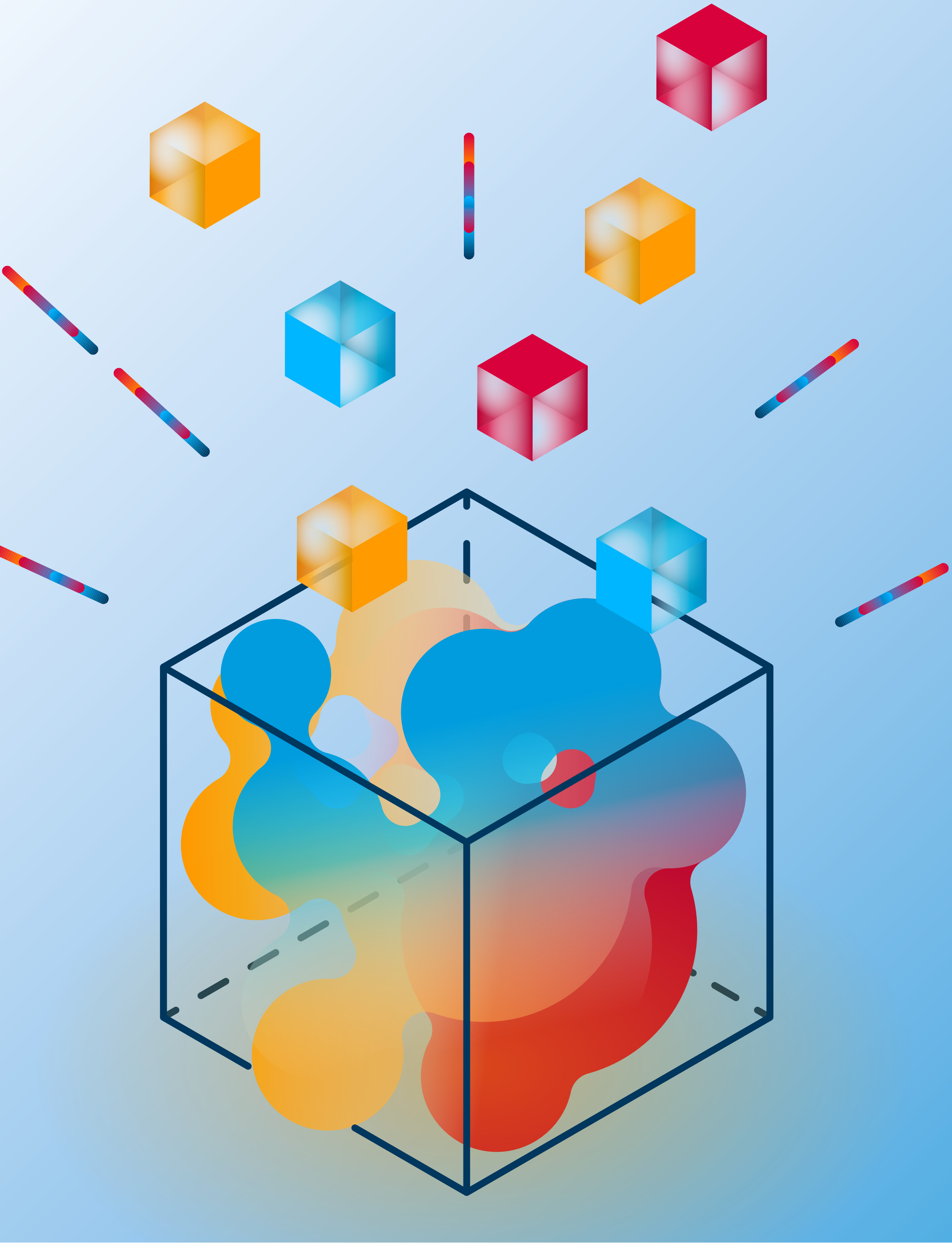
Public AI – White Paper



Supported by

OPEN
_FUTURE

Commissioned by

BertelsmannStiftung

Legal notice

Commissioned by
© Bertelsmann Stiftung, Gütersloh
May 2025

Publisher
Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Phone +49 5241 81-0
www.bertelsmann-stiftung.de

Supported by
Open Future

Authors
Dr. Felix Sieker, Bertelsmann Stiftung
Dr. Alek Tarkowski, Open Future
Lea Gimpel, Digital Public Goods Alliance
Dr. Cailean Osborne, Oxford Internet Institute

Responsible
Dr. Felix Sieker, Bertelsmann Stiftung

Editing
Barbara Serfozo, Berlin

Infographics
Jakub Koźniewski

Layout and Typesetting
Nicole Meyerholz, Bielefeld

Recommended citation style
Sieker/Tarkowski/Gimpel/Osborne (2025). Public AI – White Paper. Bertelsmann Stiftung. Gütersloh.

DOI 10.11586/2025040

Public AI – White Paper

Dr. Felix Sieker,
Dr. Alek Tarkowski,
Lea Gimpel,
Dr. Cailean Osborne

Reviewer list

Albert Cañigueral, Barcelona Supercomputing Center
Amin Oueslati, The Future Society
Ben Burtenshaw, Hugging Face
Huw Roberts, Oxford Internet Institute, University of Oxford
Isabel Hou, Taiwan AI Academy
Jakob Mökander, Digital Ethics Center, Yale University
Jennifer Ding, Boundary Object Studio
Laura Galindo, AI policy expert
Luca Cominassi, AI policy expert
Martin Hullin, Bertelsmann Stiftung
Martin Pompéry, SINE Foundation
Marta Ziosi, Oxford Martin AI Governance Initiative, University of Oxford
Paul Keller, Open Future
Paul Sharratt, Sovereign Tech Agency
Ravi Iyer, USC Marshall
Yacine Jernite, Hugging Face
Zoe Hawkins, Tech Policy Design Institute

Supported by

OPEN
_FUTURE

Commissioned by

BertelsmannStiftung

Table of contents

Preface 6

Executive summary 8

Glossary 11

1 | Introduction 13

2 | Technical primer: What are AI technologies and how do they work? 17

Defining artificial intelligence 17

The deep learning paradigm 21

“Attention is all you need”: Transformers and the rise of generative AI 22

The generative AI development process 24

Pretraining phase 24

Post-training phase 26

Deployment 26

AI scaling laws: The contested future of AI 26

What are AI scaling laws? 27

The evolution of AI scaling laws 27

Scaling and AI’s environmental footprint 28

What is the future of AI scaling laws? 29

3 | The generative AI stack 32

Overview of the AI stack 32

Advantages of the AI stack concept 33

Concentrations of power in the AI stack 35

Characteristics of key layers of the stack 37

Compute 37

Data 39

Models 42

4 | The public AI framework 45
The concept of public digital infrastructure 45
Public, private and civic actors in public digital infrastructure 47
Proposals for public AI 47
Public AI Network 47
Mozilla Foundation 48
Vanderbilt Policy Generator 49
Defining public AI infrastructure 49
Gradient of publicness of AI systems 52
Goals and governance principles of public AI policies 55
Governance of public AI 56

5 | AI strategy and three pathways to public AI 58
Elements of a public AI strategy 58
The public AI ecosystem and its orchestrating institution 60
Three pathways toward public AI infrastructure: compute, data and model 61
Compute pathway to public AI 61
Compute: bottlenecks 61
Compute: opportunities 63
Data pathway to public AI 65
Data: bottlenecks 65
Data: opportunities 66
Model pathway to public AI 68
Models: bottlenecks 68
Models: opportunities 69
Additional measures 71
Coda: mission-driven public AI policy 73

Preface

Artificial Intelligence stands at a pivotal crossroads. While its potential to transform society is immense, the power to shape its trajectory is becoming increasingly concentrated. Today, a small number of dominant technology firms hold sway not only over the most advanced AI models but also the foundational infrastructure – compute capacity, data resources and cloud platforms – that makes these systems possible. This consolidation of influence represents more than a market imbalance; it poses a direct threat to the principles of openness, transparency and democratic accountability.

When only a handful of actors define how AI systems are built and used, public oversight erodes. These systems increasingly reflect the values and economic incentives of their creators, often at the expense of inclusion, accountability and democratic oversight. Without intervention, these trends risk entrenching structural inequities and shrinking the space for alternative approaches.

This white paper outlines a strategic countervision: Public AI. It proposes a model of AI development and deployment grounded in transparency, democratic governance and open access to critical infrastructure. Public AI refers to systems that are accountable to the public, where foundational resources such as compute, data and models are openly accessible and every initiative serves a clearly defined public purpose.

Grounded in a realistic analysis of the constraints across the AI stack – compute, data and models – the paper translates the concept of Public AI into a concrete policy framework with actionable steps. Central to this framework is the conviction that public AI strategies must ensure the continued availability of at least one fully open-source model with capabilities approaching those of proprietary state-of-the-art systems. Achieving this goal requires three key actions: coordinated investing in the open-source ecosystem, providing public compute infrastructure, and building a robust talent base and institutional capacity.

It calls for the continued existence of at least one fully open-source model near the frontier of capability and lays out three imperatives to achieve this: strengthening open-source ecosystems, investing in public compute infrastructure, and building the talent base to develop and use open models.

To guide implementation, the paper introduces the concept of a "gradient of publicness" to AI policy – a tool for assessing and shaping AI initiatives based on their openness, governance structures, and alignment with public values. This framework enables policymakers to evaluate where a given initiative falls on the spectrum from private to public and to identify actionable steps to increase public benefit.

We extend our sincere thanks to Alek Tarkowski, Lea Gimpel and Cailean Osborne for their valuable insights and contributions to this work.

As you engage with the ideas presented here, we invite you to consider how this vision can inform your own decision-making and inspire policies that are both inclusive and forward-looking. Together, let us harness AI to keep it from deepening divisions while ensuring it broadens democratic possibility and strengthens social solidarity.

Dr. Felix Sieker
Project Manager
Digitization and the Common Good
Bertelsmann Stiftung

Martin Hullin
Director
Digitization and the Common Good
Bertelsmann Stiftung

Executive summary

Today's most advanced AI systems and foundation models are largely proprietary and controlled by a small number of companies. There is a striking lack of viable public or open alternatives. This gap means that cutting-edge AI remains in the hands of a select few, with limited orientation toward the public interest, accountability or oversight.

Public AI is a vision of AI systems that are meaningful alternatives to the status quo. In order to serve the public interest, they are developed under transparent governance, with public accountability, equitable access to core components (such as data and models), and a clear focus on public-purpose functions.

In practice, public AI projects ensure that the public has both insight into and influence over how AI systems are built and used. They aim to make the key building blocks – data, open source software and open models – accessible to all on fair terms. Crucially, public AI initiatives are oriented toward broad societal benefit, rather than private gain.

Over the past year, momentum behind Public AI proposals has been steadily growing, with a series of influential reports and initiatives by Public AI Network, Mozilla and the Vanderbilt Policy Generator demonstrating the importance of this approach. And even more importantly, various initiatives are developing various components and whole AI systems that fulfill this vision of Public AI.

This white paper builds on earlier proposals for Public AI and is aimed at policymakers and funders, with the goal of helping to turn the vision of Public AI into reality. In particular, it advances this timely conversation by making the following two novel contributions.

