THINK7 CANADA 2025

G7 2025 KANANASKIS

Policy Brief

Task Force 1: Transformative Technologies — AI and Quantum

Enabling an Open-Source AI Ecosystem as a Building Block for Public AI

Authors:

Katarzyna Odrozek

Vidisha Mishra

Anshul Pachouri

Arnav Nigam

April 2025

Key Points

This policy brief, informed by insights from 30 open dataset builders convened by Mozilla and EleutherAI and a policy analysis on open-source Artificial intelligence (AI) development, outlines four key areas for G7 action:

1. Expand Access to Open Data – Expand the availability of structured, high-quality open datasets by addressing legal ambiguities and incentivising digitization efforts.
2. Support Sustainable Governance – Establish infrastructure-focused funding and governance frameworks to support open datasets as Digital Public Goods, ensuring long-term sustainability, fair value-sharing, and regulatory alignment.
3. Encourage Policy Alignment in Open-source AI – Support policy convergence among developing and developed countries for the open-source AI development through policy dialogue, and knowledge-sharing platform
4. Local Capacity Building and identification of use cases – Enhance local capacity in open-source AI development, promote international collaborations, and identify and develop locally relevant and impactful AI use cases

These steps will enhance AI competitiveness, accountability, and innovation, positioning the G7 as a leader in Responsible AI development.

Introduction

The global AI race is no longer just about who builds the most powerful models, but who controls access to the data that fuels them. Across the G7 and beyond, major economies are taking different approaches to AI development, governance, and regulation.

The United States remains the dominant force in private-sector AI innovation in the developed world, with companies like OpenAI, Anthropic, and Google DeepMind leading the field. With the EU AI act, the European Union has positioned itself as a regulatory leader, focusing on transparency, accountability, and ethical safeguards. The UK, Canada, and Japan are carving out national AI strategies that balance innovation with public-interest oversight, while France and Germany are investing in open-source AI as a strategic counterweight to proprietary models. In the developing world. China is rapidly scaling its AI capabilities through ground-breaking models like Deep Seek. Africa is trying to adopt a regional approach to foster the development of local AI ecosystems (African Union 2024). India has developed its own AI-enabled language translation model (Bhashini) and plans to establish Centres of Excellence to drive cutting-edge research and innovation in AI (NITI Ayog, Government of India 2018).

The release of prominent open-source models reignited global discussions about the role of open-source AI (Edmond 2025), the need for transparent and reproducible training datasets, and how governments can enable AI development that is both competitive and accountable. It is a strategic chance to drive this narrative in service of open-source AI as

part of the Digital Public Infrastructure (DPI) and Public AI (Mozilla 2024). A Digital Public Infrastructure (DPI) approach—modelled on open-source, interoperable systems offers a pathway to democratize AI. Adopting DPI principles aligns with global ethical frameworks such as the UNESCO Recommendation on the Ethics of AI (2021), which emphasizes transparency, accountability, and inclusivity to ensure AI serves humanity. By leveraging DPI approach to build open datasets and open AI models, the G7 can bridge the AI divide among countries and ensure AI advances align with global public goods.

The Need for Open Datasets and Open-Source AI Models in AI Development

AI is increasingly shaping economies, societies, and governance, yet the datasets that power AI models remain largely opaque, inaccessible, and legally ambiguous (Vincent 2023). Today, large AI companies dominate the landscape by leveraging vast amounts of proprietary and publicly scraped data, often without clear consent mechanisms (Grynbaum, Ryan 2023). On one hand, smaller actors—including researchers, startups, and public institutions—face significant barriers in accessing high-quality, legally compliant datasets; on the other, they lack effective incentive mechanisms to train and deploy small sectoral and regional AI models that cater to the needs of people. As a result, competition, innovation, and independent oversight are stifled. This lack of openness has deep consequences:

- **Public accountability is compromised**, as researchers and auditors are unable to assess how AI systems are trained or whether they reinforce bias and misinformation.
- **Legal uncertainties surrounding data use** deter responsible dataset creation and prevent smaller players from engaging in AI development.
- **The decline of the open web**, driven by widespread opt-outs and copyright disputes, threatens the availability of legally permissible, high-quality data for AI training.
- **Public mistrust in AI is growing**, fuelled by concerns about exploitative data collection and the increasing opacity of AI decision-making.
- **Lack of sectoral and regional small AI models** that are crucial for making societal impacts in areas like healthcare, education, agriculture, etc.
- **Limited private sector innovation** due to enormous resource requirements to build training datasets and build AI models.

However, the opportunities of open datasets are significant, e.g.:

- **Public health advancements** – AI-powered diagnostics and epidemiological research rely on open datasets to ensure equitable access to medical innovations.
- **Climate resilience** – Open geospatial data enables AI-driven environmental monitoring and disaster response planning.

