciprocity and compensation

Both individual creators, as well as collective and institutional stewards of various collections and datasets, are increasingly worried that their data will be exploited, with generated added value going only to the large tech companies. This vulnerability of data shared as a commons is both a real threat and a factor that can reduce the incentives to share data. Mechanisms that introduce reciprocal value sharing were discussed at the workshop as needed to address this challenge. They can be seen as a solution that is complementary to ongoing work on preference signaling. While the latter aims to limit, in a controlled way, the use of data for AI training, reciprocity measures are meant to change the way value is given back to the commons, while data is being shared.

Copyleft licensing — both of software and content — has traditionally been the key mechanism to establish some form of reciprocity. There is growing awareness that these copyright-based mechanisms might not be legally enforceable across the AI stack, as training data is transformed into model parameters that later determine the outputs of generative AI systems. There is an urgent need to determine whether copyleft licenses are still fit for purpose and to explore alternative means of securing copyleft requirements.

At the same time, the issue of reciprocity and value sharing should not be seen just as a matter of individual rights and licensing tools that build on these rights. There is a need to develop collective mechanisms for exercising rights and securing the return of value. These have been conceptualized either in terms of collective rights of communities or in terms of care for the data commons as a collective good. The issue therefore should be framed as one of economic justice, rather than just copyright compliance.

Explorations of mechanisms for ensuring reciprocity in data sharing consider institutional frameworks like data trusts and data cooperatives (and other forms of data intermediaries) and regulatory solutions like compulsory licensing with remuneration, or some form of taxes or levies. Reciprocity could also be embedded in public procurement arrangements. Any such mechanisms need to consider exceptions for nonprofits, public interest use such as research, and commons-based, community-driven Open Source development.

Examples

- Licensing African Datasets initiative
- Wikimedia Enterprise

Focus area: policy interventions

While most of the workshop was devoted to solutions that can be deployed by Open Source AI developers, dataset curators and stewards of collections, policy interventions were also considered. These can help create incentives for sharing data openly and more generally for Open Source AI development. Among measures discussed at the workshop, several were considered most relevant, based on the dual criteria of importance and likeliness of success:

- Policies that introduce public procurement of new, open data sets and that leverage government procurement to encourage the use of Open Source AI systems and their components in public digital infrastructures;
- Policies that build on existing Open Data strategies and mandate that public data is findable, reusable, AI-ready — where possible — open. This also entails the creation of specialized institutions that steward other types of data that are obtainable or unshareable, such as health data;
- Policies that introduce stronger requirements for transparency on training data. A transparency norm would create incentives for using open data, and transparency mandates can go hand in hand with the OSAID's requirement of sharing Data Information;
- Policies that mandate or incentivize the deployment of signal preference frameworks and their adoption by rightsholders;
- Policies that mandate a data stewardship function within organizations that have data that could be leveraged for public interest purposes; and
- Policies that mandate public archival of publicly available (but not necessarily open) data, so that they remain available in the future;

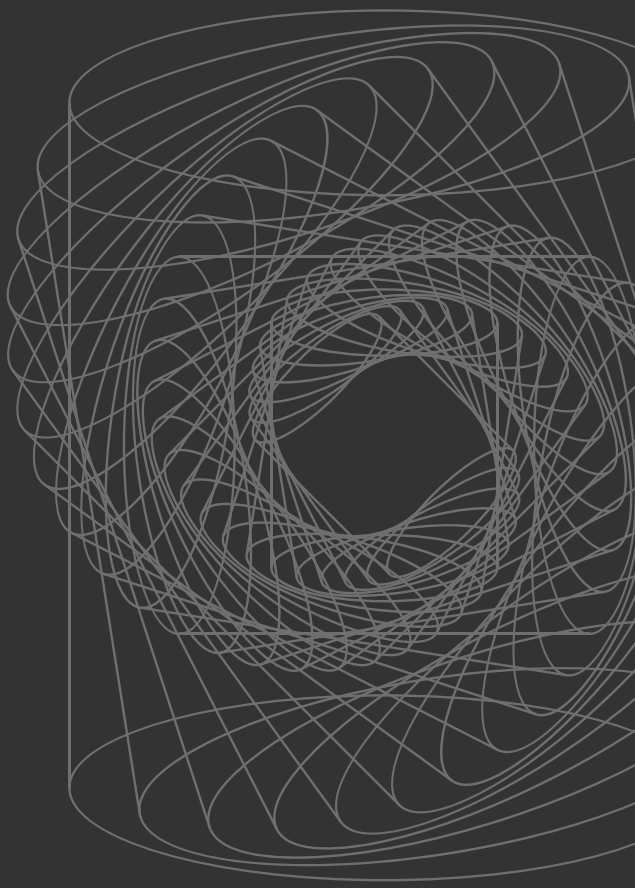
NOTES

59 Verhulst, Stefaan G., Laura Sandor, and Julia Stamm. "The Urgent Need to Reimagine Data Consent." Stanford Social Innovation Review. https://ssir.org/articles/entry/the_urgent_need_to_reimagine_data_consent.

60 Open Knowledge. "From Open Data to Sustainable Data Commons." Open Knowledge. https://okfn.org/en/projects/sustainable-data-commons/.

61 Verhulst, Stefaan G. "Data Stewardship Re-Imagined — Capacities and Competencies." Medium. https://medium.com/data-stewards-network/data-stewardship-re-imagined-capacities-and-competencies-d37a0ebaf0ee.

62 Verhulst, Stefaan G. "The Case for Including Data Stewardship in ESG." Barron's. https://www.barrons.com/articles/esg-data-companies-governance-9a3bd57d.



Paths forward

Taken together, work in these various focus areas serves two goals. First, it serves the purpose of increased data sharing, by making various types of data easier to use, by increasing the quality of datasets and by ensuring that more data is available openly. Second, it protects knowledge commons by acknowledging a broad range of social aspects of data generation and associated legal frictions and deploying mechanisms other than licenses to offer adequate governance.

Further efforts should be applied both to existing and new datasets. There is an opportunity to improve the governance of existing datasets created as tools by AI developers — especially those that are most often reused for Open Source AI solutions. There is also a need to design and build datasets that both increase the volume of data available for AI training and set stronger data governance standards. Finally, there is a need to recognize various existing collections as potential sources of AI training data, while recognizing their inherent value and acknowledging existing forms in which they are governed and maintained.

In addition, collective efforts should continue to establish, where possible, standards related to data governance and to provide guidance on ways that they should be implemented.

About the white paper

This white paper is largely based on insights gathered during a two-day, in-person convening organized by OSI and Open Future on 10 – 11 October 2024. The convening brought together a group of around twenty experts from organizations involved both in Open Source AI and in various types of data sharing initiatives.

The convening allowed the authors to refine the problem definition and to identify and co-design focus areas for intervention and even specific solutions. The goal of the workshop was not to provide a complete mapping of the various ways in which data sharing can be increased and improved. Instead, the goal was to identify specific solutions that are particularly relevant in relation to Open Source AI development.

We are grateful to the co-design participants for their insights.

Dr. Alek Tarkowski

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.

Open Source Initiative

The OSI is the authority that defines Open Source, recognized globally by individuals, companies, and by public institutions.

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.