A vision for Public AI grounded in the reality of the AI stack

A vision for public AI needs to take into account today's constraints at the compute, data and model layers of the AI stack, and offer actionable steps to overcome these limitations. This white paper offers a clear overview of AI systems and infrastructures conceptualized as a stack of interdependent elements, with compute, data and models as its core layers. It also identifies critical bottlenecks and dependencies in today's AI ecosystem, where dependency on dominant or even monopolistic commercial solutions constrains development of public alternatives. It highlights the need for policy approaches that can orchestrate resources and various actors across layers, rather than attempting complete vertical integration of a publicly owned solution.

To achieve this, it proposes three core policy recommendations:

1. Develop and/or strengthen fully open source models and the broader open source ecosystem

2. Provide public compute infrastructure to support the development and use of open models

3. Scale investments in AI capabilities to ensure that sufficient talent is developing and adopting these models

In order to achieve this, complementary pathways for Public AI development need to be pursued, focused on the three core layers of the AI stack: compute, data, and models:

1. **Compute Pathway:** It focuses on providing strategic public computing resources, particularly supporting open-source AI development. Key recommendations include ensuring computing access for fully open projects, expanding compute for research institutions, and improving coordination between public compute initiatives.
2. **Data Pathway:** It emphasizes creating high-quality datasets as digital public goods through commons-based governance. This includes developing datasets as publicly accessible resources while protecting against value extraction, and establishing public data commons with appropriate governance mechanisms.
3. **Model Pathway:** It centers on fostering an ecosystem of fully open source AI models, including both a state-of-the-art "capstone model" and specialized smaller models. The strategy emphasizes building sustainable open source AI development capabilities rather than simply competing with commercial labs.

Several additional measures are highlighted that do not fit within one of the three pathways, but help secure key public interest goals. This includes investing in AI talent and capabilities to develop and deploy AI systems in the public interest, supporting paradigm-shifting innovation toward more efficient technologies, funding open-source software and tools, and building effective deployment pathways for public AI applications.

This approach acknowledges the importance of the various layers and different paths that can be pursued to attain Public AI. It also argues for coordinated interventions across the entire AI stack, orchestrated by new public institutions capable of managing decentralized AI development ecosystems.

The "gradient of publicness": A framework for Public AI

The white paper also offers a "gradient of publicness" framework, rooted in Public Digital Infrastructure principles. This framework can guide decision-making around investments in AI infrastructure and help increase public value while acknowledging existing constraints and limitations to building fully Public AI.

This framework maps AI interventions along a continuum – from fully public to fully private – based on their attributes (e.g. accessibility, openness, interoperability), functions (e.g. enabling social or economic goals) and modes of control (e.g. democratic governance and accountability). It serves as both a diagnostic and strategic tool for assessing where an intervention falls along this continuum, and for identifying interventions that could strengthen its public value.

The gradient of publicness consists of the following six distinct levels, each representing different degrees of public attributes, functions, and control:

Level 1: Commercial provision of AI components with public attributes
Commercial entities develop and share open source components (e.g., Meta's open-sourcing of PyTorch or Llama) with high public accessibility but limited public function and control.

Level 2: Commercial AI infrastructure with public attributes and functions
Privately controlled platforms like Hugging Face Hub that democratize access to AI tools while maintaining commercial oversight but serving public interest goals.

Level 3: Public computing infrastructure
Government-funded supercomputers and data centers (e.g., EU AI Factories) that provide computing resources through public-private partnerships with moderate to high public control.

Level 4: Public provision of AI components
Publicly funded datasets, benchmarks, and tools (e.g., Mozilla's Common Voice) developed specifically as digital public goods with high public control and clear public functions.

Level 5: Full-stack public AI infrastructure built with commercial compute
AI systems like the OLMo model by Allen Institute for AI that are fully open source but rely on commercial computing infrastructure, limiting public control at the compute layer.

Level 6: Full-stack public AI infrastructure
Fully autonomous public AI systems like Spain's Alia, built with public data, models, and computing infrastructure, achieving the highest level of publicness across all layers.

Reading Guide

We encourage readers to explore the full report to gain a comprehensive understanding of the public AI vision and its implications. Depending on your specific interests, we recommend the following entry points.

If you are interested in the technological foundations of AI and the factors contributing to the rise of generative models:

- Begin with the Introduction, then read Chapter 2. These sections provide a technical primer on AI technologies and the key breakthroughs that have led to today's generative AI models.

If you are a policymaker working on AI policy and/or (investment) strategy:

- Start with the Introduction, then read chapter 3 for an overview of the AI stack (compute, data and models). Focus on the "gradient of publicness" in chapter 4, and turn to chapter 5 for an overview of pathways and recommendations for building public AI.

If you are a policymaker or funder seeking concrete policy or funding guidance:

- Begin with the Introduction, then focus on chapter 4 for the gradient of publicness framework and chapter 5 for specific policy recommendations.

Glossary

AI model

A mathematical and computational system trained to recognize patterns or make decisions based on input data.

AI scaling laws

An observed pattern showing that AI model performance improves predictably with more data, larger models (more parameters) and more compute resources.

AI system

A machine-based system that infers how to generate outputs – like predictions, content recommendations, or decisions – to influence physical or virtual environments. It combines models, data, infrastructure and interfaces and can vary in autonomy and adaptiveness.

Artificial general intelligence

A theoretical form of AI capable of understanding, learning and performing any intellectual task that a human can do.

Artificial intelligence

The broad field of computer science focused on creating machines or software capable of tasks that normally require human intelligence (e.g., perception, reasoning, decision-making).

Computer vision

A branch of AI that enables computers to interpret and process visual information from the world, such as images or video.

Deep learning

A subset of machine learning that uses multi-layered neural networks (often with many hidden layers) to automatically learn complex representations of data.

Distillation

A technique where a smaller "student" model is trained to replicate the behavior of a larger "teacher" model, enabling similar performance with lower computational requirements.

Fine-tuning

The process of adapting a pretrained model to specific tasks through additional training on task-specific datasets, preserving general knowledge while optimizing for particular applications.

Foundation model

A large AI model trained on broad datasets and designed for adaptability across many tasks. Foundation models can be unimodal or multimodal; large language models are a subset.

Generative AI

A class of AI systems designed to create new content – such as text, images or music – based on learned patterns from training data.

Graphics processing unit (GPU)

A specialized processor designed for handling parallel computations, originally developed for video games, now crucial for efficiently training and running deep learning models.

Large language model (LLM)

A type of foundation model trained on massive text corpora, capable of generating or understanding natural language, often containing billions of parameters.

Machine learning

A branch of AI where systems improve their performance on tasks over time by learning from data including methods like supervised, unsupervised and reinforcement learning.

Model parameters or weights

Numeric values within a model (e.g., connection strengths in a neural network) that are adjusted during training to enable the model to perform its tasks.

Moore's law

A prediction made by Intel co-founder Gordon Moore in 1965 that the number of transistors on a chip would double approximately every two years, leading to exponential growth in computing power.

Natural language processing (NLP)

The field at the intersection of AI and linguistics focused on enabling computers to understand, interpret and generate human language.

Neural network

A type of AI model inspired by the brain's interconnected neurons, designed to recognize patterns and relationships in data.

Open model

In this white paper, open models refer to AI models for which the model architecture, trained parameters (i.e., weights and biases) and some documentation are released under open-source licenses.

Open source license

An open source license is a legal agreement that allows users to freely use, modify, and distribute software or other works, while specifying certain conditions that must be met when using or redistributing the software.

Open source model

There are competing definitions of open-source models. In this white paper, we define an open-source model as an AI model whose trained parameters (i.e., weights and biases) – referred to as an open model – are released alongside the code, data and accompanying documentation used in its development, all under free and open-source licenses.

Public AI

Public AI refers to AI systems developed with transparent governance, public accountability, open and equitable access to core components, such as data and models, and clearly defined public functions at their core.

Reinforcement learning

Reinforcement learning is a machine learning paradigm where a system learns by interacting with an environment and improving through rewards. Variants reinforcement learning from human feedback (RLHF) and reinforcement learning from AI-generated feedback (RLAIF) are used to align AI models with desired behaviors.

Quantization

A technique that reduces the computational and memory costs of AI models by representing weights and activations with lower-precision data types, typically without requiring additional training data.

Small model

A deliberately compact AI model with millions to billions of parameters, optimized for resource efficiency and often adapted from larger models through distillation or quantization.

Transistor

A tiny semiconductor device that acts as a switch or amplifier in electronic circuits, forming the fundamental building block of virtually all modern electronic devices.

Transformers

A deep learning architecture introduced in 2017, based on attention mechanisms, that underpins the recent surge in AI advancements by enabling models to process and generate sequence data with unprecedented effectiveness.

1 | Introduction

Artificial Intelligence (AI) stands as one of the most prominent and potentially transformative technologies of the past decade. With the rapid ascent of new industry leaders like OpenAI and Anthropic, alongside a strategic pivot toward AI by incumbents such as Microsoft, Google and Meta, AI is increasingly shaping the future of sectors ranging from healthcare and finance to education and government. At the same time, critics warn that AI may be yet another hype-driven, extractive and unsustainable technology lacking a clear social purpose.

Public debate often swings between uncritical enthusiasm for AI and deep concern over its existential risks. Despite these polarized narratives, the underlying reality is that AI technologies are becoming deeply embedded in social, economic and political systems. Hype or not, AI is likely to remain a lasting force – making it all the more urgent to shape its development around clear public values.

In late 2022, just as commercial AI labs were launching the first public-facing applications based on generative AI models, Mariana Mazzucato and Gabriela Ramos published an op-ed arguing for public policies and institutions "designed to ensure that innovations in AI are improving the world." They warned, instead, that a new generation of digital technologies is being "deployed in a [policy] vacuum." According to Mazzucato and Ramos, public interventions are essential to steer the technological revolution in a direction that turns technical innovation into outcomes that serve the public interest.[1]

Emerging concentrations of power

Three years on, concentrations of power in AI have only deepened. A small group of dominant technology companies now control not only most state-of-the-art AI models, but also the foundational infrastructure that shapes the field. This emerging AI oligopoly builds on existing monopolies in cloud computing and digital platforms, reinforcing the dominance of hyperscalers and platform giants.

This concentration is not merely structural – it has far-reaching consequences. When a handful of actors define how AI is built and deployed, the benefits are captured by the few, while the risks are borne by the public. AI systems increasingly reflect the values, incentives and worldviews of their creators, often at the expense of public inclusion, accountability and democratic oversight.

Compounding the issue is the rapid pace of AI development, characterized by two destabilizing trends. First, corporate competition has triggered a relentless race toward ever more capable AI models, often framed as a pursuit of so-called "superintelligence." As a result, billions are being invested in systems designed to achieve commercial dominance, often with little regard for societal benefit or safety. These investments in compute infrastructure frequently lack a clear articulation of the public needs they are meant to address, and can have significant environmental costs.