- **Economic growth** – Small and medium enterprises (SMEs) can leverage public AI models for financial inclusion, digital services, and workforce automation.

If AI is to serve the public good, it must be built on transparent, well-governed, and open datasets and open AI models that empower a broader range of stakeholders. The G7 has a unique opportunity to take leadership in shaping the policies, legal frameworks, and funding mechanisms needed to build a robust open dataset and open AI models ecosystem—ensuring that AI remains competitive, fair, and accountable.

With the following recommendations, we share insights from 30 open dataset builders convened by Mozilla and EleutherAI (Baack, Biderman, Odrozek, Skowron, et al. 2025), and policy analysis on open-source AI development.

Key Policy Recommendations for the G7

Expand Access to Open Data for AI Training

Access to high-quality, legally compliant datasets is a fundamental enabler of AI innovation and competition. However, vast amounts of publicly funded knowledge and cultural assets remain locked behind restrictive access policies. Even when data is technically open, it is often fragmented, poorly documented, or locked in inaccessible or gated repositories or formats, and the one that is accessible is often unstructured. The process of identifying licensing status and metadata across jurisdictions can be overwhelming. Many companies or institutions don't even know that and how they could release their data into the open. Finally, massive opt-outs of AI crawlers threaten to significantly reduce open data availability.

Legal uncertainty is one of the biggest barriers to open dataset development. Especially for volunteer-driven organizations without substantial legal support it remains a significant barrier and a chilling effect for the ecosystem. Many organizations and researchers fear legal repercussions when curating or using datasets due to unclear copyright frameworks and lack of global and consistent machine-readable licensing and consent information across jurisdictions that would allow to use openly licensed and public domain data without hesitancy.

To address all of the above, the G7 should:

- Mandate and fund large-scale digitization efforts for public domain and government datasets, ensuring they are structured in machine-readable formats and encourage cultural and scientific institutions to contribute to the open dataset ecosystem by providing sustainable incentives for making their collections openly accessible. This might include public-private partnerships, tax incentives or research grants that facilitate AI training on ethically sourced, well-curated datasets while maintaining respect for data rights holders and upholding security measures.

- Invest in open-source tooling, such as tools for extracting openly licensed content from difficult formats like PDFs. Providing these tools as open-source software would accelerate access to quality training data for AI systems.

- Clarify the legal status of AI training data to enable responsible use by investing in development and standardization of metadata and licensing requirements to facilitate cross-border AI research and open dataset building. AI training data should clearly document its source, legal status, and applicable permissions, reducing legal uncertainty for developers. Establishing harmonized "safe harbour" provisions would allow dataset creators to correct licensing errors without immediate legal liability. This would protect non-commercial and research-focused dataset initiatives, which often operate with limited legal resources.

By prioritizing open access to high-quality training data, the G7 can reduce reliance on proprietary datasets controlled by a few AI incumbents, ensuring that AI development remains competitive, diverse, and fair. However, it is imperative to recognize that not all datasets should be fully open—tailored data governance mechanisms and robust security measures must reflect the needs of data subjects and specific risk use cases.

For example, public data on infrastructure, such as energy grids or transport networks, can expose vulnerabilities to cyber threats, while Indigenous cultural heritage data may be exploited without appropriate protections (Donavyn 2021). To balance openness with security and ethical considerations, more targeted access privileges—combining transparency with responsible data stewardship, security, and data protection measures—should be applied on a case-by-case basis.

Support Sustainable Governance for Open Datasets for AI

Unlike proprietary AI datasets, open datasets lack clear financial sustainability models. Because they are freely accessible, they cannot rely on traditional revenue streams, leaving their maintenance underfunded and dependent on short-term grants or volunteer efforts. Without sustainable funding, even well-intentioned open dataset projects struggle to remain viable over time. Fair management of such datasets that respects the rights of individuals and communities need to go hand in hand with such efforts.

To ensure the long-term resilience and ethical governance of open datasets, the G7 should:

- Fund open dataset infrastructure as digital public goods. Public financing should prioritize high value, widely used datasets while supporting essential tools, ongoing maintenance, and sustainable governance. A key component of this effort is the development of public dataset repositories ("AI Data Commons") that provide trusted, high-quality training data for AI research and development. These repositories must adhere to clear metadata and licensing standards to facilitate responsible use and simplify the identification of public domain data.
- Promote fair value-sharing mechanisms that incentivize contributors while maintaining open dataset accessibility. This could include public-private funding

models, where companies benefiting from open datasets contribute to their upkeep, or exploring models like Wikimedia Enterprise, which provides structured data access to commercial users while keeping core datasets open for public use.
- Develop and legally enshrine governance frameworks that empower communities and institutions to implement dataset stewardship. Emerging models like data trusts and regulated data intermediaries can ensure fair management, transparency, and accountability in AI dataset governance.