Second, geopolitical shifts are pushing states to treat AI as a zero-sum game. Governments are engaging in AI nationalism – walling off innovation behind borders in the name of digital sovereignty – rather

[1] Mariana Mazzucato and Gabriela Ramos. "AI in the Common Interest." Project Syndicate, 26 Dec. 2022. https://www.project-syndicate.org/commentary/ethical-ai-requires-state-regulatory-frameworks-capacity-building-by-gabriela-ramos-and-mariana-mazzucato-2022-12

than fostering international cooperation. Sovereign AI strategies often mirror the priorities of dominant commercial actors, focusing heavily on large-scale investments in compute infrastructure under the assumption that public benefits will follow. However, this approach risks entrenching existing power asymmetries and deepening dependencies on the providers of critical AI components.

As the AI Now Institute has observed, current AI policy tends to "conflate public benefit with private sector success,"[2] leading to strategies that prioritize industrial competitiveness over accountability, transparency or equitable access. Without a course correction, this trajectory could lock in structural inequities and limit democratic control over foundational AI systems.

The need for a countervision: Public AI

In response to this growing imbalance, interest is building around alternatives to the dominant commercial AI paradigm – a concept broadly described as public AI. While the definition of public AI remains fluid, it generally refers to AI systems developed with transparent governance, public accountability, open and equitable access to core components such as data and models, and clearly defined public-purpose functions.

A central pillar of public AI is the development of fully open source models, which involves releasing the trained parameters (i.e., weights and biases) of models – commonly referred to as open models – along with the code, data and documentation used in their creation under open source licenses. Open models ensure that foundational AI systems are accessible, inspectable and modifiable by a wide range of actors, including researchers, public institutions and civil society. Unlike proprietary models, fully open source models offer transparency into model weights and training data, allowing for reproducibility and independent auditing. They also lower barriers to experimentation and adaptation, especially for academic researchers, startups and non-commercial applications.

While open source models do not eliminate market concentration, they can reduce dependency on a few dominant industry players and allow public interest applications to be developed without restrictive licenses or opaque constraints. In this way, open source AI serves as critical infrastructure for aligning AI development with democratic values and public goals.

Moreover, visions for public AI emphasize the need to strengthen the broader ecosystem of AI infrastructure and tools. Without a sustainable open source AI ecosystem, efforts to keep AI development transparent and aligned with the public interest will remain fragile and limited.

Two urgent realities demand immediate action. First, the window for intervention is closing. As a handful of corporate and state actors consolidate control over critical infrastructure and resources – such as compute, datasets and talent – the cost of building alternatives continues to rise. Second, the stakes are global. AI's impact on labor, healthcare, education, the environment and democracy transcends borders. Without proactive measures, its benefits will accrue to a privileged minority, while the risks – from disinformation to algorithmic bias – will disproportionately affect marginalized groups.

[2] Amba Kak and Sarah Myers West "2023 Landscape. Confronting tech power." AI Now Institute. 2023. https://ainowinstitute.org/wp-content/uploads/2023/04/AI-Now-2023-Landscape-Report-FINAL.pdf

INFOBOX | Open source AI and degrees of openness

The development of publicly controlled, state-of-the-art open source AI models should be a central goal of public AI policies. These models provide a public alternative to commercial offerings and help foster an ecosystem in which smaller, complementary models can thrive.

The benefits and risks of open source AI development have been the subject of intense policy debate in recent years. While early discussions focused largely on the safety risks of open source AI systems, there is growing support for such models and increasing recognition of their potential to democratize AI development.

However, the term "open source AI" is currently used to describe models with varying degrees of openness. In some cases, "open washing" occurs – when models that are not truly open are marketed as meeting open source standards.[3] Public AI policy must therefore be grounded in a clear and consistent definition of what constitutes open source AI. The Open source AI Definition Initiative proposed a binary definition: AI systems released under licenses that permit use, study, modification and redistribution. This definition, however, has not yet been widely adopted.[4] Alternatively, Irene Solaiman of Hugging Face has conceptualized openness as a gradient – from fully closed to fully open.

In this white paper, we focus on open models – that is, models in which both the architecture and parameters (i.e. weights and biases) are released under permissive licenses. It is important to note, however, that AI models are made up of several components beyond architecture and weights. For example, the Model Openness Framework[5] breaks models down into 16 components spanning code, data and documentation, and outlines three tiers of model openness depending on how fully these components are shared under permissive licenses.

Today, many models described as "open" or "open source" share only a limited set of components. In most cases, only model parameters are released, accompanied by some documentation. While this allows for reuse, it offers insufficient transparency into the training data and the development process.[6] These models are more accurately described as open-weight models.[7] In this white paper, we use the generic term open model to refer to a broad spectrum of models released under open terms, including commercial open-weight models from companies such as Mistral or DeepSeek.

At the same time, a key recommendation of this white paper is the development of fully open source AI models, which we define as the complete release of model parameters, architecture, code, datasets and associated documentation under open source licenses. Currently, very few state-of-the-art models meet this definition – among them, OLMo 2 by the Allen Institute for AI.

[3] Sarah Kessler. "Openwashing." New York Times. 19 May 2024. https://www.nytimes.com/2024/05/17/business/what-is-openwashing-ai.html

[4] Open Source Initiative. "Open Source AI Definition." Accessed 27 April 2025. https://opensource.org/ai

[5] Model Openess Framework. https://isitopen.ai/ Accessed 27 April 2025.

[6] Zuzanna Warso, Paul Keller and Max Gahntz. "Towards Robust Training Data Transparency." Open Future. 19 June 2024. https://openfuture.eu/publication/towards-robust-training-data-transparency/

[7] "Open weights: not quite what you've been told." Open Source Initiative. https://opensource.org/ai/open-weights Accessed 23 April 2025.

Why this report?

This white paper serves as a guide for policymakers and funders at a critical juncture. It aims to:

- **Demystify AI:** By breaking down generative AI technology into a “stack” comprising **compute, data and models** layers, we cut through the hype, make its complexities more graspable and expose underlying power imbalances.
- **Identify bottlenecks and dependencies:** We identify choke points in today’s AI ecosystem that reinforce concentration, from GPU shortages to data monopolies.
- **Outline pathways to public AI:** We propose a vision of AI as public digital infrastructure and spotlight policy and funding interventions that could enable a public AI ecosystem to emerge.

In chapter 2, we offer a technical primer that defines AI and distinguishes key types of AI technologies, with a focus on machine learning and generative AI. We also describe the dominant development paradigm for generative AI – the transformer architecture – and explain the scaling laws that underpin it.

In chapter 3, we define the generative AI stack and provide an overview of its three key layers: compute, data and models. These layers shape the conditions for building public AI solutions. We pay particular attention to the concentration of power at each layer and the implications this has for public AI efforts.

In chapter 4, we introduce a framework for public AI grounded in the concept of public digital infrastructure. Drawing on existing proposals, we analyze the attributes, functions and ownership structures of components in public AI systems. We then introduce a gradient of publicness for AI solutions, which accounts for varying degrees of dependency, especially at the compute layer. This chapter also sets out governance principles for public AI.

In chapter 5, we outline three pathways to public AI – starting at the compute, data and model layers respectively. For each layer, we identify key bottlenecks and propose solutions that advance the vision of public AI. The chapter concludes with strategic recommendations to connect these pathways into a cohesive public AI agenda.

2 | Technical primer: What are AI technologies and how do they work?

A lack of precision around AI technologies and the technical aspects of their development remains a major limitation in current policy debates.

Understanding AI – including both the technical and economic dimensions of its development – is essential for crafting targeted, effective and efficient policy interventions. Yet there is often little clarity in discussions about how these technologies are developed, including the resources required and the scientific or engineering breakthroughs that have made this development possible. The term "AI" is frequently used to refer to a wide array of technologies with very different characteristics, without distinguishing clearly between them.

Greater clarity is needed to design realistic pathways for implementing public AI strategies. This section outlines key technical aspects of AI development and lays the groundwork for the analysis in the following chapters, which explore how to create public infrastructures and solutions in the AI space, with a particular focus on generative AI.

To this end, the section provides a technical explainer of AI technologies. It includes a basic typology, an overview of the transformer paradigm – the dominant development model for generative AI – and a discussion of AI scaling laws, which help explain why generative AI requires such vast amounts of compute. These scaling laws are a critical factor in determining which actors are able to develop generative AI and help explain why creating truly public AI remains so challenging.

Defining artificial intelligence

Broadly speaking, AI research and development is concerned with building computer systems capable of performing tasks that typically require human intelligence.[8] The origins of the term are attributed to an academic workshop held in 1956 at Dartmouth College in the United States.[9] Since then, the field has gone through multiple cycles of optimism and disillusionment, with various approaches to training machines to replicate or approximate human cognitive tasks. As the field has evolved, so too have definitions of AI – often shifting in response to changing capabilities and expectations. What once marked the cutting-edge of AI, such as speech or image recognition, has now become routine in many applications.

A widely accepted definition comes from the OECD, which describes an AI system 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. AI systems vary in their levels of autonomy and adaptiveness after deployment."[10] This definition has been adopted by numerous govern-

[8] Stuart Russell and Peter Norvig. "Artificial Intelligence: A Modern Approach, 4th US Ed." (Berkeley: University of Berkeley, 2022).

[9] Haroon Sheikh, Corien Prins, and Erik Schrijvers, "Artificial Intelligence: Definition and Background," in Mission AI, Research for Policy (Cham: Springer, 2023), 15–41, https://doi.org/10.1007/978-3-031-21448-6_2

[10] "Explanatory memorandum on the updated OECD definition of an AI system." OECD, 2024. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/03/explanatory-memorandum-on-the-updated-oecd-definition-of-an-ai-system_3c815e51/623da898-en.pdf. Accessed 23 April 2025.

Figure 1 | AI terminology

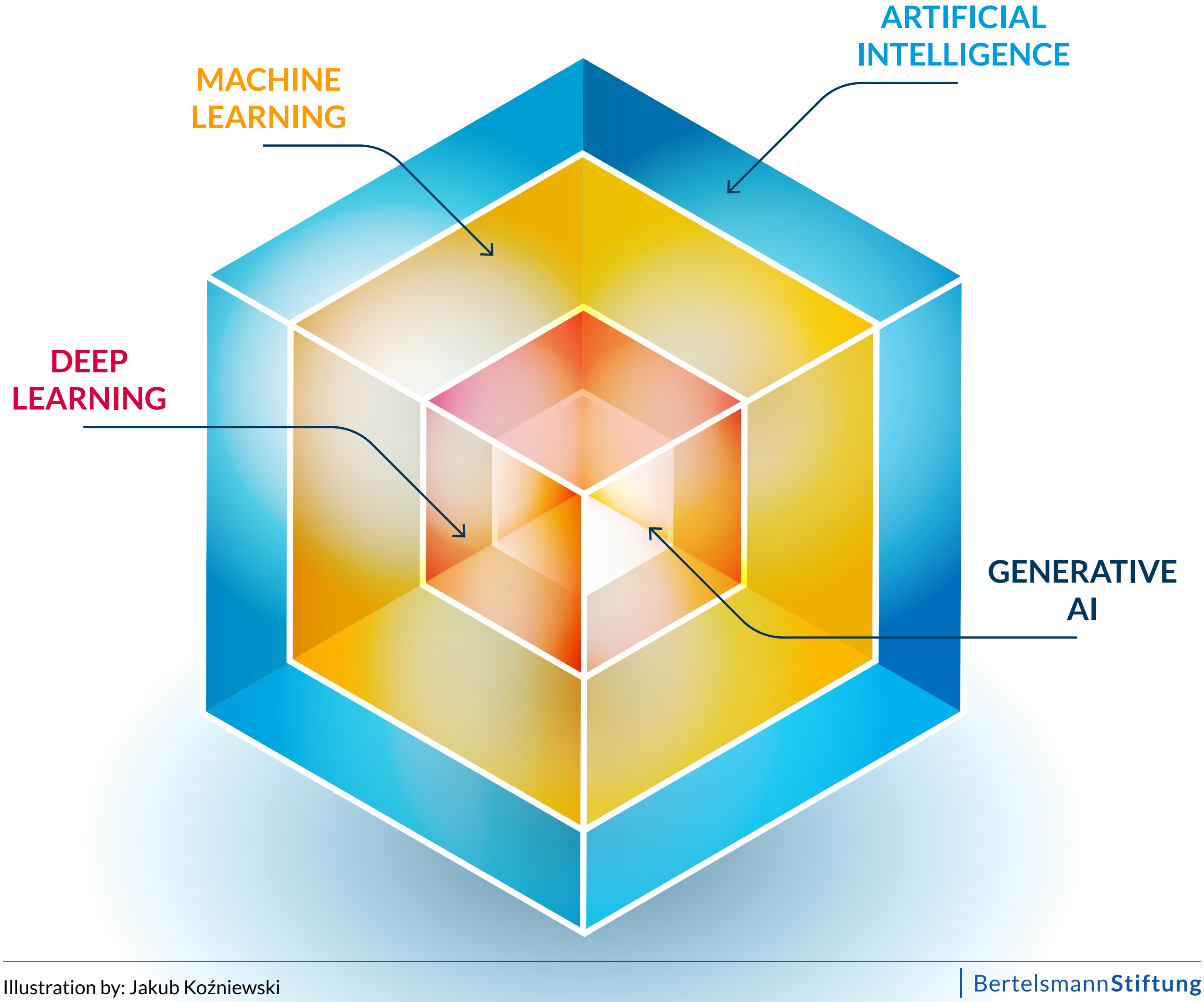


Illustration by: Jakub Koźniewski

BertelsmannStiftung

ments and international institutions in in the development of AI governance frameworks.[11]

While we adopt the OECD's definition as a general framing of AI, this report focuses on machine learning (ML) approaches to AI, and in particular the contemporary paradigm of generative AI. In this section, we define these technical concepts before discussing them in greater detail in the sections that follow.

Machine learning (ML) is central to contemporary AI and refers to the development of computer programs – specifically, statistical models – that learn from data rather than relying on explicitly programmed instructions. Tom Mitchell defines ML as follows: "a computer program is said to learn from experience (E) with respect to some class of tasks (T) and performance measure (P), if its performance at tasks in T, as measured by P, improves with experience E."[12] ML encompasses a variety of statistical learning ap-

[11] CEIMIA. "A Comparative Framework for AI Regulatory Policy." https://ceimia.org/en/projet/a-comparative-framework-for-ai-regulatory-policy/ Accessed 23 April 2025.

[12] Tom Mitchell. "Machine Learning." (McGraw Hill, 1997). https://www.cs.cmu.edu/~tom/mlbook.html

proaches, including supervised learning (learning from labeled data), unsupervised learning (identifying patterns in unlabeled data), and reinforcement learning (learning through interaction with an environment).

Deep learning is a branch of ML that uses artificial neural networks with multiple layers – hence "deep" – to detect patterns in raw data.[13] deep learning is especially effective for handling unstructured data such as text or images, making it popular in fields like natural language processing and computer vision. It also plays a key role in generative AI applications, as described below.

Generative AI refers to the application of deep learning techniques to build models that, given an input – typically a natural language prompt – can generate novel outputs such as text, images, audio or code, without being explicitly programmed for each task.[14] Recent advances in generative AI have been driven by the emergence of transformer-based architectures and scaling laws, which we explain in more detail below, as well as the rise of user-facing tools, such as OpenAI's ChatGPT.[15] Generative AI includes both unimodal models, such as language or vision models, and multimodal models, which can process and generate multiple types of data – such as text, images or audio – either as inputs or outputs.

Foundation models represent a major category within generative AI. They are characterized by their large scale (with up to trillions of parameters), training on vast and diverse datasets and adaptability to a wide range of downstream applications.[16] Foundation models can be either unimodal or multimodal. For example, OpenAI's GPT-3 is a unimodal foundation model focused solely on text, while GPT-4o is a multimodal model capable of processing and generating text, audio, images and video. While the terms "large language model" (LLM) and "foundation model" are often used interchangeably, LLMs are technically a subset of foundation models that specialize in language processing. Foundation models more broadly include systems focused on other modalities, such as vision or audio, or combinations of them.

Although generative AI is currently receiving significant attention – with high-profile tools like ChatGPT and Claude capturing public attention – it is important to note that the most common AI use cases still rely on more traditional ML approaches, such as regression models, random forests and clustering algorithms. According to the AI Mapping 2025 report, which studied 750 French AI startups, ML remains the most widely used AI technique (28%), followed by deep learning (20%) and generative AI (15%).[17] The relative popularity of such ML approaches is reflected in the download statistics of widely used open source software libraries for AI. For example, scikit-learn, a Python library that implements ML algorithms and known as "the Swiss army knife for ML," is downloaded up to 3 million times per day.[18, 19] It is followed by PyTorch, a popular framework for training deep learning models, and transformers, a library for accessing and fine-tuning models hosted on Hugging Face Hub, which are both downloaded up to 1.5 million times per day.[20, 21]

[13] Yann LeCun, et al. "Deep Learning." Nature, vol. 521, no. 7553, May 2015, pp. 436–44. www.nature.com, https://doi.org/10.1038/nature14539 Accessed 3 April 2025.

[14] McKinsey. What Is ChatGPT, DALL-E, and Generative AI? https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-generative-ai Accessed 23 April 2025.

[15] OpenAI. Introducing ChatGPT. https://openai.com/index/chatgpt/ Accessed 13 March 2024.

[16] Rishi Bommasani, et al. On the Opportunities and Risks of Foundation Models. arXiv:2108.07258, arXiv, 12 July 2022. arXiv.org, https://doi.org/10.48550/arXiv.2108.07258

[17] France Digitale. "Mapping Des Startups Françaises de l'IA." https://mappings.francedigitale.org/ia-2025 Accessed 23 April 2025.

[18] PyPI Stats. "scikit-learn." Accessed 23 April 2025. https://pypistats.org/packages/scikit-learn

[19] Inria. "The 2019 Inria-French Academy of Sciences-Dassault Systèmes Innovation Prize: Scikit-Learn, a Success Story for Machine Learning Free Software." https://www.inria.fr/en/2019-inria-french-academy-sciences-dassault-systemes-innovation-prize-scikit-learn-success-story Accessed 23 April 2025.

[20] PyPI Stats. "transformers." https://pypistats.org/packages/transformers Accessed 23 April 2025.

[21] PyPI Stats. "torch." https://pypistats.org/packages/torch Accessed 23 April 2025.

Generative AI models are increasingly seen as potential general-purpose technologies (GPTs) – foundational innovations that reshape society and the economy across multiple sectors, much like the printing press, electricity, computers or the internet.[22] More broadly, ML techniques are defined by their versatility and capacity to generate innovation across domains, thanks to their core ability to identify patterns and make predictions from data. As such, ML represents a paradigm shift from earlier single-purpose AI systems. At the same time, these models can be easily adapted for domain-specific applications without requiring major redesigns.

This dual nature – general-purpose but also highly adaptable – sets ML apart from other historical GPTs and adds complexity to policy debates. Policymakers must grapple with this duality when designing frameworks that support both broad societal benefits and context-specific applications of ML and generative AI technologies.

[22] Sabrina Küspert, Nicolas Moës, Connor Dunlop. “The Value Chain of General-Purpose AI.” Ada Lovelace Institute. 10 February 2023. https://www.adalovelaceinstitute.org/blog/value-chain-general-purpose-ai/

Figure 2 | Neural network

An example of a simple neural network capable of recognizing letters of the alphabet.

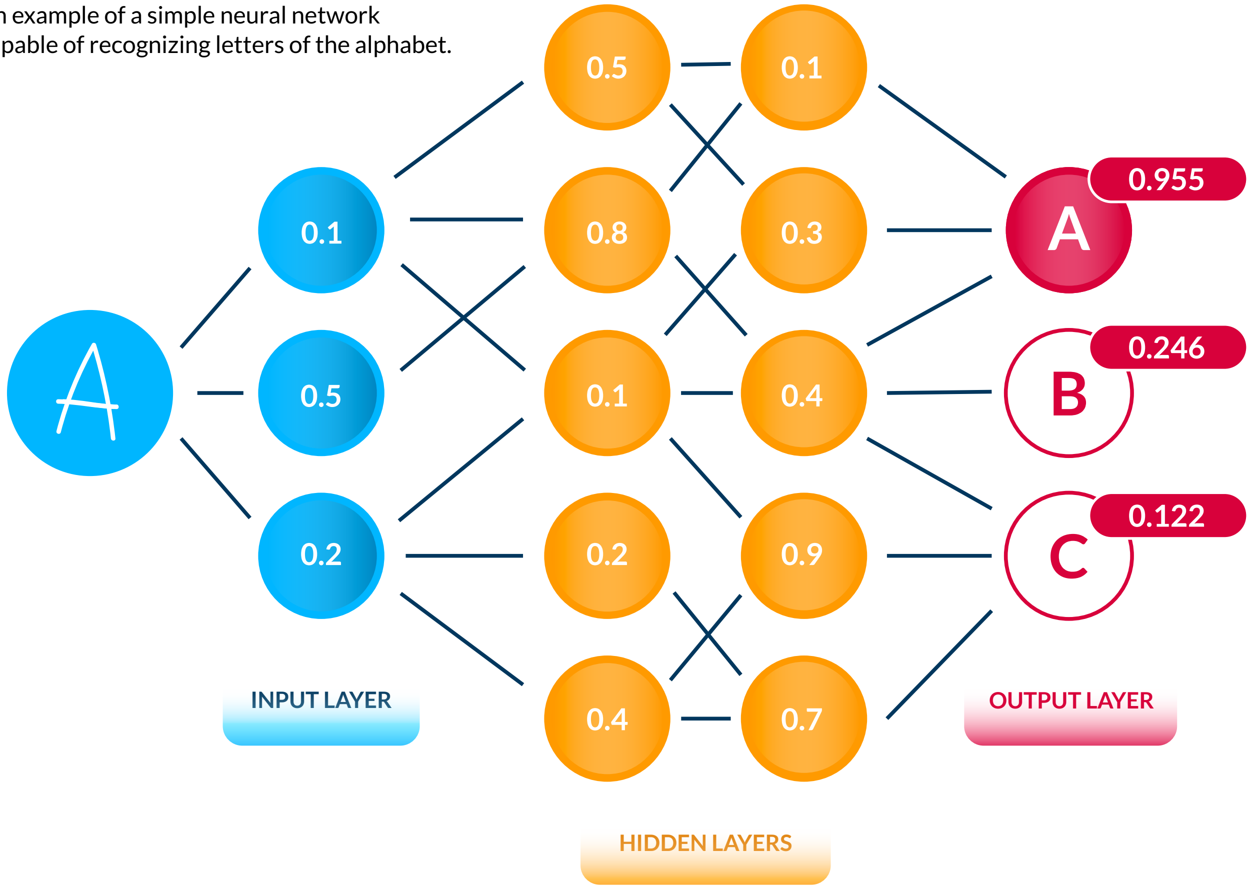


Illustration by: Jakub Koźniewski

The deep learning paradigm

Various ML paradigms have been developed to train models to learn from data and perform specific tasks. This ability to train models based on data, and the applications that can be built on its basis, are the key technological capacities that public AI policies aim to secure for the public interest. In this section, we focus on deep learning approaches, which involve the use of neural networks. Deep learning has become a central paradigm for generative AI development. It gained significant momentum following a breakthrough research paper published in 2006,[23] and the pioneering application of that research in the development of AlexNet in 2012.[24] As the paradigm evolved over the last decade, three types of resources emerged as particularly critical to its success: compute, data and model architectures (including model size).