By recognizing open datasets as essential digital infrastructure, the G7 can ensure long-term sustainability, promote ethical governance, and prevent valuable public data from being enclosed by private actors. However, significant challenges need to be overcome, including the legal and social complexities of data stewardship, and concerns over data sovereignty in the creation of AI Commons.

Encourage Policy Alignment on Open-Source AI Development

Many developed and developing countries often lack platforms, subject matter expertise and adequate resources to frame effective policy and technology responses to enable open-source AI development, despite its potential to democratize access and foster self-reliance. G7 countries should champion global policy alignment on open-source AI development by establishing multilateral dialogue platforms and global knowledge-sharing forums. These forums would enable policymakers, researchers, and industry leaders globally to exchange best practices, co-design governance frameworks, and address challenges in strengthening and scaling open AI ecosystems. It's an important priority for G7 as fragmented AI policies risk deepening the global innovation divide.

A G7-backed initiative would empower both developing and developed countries to support open-source AI development and build on global AI governance while advancing interoperable standards, ensuring AI systems reflect diverse socio-cultural contexts (Responsible AI working group report, GPAI 2023). By prioritizing policy convergence, the G7 can catalyze a shift from isolated capacity-building to collective problem-solving, ensuring open-source AI becomes a pillar of inclusive digital transformation.

To ensure inclusive open AI development, the G7 should:

- **Support global policy convergence on open-source AI development** through a dedicated multi-year discussion forum to collaboratively explore regulatory, ethical, and technical dimensions of open-source AI. This would include strategies to incentivize shared innovation, protect digital sovereignty, and mitigate biases in open-source AI models.
- **Establish an open-access knowledge-sharing platform** in partnership with G20, GPAI, etc. that can be readily used by the policymakers and other stakeholders globally to learn about different policy approaches and tools to support open-source AI development. This knowledge-sharing platform can be updated periodically and should cater to the needs of both developed and developing countries.

Strengthen Local Capacity and Use Cases in Open-Source AI Development

Countries within G7 and outside have different levels of maturity in AI technology talent. The availability of local open-source AI development talent significantly impacts the capacity and capability of the countries to develop locally contextualized open-source AI models. Local capacity building in AI is critical to ensure equitable participation in the global digital economy, avoid dependency on foreign technologies, and address region-specific challenges (e.g., healthcare, agriculture, climate resilience). Local expertise in AI ensures ethical AI that reflects diverse values (Responsible AI & Innovation and Commercialization Working Group report, GPAI, 2023).

Many regions and countries have recognized the importance of local AI capacity building as outlined in AI policy and strategy documents. For instance, the Continental AI Strategy of the African Union emphasizes research and innovation in AI by fostering partnerships between academia, the private sector, and public institutions in Africa. This approach ensures AI development aligns with the local and regional priorities and socio-economic contexts. Such policies underscore the necessity of local talent development, infrastructure enhancement, and governance frameworks tailored to specific national and regional needs.

To support local capacity building and use cases in open-source AI development, G7 should:

- **Facilitate and co-fund collaborations/ partnerships among research communities, academia, developer communities, and think tanks across developed and developing countries** to create open-source AI training programs, research initiatives, and innovation hubs. The G7 can partner with other multilateral and development institutions, philanthropic donors, national governments to establish such programs and initiatives and incentivize local research ecosystems through grants, fellowships, tax-exemption, etc.
- **Establish a global AI application use cases library by sectors and geographies.** The G7 can partner with multilateral institutions including G20 and GPAI, academia, and research institutions to populate the repository with context-specific examples and case studies. This library would serve as a dynamic knowledge hub, showcasing open-source AI innovations across different sectors and geographies (e.g., India's Bhashini for language inclusivity (National Language Translation Mission, n.d.), Brazil's deforestation monitoring systems ("Previsia", n.d.)). By aggregating diverse regional experiences from the developed and developing world, the library would enable the development of locally led use cases of open-source AI models.
- **Develop a framework to understand the risk-benefit and cost associated with the training and deployment of open-source AI models** by introducing standardized metrics to evaluate each use case's ethical, economic, and environmental trade-offs, including participatory assessments from affected communities (e.g., farmers using AI climate models or gig workers impacted by automation). This aligns with the voluntary reporting framework of the G7 Hiroshima AI Process to encourage

transparency and accountability among organizations developing advanced AI systems (G7 2023).