The underlying design of neural networks is based on computational model architectures inspired by biological neurons in the brain. These networks consist of interconnected artificial neurons organized into layers that process and transform data through a series of mathematical operations. Each network has an input layer that receives raw data, hidden layers that process this information and an output layer that produces results. The “deep” in deep learning refers to the presence of multiple hidden layers between the input and output layers, which enables the model to learn increasingly complex and abstract representations of data.

Figure 2 shows an example of a simple neural network capable of recognizing letters of the alphabet. The network learns by adjusting the weights – the strengths of the connections between neurons – through a process called backpropagation. In this process, the network makes predictions, compares them to correct answers, calculates the error and updates the weights to reduce this error.[25] Importantly, the total number of weights – often expressed as 2B, 7B, or 40B (to indicate billions of parameters) – determines the size of the model and plays a crucial role in determining its capabilities and effectiveness. This architecture has proven powerful for complex data types like images, text and audio, enabling major breakthroughs in fields from computer vision to natural language processing.

Neural networks were first developed in the 1950s and, while promising, were long constrained by two major limitations: insufficient computing power and limited access to data. Over the next decades, development of machine learning technologies – including recent advances in generative AI – depended on securing these key resources. The field experienced several cycles of enthusiasm and disappointment until three key developments converged in the 2010s:

- **Model architectures:** The development of novel architectures, such as convolutional neural networks, initially used for computer vision and later adopted for machine translation. Ultimately, the transformer architecture enabled the emergence of generative AI applications.

- **Compute:** Advances in compute infrastructure, including the use of Graphics Processing Units (GPUs) for machine learning and the development of the CUDA programming platform, allowed for the massive parallel computations required for AI training.

- **Data:** The creation of large labeled datasets enabled a data-centric approach to AI and reinforced the dominance of supervised learning. Many of these datasets were built from publicly available or openly shared web content, highlighting how open data can facilitate the collection and curation of training examples.

[23] Geoffrey E. Hinton, et al. “A Fast Learning Algorithm for Deep Belief Nets.” Neural Computation, vol. 18, no. 7, July 2006, pp. 1527–54. DOI.org (Crossref), https://doi.org/10.1162/neco.2006.18.7.1527

[24] Alex Krizhevsky, et al. “ImageNet Classification with Deep Convolutional Neural Networks.” Advances in Neural Information Processing Systems, vol. 25, Curran Associates, Inc., 2012. Neural Information Processing Systems, https://papers.nips.cc/paper_files/paper/2012/hash/c399862d3b9d6b76c8436e924a68c45b-Abstract.html

[25] David E. Rumelhart, et al. “Learning Representations by Back-Propagating Errors.” Nature, vol. 323, no. 6088, Oct. 1986, pp. 533–36. https://doi.org/10.1038/323533a0

A pivotal moment came in 2012 with AlexNet, a deep neural network that significantly outperformed existing methods in image recognition. Its success demonstrated that deep learning models could achieve breakthrough results when trained on large datasets using sufficient computational power. This marked the beginning of the modern deep learning era.

AlexNet would not have been possible without ImageNet, a dataset containing millions of labeled images across thousands of categories. ImageNet provided the scale necessary to train deep neural networks in ways that had not previously been feasible. Its success relied on publicly available web content that could be scraped, as well as on Amazon's Mechanical Turk platform, where tens of thousands of low-paid workers labeled the images. According to Fei Fei Li, who led the work, the ImageNet project was the first initiative to demonstrate the critical role of large-scale datasets for AI development.[26]

Equally important was the leap in computational capabilities. A key breakthrough came with GPUs, which were originally designed for computer games but proved transformative for AI due to their superior ability to handle multiple calculations simultaneously. Nvidia‘s introduction of the CUDA computing platform in 2006 was especially important, as it allowed GPUs to be used for general-purpose machine learning tasks. This convergence of massive datasets and GPU-powered parallel processing capabilities proved decisive in accelerating AI development.

Already with ImageNet, data governance issues that now plague AI development were beginning to surface. The labor of some 49,000 underpaid Mechanical Turk workers remained largely invisible. The work involved in labeling images scraped from the web, and the rights of the individuals depicted in those images, were also poorly acknowledged.[27] Later, critical analyses of ImageNet revealed that the labeling process introduced multiple forms of bias, which went unaddressed even as the dataset became foundational to AI research.[28] Similar problems applied to other computer vision datasets built at that time, and these challenges have continued with the creation of training datasets for generative AI models.[29]

"Attention is all you need": Transformers and the rise of generative AI

Among recent technological innovations, the development of the transformer architecture stands out as arguably the most significant breakthrough driving the rapid progress of generative AI over the past decade. Introduced by machine translation researchers at Google in 2017, this architecture marked a major leap forward in machine learning for language processing.[30]

Unlike earlier models that processed text sequentially, transformers can process entire sequences simultaneously through a mechanism called self-attention. This mechanism allows the model to weigh the relevance of each word in relation to all others in a sentence, greatly enhancing its capacity to understand context and meaning.

The complexity of the self-attention mechanisms means that transformers require compute and data resources at a scale not seen in any prior algorithm. This characteristic has key consequences for how AI development is resourced – an issue that public AI policies must take seriously.

[26] Timothy B. Lee. "Why the Deep Learning Boom Caught Almost Everyone by Surprise." Understanding AI. 5 November 2024. https://www.understandingai.org/p/why-the-deep-learning-boom-caught

[27] Vi Hart. "Changing My Mind about AI, Universal Basic Income, and the Value of Data." The Art of Research. 30 May 2019. https://theartofresearch.org/ai-ubi-and-data/

[28] Kate Crawford and Trevor Paglen. "Excavating AI. The Politics of Images in Machine Learning Training Sets." 19 September 2019. https://excavating.ai/

[29] Alek Tarkowski and Zuzanna Warso. "AI Commons. Filling the governance vacuum on the use of information commons for AI training." 12 January 2023. https://openfuture.eu/publication/ai-commons/

[30] Ashish Vaswani, et al. Attention Is All You Need. arXiv:1706.03762, arXiv, 2 Aug. 2023. arXiv.org, https://doi.org/10.48550/arXiv.1706.03762

INFOBOX | The attention mechanism in transformer architectures

The attention mechanism starts by converting each token – a word or subword unit – into a high-dimensional vector. It then computes three matrices, known as queries, keys and values from these vector tokens. These matrices help the model determine how information flows between tokens, allowing it to capture both short-range and long-range dependencies in the text. This architecture enables models to learn increasingly sophisticated language patterns.

Previous approaches to language processing relied primarily on techniques like recurrent neural networks (RNNs), which processed text sequentially (i.e., one word at a time). Consider the sentence:

“Watching the water flow gently, she sat down by the bank.”
RNN-based models would process the sentence word by word. By the time it reaches the word “bank,” the model might struggle to recall earlier words like “water” and “flow” because RNNs tend to forget information as the sequence grows longer. While more sophisticated architectures like long short-term memory networks (LSTMs) improved on basic RNNs, they still faced fundamental limitations in processing long sequences.

The key innovation of transformers was the introduction of the attention mechanism, which enables the model to process all words in a sequence simultaneously and compute their relationships. For instance, in the sentence:

“The water continued to flow steadily, gradually eroding the bank.”
The attention mechanism allows the model to recognize that “water” and “flow” are highly relevant for interpreting the meaning of “bank,” helping it infer that the term refers to a riverbank rather than a financial institution – even though the relevant words appear several tokens earlier. When processing each word, the model calculates attention scores to determine how much weight to assign to every other word in the sequence.

It is important to note that transformer architectures come in different variants that use the attention mechanism in task-specific ways. Encoder-only models process entire inputs at once to build rich contextual representations (as in the example above), making them well suited for tasks such as text classification and comprehension. Decoder-only models generate outputs one token at a time, attending to previously generated tokens, and are used in text generation and completion systems. Encoder-decoder models combine both approaches by first encoding the full input before the decoder produces an output, enabling tasks like translation, summarization and question answering.

Since 2017, the transformer architecture has become the foundation for a wide range of language tasks. Its most prominent application has been in LLMs, notably the Generative Pretrained Transformer (GPT) family developed by OpenAI, beginning with GPT-2 in 2019. Since then, generative AI development has focused on extending and improving model capabilities. In this development paradigm, progress is achieved by scaling three key resources: data, compute and the size of the resulting model. The characteristics of transformer architectures – and their high demands for compute and data – are crucial considerations for public AI policies.

Recent architectural innovations have extended transformer capabilities to handle multiple modalities simultaneously. These multimodal models are able to process combinations of text, images, audi and video through specialized adaptations of the original architecture. At the same time, researchers are exploring novel approaches like domain-specific fine-tuning, which improves model performance for specific applications without relying solely on increasing model size. Advances in quantization and pruning re also making it possible to deploy larger models more efficiently on resource-constrained hardware.

The generative AI development process

The development of generative AI models is a multistage process that combines different training approaches to create capable AI systems. This process involves multiple components, including the model architecture, code used to train or evaluate a model, code used to preprocess training datasets, and datasets used for model training, evaluation or alignment. Throughout the process, computing resources are used to train the generative AI model on large volumes of data.

In this section, we provide an overview of the generative AI development process, from the pretraining of base foundation models to their post-training fine-tuning and eventual deployment in user-facing applications. The technical characteristics of this process determine the resources needed to create generative AI. Securing access to these resources is a key objective of public AI policies. The aim here is to provide a conceptual framework that illustrates the core stages and decision points in developing generative AI systems. Understanding how the components interconnect – and where dependencies arise – is essential for designing realistic and actionable pathways toward public AI.

The diagram below offers a simplified illustration of the generative AI development process, divided into three phases: pretraining, post-training and inference.

Pretraining phase

Pretraining is the first step in the model development pipeline, in which models are trained through self-supervised learning to identify general patterns from large datasets.

Before any training begins, the data must be prepared. This process starts with data collection, followed by the cleaning and validation of raw data. At this stage, legal issues – such as licensing – are addressed. The cleaned dataset is then filtered to remove low-quality or inappropriate content. The data is further processed; for example, text data used to train LLMs is tokenized to ensure it can be efficiently processed by the model. If sensitive data is involved, privacy and security measures are also implemented. In parallel, the model architecture, software and other tooling are designed and prepared.

This phase requires substantial computational resources and, since 2020, has been shaped by scaling laws – a trend in which increasing model size, dataset size and compute power leads to improved performance (these laws are described in more detail on page 27). Each new generation of frontier models requires greater quantities of these key resources. However, computational demands vary significantly depending on model scale, and recent advances in training efficiency have begun to reduce these requirements. A new small model paradigm is emerging, shifting focus from ever-larger models to more efficient and resource-conscious approaches.[31]

In each case, pretraining is the most resource-intensive phase of model development. Its output is a pretrained base model – not yet suitable for deployment, but capable of serving as a general-purpose foundation for further training.

[31] Rina Diane Caballar. “What Are Small Language Models (SLM)?.” IBM. 31 October 2024. https://www.ibm.com/think/topics/small-language-models Accessed 3 April 2025.

Figure 3 | Development of a generative AI model

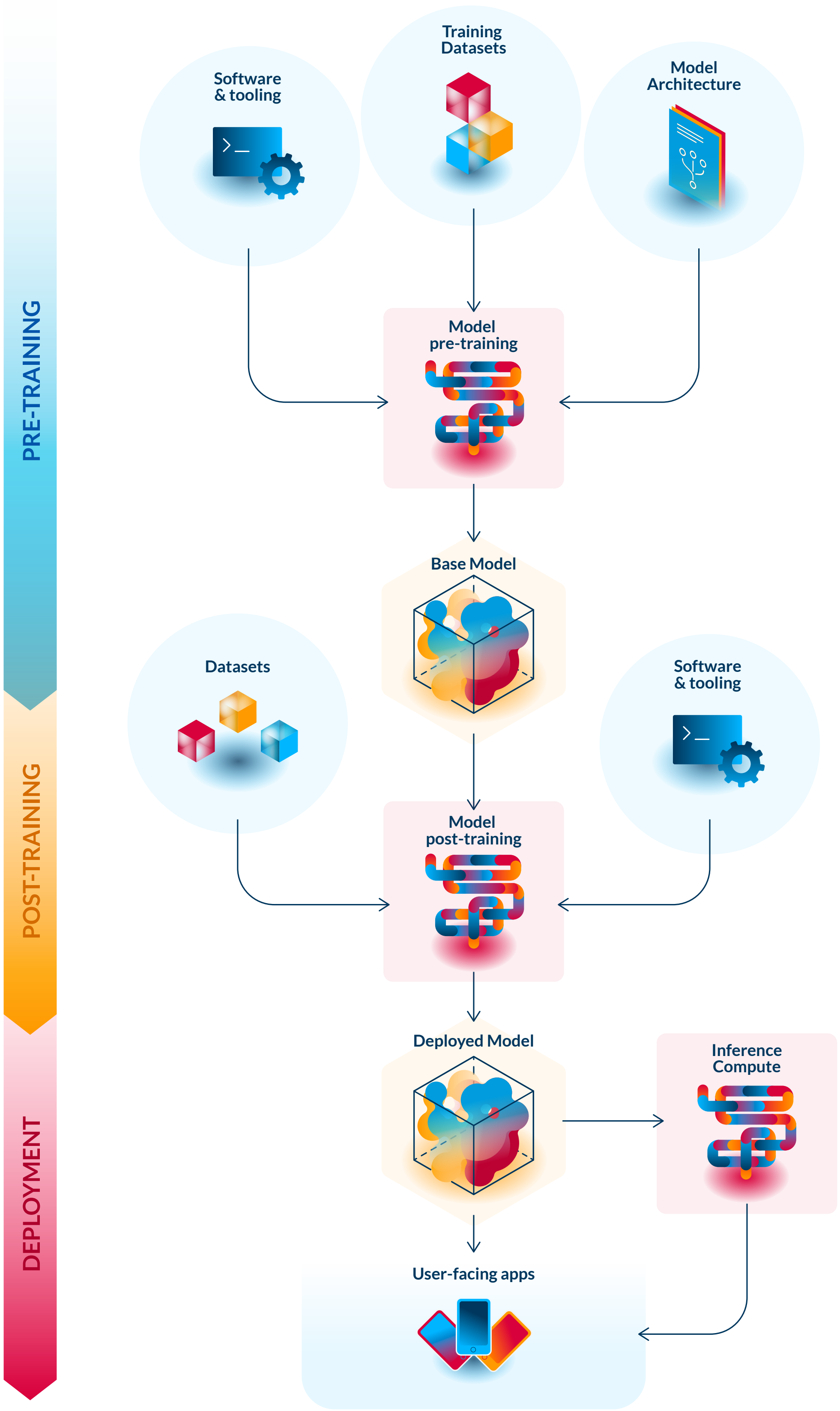


Illustration by: Jakub Koźniewski

Post-training phase

During this phase, the base model is refined through additional training to improve its capabilities for specific domains and align its behavior with specific objectives. At this stage, various datasets are used to further train the model, including validation, instruction and benchmark datasets.

Development methods in this phase have evolved rapidly, with two key approaches becoming prominent in 2024. The first is supervised fine-tuning, in which models are trained on domain-specific or task-specific datasets. For example, a model might be finetuned on medical literature to enhance its healthcare-related capabilities. The second approach relies on reinforcement learning to enhance model reasoning and decision-making. Two primary methods are reinforcement learning from human feedback (RLHF), where human preferences guide the learning process, and reinforcement learning from AI feedback (RLAIF), which uses other AI systems to provide training signals.[32]

This phase also includes evaluation and deployment optimization, ensuring that the models meet requirements for accuracy, reliability, computational efficiency and safety before deployment in real-world applications. This includes testing on standardized benchmarks, which measure model capabilities across diverse domains including reasoning, knowledge and safety.[33] Models are also evaluated through specialized testing methodologies like adversarial testing and hallucination detection.

The result of this phase is a fully trained model, ready for deployment.

Deployment

In the so-called inference stage, the trained model is deployed for use in user-facing applications, requiring additional computational resources known as inference compute. Alternatively, a model can be hosted on a cloud platform and made available via an API, enabling third parties to build their own applications on top of it. Inference compute is essential to run the model and, in the case of large frontier models, can involve significant costs and environmental impact. This is also the reason why small, more sustainable models are being developed.

The overview presented here simplifies what is often a far more complex development process. In practice, model development is rarely a one-time effort. AI labs typically seek to aim to improve their models over time, revisiting earlier stages of the process. As a result, development is often iterative, with feedback from later stages informing adjustments to earlier components. Developers frequently cycle between phases as they refine their systems. Finally, compound AI systems combine multiple models to build holistic workflows and applications.

This overview also does not fully reflect the diversity of open model development ecosystems, such as those found on platforms like Hugging Face. Such development entails the combined use of various models. In such cases, work often begins with an openly available base model from which derivative models are produced, such as through fine-tuning. Furthermore, smaller models can be created from large models through techniques like distillation or quantization (see Section 3).

AI scaling laws: The contested future of AI

In the previous section, we provided an overview of how generative AI is developed, and demonstrated how compute, data, software and architectural components are used in this process. In this section we explain AI scaling laws – a defining feature of transformer-based model development. The discovery of

[32] Nathan Lambert. “Beyond Human Data: RLAIF Needs a Rebrand.” Interconnects. 26 April 2023. https://www.interconnects.ai/p/beyond-human-data-rlaif

[33] Reginald Martyr. “LLM Benchmarks Explained: Significance, Metrics & Challenges.” Orq.ai. 26 February 2025. https://orq.ai/blog/llm-benchmarks

these empirical relationships has made AI development increasingly dependent on securing ever-larger amounts of compute power and training data. Ensuring access to these resources must be a core focus of any public AI policy aiming to support state-of-the-art model development.

What are AI scaling laws?

A 2020 research paper by OpenAI researchers observed that there are scaling laws[34] inherent in transformer-based model development. The paper identified a power law relationship between three scalable resources for AI training – model parameter count, training data size and computational power – and model performance. The approach has been described as the "bigger is better" paradigm in AI.[35]

Scaling laws are not natural laws. They are based on empirical observations: performance on benchmark tasks improves when all three inputs – model size, dataset size and compute – are increased together during the pre-training phase. In other words, transformer models – known for their resource-intensive nature due to their ability to process vast amounts of data in parallel – are central to this paradigm. These improvements only emerge when all three components are scaled up together during the pretraining phase. Researchers found that increasing any single factor in isolation leads to diminishing returns.[36]

The so-called Chinchilla scaling law introduced in 2022 by DeepMind, refined this model by proposing a more optimal ratio between model size, training data and compute.[37] The research demonstrated that earlier models, in particular GPT-3, had been undertrained: the models were too large relative to the amount of data and compute used. A new approach, suggested in the paper, allowed smaller models to be as effective as larger ones, if higher quality data was used for training over extended periods of time.

The evolution of AI scaling laws

While scaling laws have driven significant gains in AI performance, recent research suggests that the benefits of continued scaling may be slowing.[38] The key constraint is the limited availability of high-quality training data. Ilya Sutskever, one of the co-founders of OpenAI and a key figure in transformer-based model development, argues that we have reached "peak data,"[39] as all frontier AI labs rely on scraping web data, and scaling laws are beginning to plateau. The stakes in this debate are high. Proponents of artificial general intelligence (AGI), a form of artificial super-intelligence, often place their bets on the continued validity of scaling laws. This belief in the power of AI scaling also underpins recent massive investments into compute, such as the $500 billion investment announced by OpenAI, Oracle and Softbank in January 2025.[40] Critics, meanwhile, view the current wave of AI development as yet another hype cycle destined to collapse because of the consequences of scaling laws.[41]

Opinions about the future of AI scaling laws vary. Some experts argue that scaling laws plateau naturally over time and that the current slowdown is a natural progression of pretraining scaling laws and

[34] Jared, Kaplan, et al. Scaling Laws for Neural Language Models. arXiv:2001.08361, arXiv, 23 Jan. 2020. arXiv.org, https://doi.org/10.48550/arXiv.2001.08361

[35] Gaël Varoquaux, et al. Hype, Sustainability, and the Price of the Bigger-Is-Better Paradigm in AI. arXiv:2409.14160, arXiv, 1 Mar. 2025. arXiv.org, https://doi.org/10.48550/arXiv.2409.14160

[36] Cameron R. Wolfe. "Scaling Laws for LLMs: From GPT-3 to O3." Deep (Learning) Focus, 6 Jan. 2025, https://cameronrwolfe.substack.com/p/llm-scaling-laws

[37] Jordan, Hoffmann, et al. Training Compute-Optimal Large Language Models. arXiv:2203.15556, arXiv, 29 Mar. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2203.15556

[38] Gaël Varoquaux, et al. ibid.