Author Biographies

Katarzyna Odrozek is a technologist, researcher, and advisor specializing in AI ethics, research strategy, and open community engagement. As Director of the Insights team at Mozilla Foundation, she led research and policy work on responsible AI while supporting ethical innovators. Kasia specializes in data for AI and regularly advises on funding in grant making bodies such as the EU AI & Society Fund or the Prototype Fund. Previously, she worked with Wikipedia communities on open culture and software, led product strategy at her podcasting platform TapeWrite, and founded the Berlin Chapter of Zebras Unite to promote ethical startup practices. She holds a background in law, political science, and product management.

Vidisha Mishra leads programmatic and community engagement at GSI, focusing on partnerships with the G20, G7, and related multilateral groups. She has over a decade of experience in policy advice, program management, and strategic engagement across global think tanks. Previously, she worked at UNU-IIGH on global health partnerships and led research on gender in economic diplomacy at ORF. Vidisha contributes to multilateral processes like W20 and T20 and has been recognized through fellowships in foreign policy. She holds degrees from LSE and Warwick and is pursuing an Executive MPA at the Hertie School, specializing in digitization and big data.

Anshul Pachouri is a Global Sector Lead for Technology Policy and Innovation at the MSC's Centre for Responsible Technologies. He works at the intersection of AI, public policy, DPI (Digital public infrastructure), and innovation. He has led many technology policy and DPI engagements with national governments, digital ID authorities, and donors across Africa and Asia. He also co-designed and co-lead the Digital ID Hackathon Africa in partnership with Carnegie Mellon University Africa Upanzi Network. He has written extensively on AI policy and regulation for multilateral forms such as Think 7 Italy, UN ST&I Forum, and Global Partnership on AI (GPAI). He has also presented and published his papers at leading academic institutions such as Harvard Kennedy School, National University of Singapore etc. and multilateral institutions such as Asian Development Bank Institute.

Arnav Nigam is an associate in the Center for Responsible Technologies, Microsave Consulting (MSC). His areas of interest include digital public infrastructure, technology policy, data governance and human centred design. With MSC, he led and worked on several projects, including DPI for AI strategy in LMICS, digital ID hackathons in Eastern, Western, Northern and Southern Africa - aimed to build local capacity and develop new use cases in the digital ID ecosystem, Aadhaar Innovation Sandbox and Aadhaar non-personal data sharing policy. He holds a master's in public policy and bachelor's in computer science and engineering.

References

Baack, Stefan, Stella Biderman, Kasia Odrozek, Aviya Skowron, Ayah Bdeir, Jillian Bommarito, Jennifer Ding, et al. 2025. "Towards Best Practices for Open Datasets for LLM Training." arXiv, January 14, 2025. https://arxiv.org/abs/2501.08365.

Vincent, James. "OpenAI's GPT-4: How Closed AI Research Is Shaping the Future." The Verge, March 15, 2023. https://www.theverge.com/2023/3/15/23640180/openai-gpt-4-launch-closed-research-ilya-sutskever-interview.

Grynbaum, Michael, and Ryan Mac. "The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work." The New York Times, December 27, 2023.

Edmond, Charlotte. "What Is Open-Source AI and How Could DeepSeek Change the Industry?" World Economic Forum, February 5, 2025. https://www.weforum.org/stories/2025/02/open-source-ai-innovation-deepseek/.

Coffey, Donavyn. "Māori Are Trying to Save Their Language from Big Tech." WIRED, April 28, 2021. https://www.wired.com/story/maori-language-tech/.

Mozilla. Public AI: Governing AI as a Public Good. Mozilla Foundation, 2024. https://assets.mofoprod.net/network/documents/Public_AI_Mozilla.pdf.

African Union. 2024. "Continental Artificial Intelligence Strategy." https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy.

NITI Ayog, Government of India. 2018. "National Strategy for Artificial Intelligence." https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf.

UNESCO. 2021. "Recommendation on the Ethics of Artificial Intelligence." https://unesdoc.unesco.org/ark:/48223/pf0000381137.

GPAI. 2023. Responsible AI Working Group Report. https://gpai.ai/projects/responsible-ai/Responsible%20AI%20WG%20Report%202023.pdf.

GPAI. 2023. "Innovation and Commercialization Working Group Report." https://gpai.ai/projects/innovation-and-commercialisation/2023%20IC%20WG%20Report%20VF%20(002).pdf.

National Language Translation Mission. n.d. Bhashini. Accessed February 2025. https://bhashini.gov.in/.

"Previsia." n.d. Accessed February 2025. https://previsia.org.br/.

G7. 2023. “Hiroshima Process International Code of Conduct for Organizations Developing Advanced AI Systems.” https://transparency.oecd.ai/.

© 2025 Katarzyna Odrozek, Vidisha Mishra, Anshul Pachouri and Arnav Nigam