[39] Jeffrey Dastin. "AI with Reasoning Power Will Be Less Predictable, Ilya Sutskever Says." Reuters. 14 December 2024. https://www.reuters.com/technology/artificial-intelligence/ai-with-reasoning-power-will-be-less-predictable-ilya-sutskever-says-2024-12-14/

[40] Reuters. "SoftBank to Invest $500 Mln in OpenAI, The Information Reports." 30 September 2024. https://www.reuters.com/technology/softbank-invest-500-mln-openai-information-reports-2024-09-30/

[41] Gary Marcus. "The Most Underreported and Important Story in AI Right Now Is That Pure Scaling Has Failed to Produce AGI." Fortune. 19 February 2025. https://fortune.com/2025/02/19/generative-ai-scaling-agi-deep-learning/

was always to be expected. Others predict that AI scaling laws will not diminish but rather evolve, as the scaling effects depend on multiple factors.[42]

An evolution in scaling is already underway. The singular focus on model pretraining –dominant between 2020 and 2023 – is giving way to a new paradigm that focuses on advantages gained in later stages of development, or even in the deployment phase.[43] Even if the original pretraining scaling laws are beginning to level off, these other phases are increasingly seen as the next frontier. Some researchers argue that the future of AI capabilities might depend more on finding the right balance between these three scaling dimensions rather than pushing any single dimension to its limits.[44]

Starting in 2024, leading AI companies have shifted their focus to scaling during the post-training phase. In this phase, optimization involves refining models after their initial training through reinforcement learning techniques and other fine-tuning methods.[45] These methods help align models with human preferences and specific tasks, and enable the development of models capable of more complex reasoning. Research suggests that the relationship between resources invested in post-training and resulting performance improvements follows its own distinct scaling patterns. Some researchers argue that there may be more headroom for improvement in this phase than in pretraining, particularly as reinforcement learning methods continue to advance.[46]

A related trend is inference compute scaling, where greater computational resources are allocated during model use to enhance performance. This has emerged as a new frontier, particularly in the development of reasoning models that generate multiple candidate outputs and internally select the best one.[47] Early results indicate this approach can significantly boost performance without increasing model size or requiring additional training data.[48] The approach gained attention with the release of OpenAI's reasoning models, such as O1 and O3, which demonstrated strong benchmark performance using this method.

Even if the original scaling laws – now referred to as pretraining scaling – begin to plateau, this does not necessarily imply a reduction in computational demand. Inference scaling continues to gain traction and could sustain high resource requirements. Consequently, the development and deployment of transformer-based models may remain costly, even as architectures and training methods evolve.[49]

As described in chapter 2, AI scaling laws are a key driver of concentration in the AI ecosystem. An AI development approach that is based on transformer architecture, and that adheres to AI scaling laws has several repercussions. First, it creates the kind of concentrations of market power outlined in the previous section. Second, it creates new and distinct forms of digital divide, which are related to uneven access to computing resources. And third, it dramatically increases the environmental footprint of AI systems.

Scaling and AI's environmental footprint

The training and deployment of large AI models comes with a major environmental footprint that extends beyond just the energy consumption of data

[42] Cameron R. Wolfe. ibid.

[43] Dario Amodei. "On DeepSeek and Export Controls." Dario Amodei (blog). January 2025. https://www.darioamodei.com/post/on-deepseek-and-export-controls Accessed 3 April 2025.

[44] Jordan Hoffmann, et al. ibid.

[45] Cameron R. Wolfe. ibid.

[46] Cameron R. Wolfe. "Basics of Reinforcement Learning for LLMs. Understanding the problem formulation and basic algorithms for RL." Deep (Learning) Focus, 25 September 2023. https://cameronrwolfe.substack.com/p/basics-of-reinforcement-learning/

[47] Maxwell Zeff. "Current AI Scaling Laws Are Showing Diminishing Returns, Forcing AI Labs to Change Course." TechCrunch, 20 November 2024. https://techcrunch.com/2024/11/20/ai-scaling-laws-are-showing-diminishing-returns-forcing-ai-labs-to-change-course/

[48] Cameron R. Wolfe. ibid.

[49] "Scaling Laws – O1 Pro Architecture, Reasoning Training Infrastructure, Orion and Claude 3.5 Opus 'Failures.'" SemiAnalysis, 11 December 2024, https://semianalysis.com/2024/12/11/scaling-laws-o1-pro-architecture-reasoning-training-infrastructure-orion-and-claude-3-5-opus-failures/

centers. For example, training a single large language model can emit up to 550 metric tons of CO2.[50] Moreover, the energy required for deployment – known as inference – accounts for a substantial share of ongoing AI-related energy use, ranging from one-third at Meta[51] to as much as 60% at Google.[52] In addition, the cooling systems required to prevent data centers from overheating often rely on large volumes of water, placing further strain on local water resources.[53] As AI systems become more widespread, these energy demands continue to grow. It is telling that the dominant AI companies are evolving into energy companies.[54] This trajectory is fundamentally unsustainable, as computational demands grow faster than improvements in model performance.[55]

The environmental costs of AI extend beyond energy consumption. Beyond data centers, the entire AI supply chain raises serious concerns about the environmental costs of contemporary approaches to AI development and deployment[56] – from the extraction of raw materials for GPUs to the mounting problem of electronic waste from discarded hardware. This creates what researchers call a Jevons paradox:[57] as individual models become more efficient, overall environmental impact increases because improved efficiency leads to broader deployment and more frequent use. This dynamic raises questions about whether the current trajectory of AI development, with its emphasis on scale, is environmentally sustainable in the long term.

These environmental concerns are an important factor that should be integrated into public AI policymaking. Any deployment of AI technologies must address the environmental impact inherent in today's generative AI systems. Commercial AI labs – operating within the paradigm of AI scaling laws and leveraging full-stack approaches – tend to embrace a "bigger is better" model of data center development, often at the expense of environmental sustainability.[58] AI's environmental impact, alongside financial limitations, is a major reason why public AI policies should not simply replicate commercial strategies focused on ever-larger models.

What is the future of AI scaling laws?

In late 2024, news about DeepSeek's V3 and R1 models led many media observers to question the scaling paradigm's future. Initial reports suggested that DeepSeek had managed to reduce development costs from from hundreds of billions to just several billion dollars. However, it has since become clear that these models do not represent a fundamental shift in AI development, contrary to what some experts initially suggested.

DeepSeek ultimately did not demonstrate that a new generation of state-of-the-art models can be built with dramatically reduced compute requirements. The widely cited $5.6 million training figure referred

[50] Gaël Varoquaux, et al. ibid.

[51] Carole-Jean Wu, et al. "Sustainable AI: Environmental Implications, Challenges and Opportunities." Proceedings of Machine Learning and Systems, vol. 4, Apr. 2022, pp. 795–813. proceedings.mlsys.org, https://proceedings.mlsys.org/paper_files/paper/2022/hash/462211f67c7d858f663355eff93b745e-Abstract.html

[52] David Patterson, et al. "The Carbon Footprint of Machine Learning Training Will Plateau, Then Shrink." Computer, vol. 55, no. 7, July 2022, pp. 18–28. IEEE Xplore, https://doi.org/10.1109/MC.2022.3148714

[53] Cindy Gordon. "AI Is Accelerating the Loss of Our Scarcest Natural Resource: Water." Forbes. 25 February 2024. https://www.forbes.com/sites/cindygordon/2024/02/25/ai-is-accelerating-the-loss-of-our-scarcest-natural-resource-water/

[54] Alex Lawson. "Google to Buy Nuclear Power for AI Datacentres in 'World First' Deal." The Guardian, 15 October 2024. https://www.theguardian.com/technology/2024/oct/15/google-buy-nuclear-power-ai-datacentres-kairos-power

[55] Gaël Varoquaux, et al. ibid.

[56] Ana Valdivia. "The Supply Chain Capitalism of AI : A Call to (Re)Think Algorithmic Harms and Resistance through Environmental Lens." Information, Communication & Society, Oct. 2024, pp. 1–17. DOI.org (Crossref), https://doi.org/10.1080/1369118X.2024.2420021

[57] Alexandra Sasha Luccioni, Emma Strubell, and Kate Crawford. 'From Efficiency Gains to Rebound Effects: The Problem of Jevons' Paradox in AI's Polarized Environmental Debate'. arXiv, 27 January 2025. https://doi.org/10.48550/arXiv.2501.16548

[58] Dara Kerr. "AI Brings Soaring Emissions for Google and Microsoft, a Major Contributor to Climate Change." NPR, 12 July 2024. NPR, https://www.npr.org/2024/07/12/g-s1-9545/ai-brings-soaring-emissions-for-google-and-microsoft-a-major-contributor-to-climate-change

only to the final training run. The total cost of DeepSeek's AI infrastructure is estimated at $1.6 billion. The company operates around 50,000 GPUs – comparable to major Western AI labs and consistent with the demands of the AI scaling laws.[59]

As explained in the technical report, DeepSeek researchers achieved performance on par with existing state-of-the-art models through two major algorithmic improvements.[60] The first was an advancement of the mixture of experts technique, which divides the model into specialized submodels. Instead of activating the full model during training and inference, only relevant submodels are engaged, reducing computational requirements.

The second breakthrough prompted by U.S. export controls, which restrict DeepSeek to using lower-performance Nvidia chips with limited memory bandwidth – a major constraint, since model training requires moving massive volumes of data between memory and processing units. In response, DeepSeek developed methods to reduce memory bandwidth demands during both pretraining and inference.[61]

Moreover, the DeepSeek researchers demonstrated that the advanced capabilities of state-of-the-art models like DeepSeek-R1 can be distilled into smaller models. These models outperform state-of-the-art models, both open and proprietary. The company has also pledged to release open-weight versions of its models – freely available for use, though without access to the original training data or certain proprietary components.

The release of DeepSeek's models offers several important lessons for public AI strategy. First, under the current scaling paradigm, access to computing power remains a critical requirement for developing state-of-the-art AI models. Even with the innovations introduced by the DeepSeek team, building state-of-the-art models remains a resource-intensive endeavor. Even if costs – and thus environmental impact – are reduced at each stage of training and deployment, the Jevons paradox still applies: efficiency gains may drive wider adoption, ultimately increasing total energy consumption.[62]

However, computing power is only one part of the equation. The DeepSeek example shows that significant gains can also be achieved through advances in machine learning techniques. Many of these depend on the availability of a state-of-the-art model, which can be used to create derivative, smaller models – enabling capability transfer without needing to replicate the full computational cost of the original model.

Therefore, concentration of compute does not automatically result in a lasting concentration of model capability. Distillation techniques used by DeepSeek show that smaller, more energy-efficient models can be created on the basis of larger models.

The key takeaway from the release of DeepSeek's models is that ultimately it confirms the economics of frontier AI development under the current scaling paradigm. Any public AI strategy must still secure access to significant computing – comparable to those of major commercial AI labs – in order to develop state-of-the-art models. This either requires significant investment in public compute infrastructure, which would only yield results over the longer term, or accepting a degree of reliance on commercial compute providers.

[59] Anton Shilov. "DeepSeek Might Not Be as Disruptive as Claimed, Firm Reportedly Has 50,000 Nvidia GPUs and Spent $1.6 Billion on Buildouts." Tom's Hardware, 2 February 2025, https://www.tomshardware.com/tech-industry/artificial-intelligence/deepseek-might-not-be-as-disruptive-as-claimed-firm-reportedly-has-50-000-nvidia-gpus-and-spent-usd1-6-billion-on-buildouts

[60] DeepSeek-AI, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948, arXiv, 22 Jan. 2025. arXiv.org, https://doi.org/10.48550/arXiv.2501.12948

[61] Ben Thompson. "DeepSeek FAQ." Stratechery. 27 January 2025. https://stratechery.com/2025/deepseek-faq/

[62] Jennifer Collins. "What Does DeepSeek Mean for AI's Environmental Impact?" DW.com. 30 January 2025. https://www.dw.com/en/what-does-chinas-deepseek-mean-for-ais-energy-and-water-use/a-71459557

At the same time, the DeepSeek example illustrates that machine learning techniques are far from reaching their full potential. Algorithmic improvements and post-training methods such as distillation can make training more efficient and enable the development of smaller, more accessible models. A public AI strategy could therefore prioritize a research and innovation agenda aimed at reducing dependence on transformer architectures and their associated scaling laws. Small model development can also complement the creation of frontier models – particularly when models are openly released – by supporting an ecosystem in which a range of models can be adapted and used by diverse actors.

3 | The generative AI stack

Overview of the AI stack

AI systems can be understood to comprise a layered technological stack in which each layer interacts with and supports the others. Each layer typically serves a distinct purpose, as different components – often developed by different actors – contribute to the creation and operation of the overall system. This modular perspective, similar to how the internet is often described as a technological stack, helps clarify the roles, dependencies and forms of collaboration involved in AI development and deployment.[63]

The basic AI stack consists of the following layers, arranged from the bottom up:

- **Compute:** This foundational layer refers to the physical and software infrastructure that enables AI development and deployment. At its core are specialized processors or chips – primarily GPUs – designed to handle the massive parallel computations required for training and running AI models. To make these chips usable at scale, two elements are essential: software frameworks that optimize GPU performance and the integration of GPUs into data centers, where they are stacked and networked into powerful, scalable compute systems, often delivered via cloud platforms.

- **Data:** This layer involves storage, processing and transfer of datasets used in both the pretraining and post-training phases of AI development.

Figure 4 | Layers of the AI stack

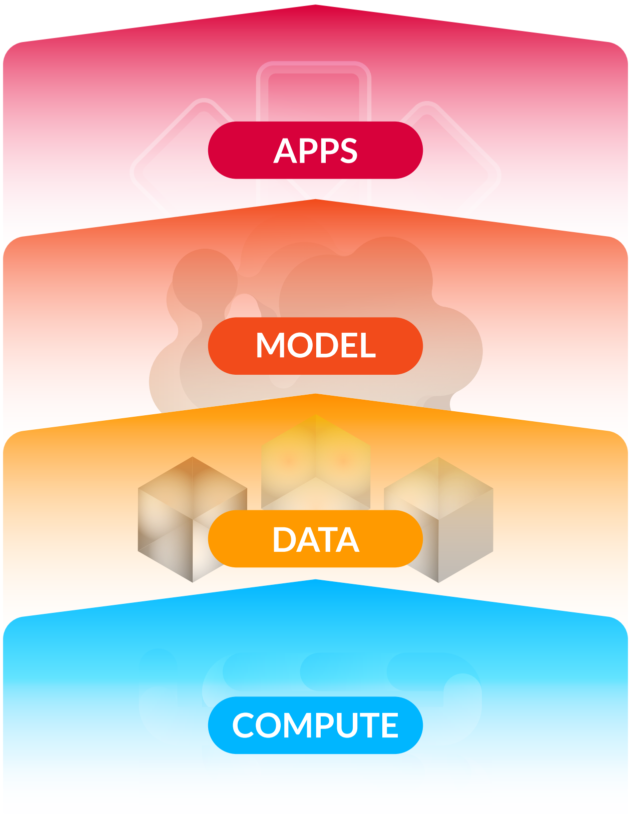


Illustration by: Jakub Koźniewski | BertelsmannStiftung

- **Models:** This layer refers to the AI models themselves. Each model consists of an architecture and a set of parameters – its weights and biases – refined through training. These models are typically deployed as cloud-based services.

[63] Cole Stryker, "What Is an AI Stack?" IBM. 29 November 2024. https://www.ibm.com/think/topics/ai-stack

- **Applications:** In this layer, AI models are embedded in user-facing systems and applications. Running these applications requires additional computing power, referred to as inference or test-time compute.

In analyzing public AI development pathways, we focus on the compute, data and model layers. We do not address the applications layer, as it lies downstream from the core layers under discussion. The existence of public AI applications depends on the availability and capabilities of AI systems built through the orchestration of resources at these foundational levels. Therefore, when considering elements of public AI policy, application development pathways will be listed as a key additional measure.

Beyond the fundamental layers – from hardware to applications – additional layers are sometimes identified to emphasize other critical functions. For example, Mozilla introduces a safeguards layer in its AI stack to highlight the importance of tools and mechanisms that ensure the safety of AI systems.[64] Similarly, a talent layer is often added to highlight the indispensable role of human talent and know-how in driving AI development.[65]

Software used in AI development can also be seen as a cross-cutting layer that spans the entire AI stack. At each level, both proprietary and open source solutions are commonly used by developers. At the compute level, software is essential for managing and orchestrating hardware resources during the development and deployment of AI models. It creates abstractions that allow developers to harness computing power without dealing directly with the complexity of low-level hardware. For example, Nvidia's CUDA provides direct access to GPUs and optimizes their use for AI workloads, while PyTorch is a deep learning framework that abstracts hardware complexity and offers high-level APIs for efficient model development.

In the data layer, software manages data ingestion, processing and organization. For models, it provides development environments as well as tools for training, evaluation and fine-tuning. Widely used open source libraries include scikit-learn and PyTorch for training machine learning models, GPT-NeoX for training large language models, vLLM for model inference or serving, transformers for fine-tuning and LM Harness and lighteval for evaluation. At the application layer, software provides APIs, monitoring systems and user interfaces that make AI models accessible and usable by downstream developers and users.

Because software development cuts across all layers of the stack, it is difficult to define a standalone public AI development pathway that focuses solely on software. Instead, support for software development should be considered a key complementary measure within each of the pathways.

The section below outlines the advantages of the stack model for designing public AI policy and highlights how it helps clarify concentrations of power in the AI ecosystem. This is followed by a closer look at the three core layers – compute, data and models. The characteristics of these layers shape dependencies on commercially provided resources that public AI initiatives must navigate. They also present key considerations for any effort to develop independent public AI systems.

Advantages of the AI stack concept

The stack model can be a useful framework for governing complex technologies. In this context, governance is understood as the exertion of control over the various layers of the stack and the orchestration of actors at each layer to achieve specific outcomes through the use of the overall technology. Typical forms of such control include regulation and volun-

[64] Adrien Basdevant, et al. "Towards a Framework for Openness in Foundation Models. Proceedings from the Columbia Convening on Openness in Artificial Intelligence." Mozilla. 21 May 2024. https://foundation.mozilla.org/en/research/library/towards-a-framework-for-openness-in-foundation-models/

[65] Ganesh Sitaraman and Alex Pascal. "The National Security Case for Public AI." Vanderbilt Policy Accelerator for Political Economy and Regulation. 24 September 2024. https://cdn.vanderbilt.edu/vu-URL/wp-content/uploads/sites/412/2024/09/27201409/VPA-Paper-National-Security-Case-for-AI.pdf

tary norms,[66] and key governance questions concern the interplay between various layers of the stack.

This model also goes hand in hand with supply chain analyses, particularly at the hardware level, by revealing how dependencies and power concentrations emerge in technological systems.[67] At each layer of the stack, power can accumulate, and the stack model helps clarify these concentrations and their broader impact on the technological system.

This perspective also illustrates the interconnections between AI systems and other digital infrastructures – such as the internet, online platforms and data centers – and shows how different actors (industry, governments, NGOs, academia and communities) both rely on and influence one another.[68]

Researchers from the Ada Lovelace Institute note that the resource-intensive nature of AI development often renders "downstream" users dependent on "upstream" providers, typically large AI companies. This dynamic underscores the need for policymakers to understand how value is created and distributed across the AI stack.[69]

In a stack controlled by a commercial actor, the company often pursues vertical integration, aiming to control the entire stack rather than incorporating third-party components. This can lead to monopolistic power, particularly in digital platforms or cloud infrastructure.[70] An alternative strategy seeks to "commoditize the complement," that is, to obtain monopoly power in a single layer, while fostering competition in – and thus commoditize – other layers. From a business perspective, the stack model offers a way to analyze where profits are generated, accounting for both dependencies (such as on GPUs) and competition in an environment where many solutions are openly shared.[71]

A public approach, on the other hand, focuses not on control, but on orchestrating the various components and layers to achieve public interest goals. A layered approach, based on the stack metaphor, allows for better governance of AI.[72] It takes into account the complexity of AI systems, while also demonstrating their interdependent nature. It allows for examination of dependencies at different layers, as well as the benefits of sharing key resources, and their impact on model development. For example, policies that consider the entire stack can address more than just compute resources. The European AI Continent Action Plan is an example of such a "full-stack" approach, as it includes measures on computing power, training data, models and deployment of AI systems.[73]

The stack metaphor also helps answer two key questions for a public AI strategy: whether a fully public AI stack is possible and, if not, what types of interventions across the AI stack can best generate public value while minimizing dependencies. In Chapter 5, we recommend pathways to public AI based on this analysis.

The first question closely relates to sovereign AI strategies, which aim to give nation-states independent control over a domestic AI stack. This idea is strongly promoted by Nvidia, which holds a domi-

[66] José Van Dijck. "Seeing the Forest for the Trees: Visualizing Platformization and Its Governance." New Media & Society, vol. 23, no. 9, Sept. 2021, pp. 2801–19. DOI.org (Crossref), https://doi.org/10.1177/1461444820940293

[67] Eleanor Shearer, Matt Davies and Mathew Lawrence. "The Role of Public Compute." Ada Lovelace Institute. 24 April 2024. https://www.adalovelaceinstitute.org/blog/the-role-of-public-compute/ Accessed 3 April 2025, Ana Validivia, ibid.

[68] Victoria Ivanova et al. "Future Art Ecosystems. Vol.4 Art x Public AI." Serpentine Labs. 2025. https://reader.futureartecosystems.org/briefing/fae4/

[69] Sabrina Küspert, Nicolas Moës, Connor Dunlop. ibid.

[70] Cecilia Rikap. "Antitrus t Policy and Artificial Intelligence: Some Neglected Issues." Institute for New Economic Thinking. 10 June 2024. https://www.ineteconomics.org/perspectives/blog/antitrust-policy-and-artificial-intelligence-some-neglected-issues

[71] Matt Bornstein, Guido Appenzeller, and Martin Casado. "Who Owns the Generative AI Platform?" Andreessen Horowitz. 19 January 2023. https://a16z.com/who-owns-the-generative-ai-platform/

[72] Jakob Mökander, et al. "Auditing Large Language Models: A Three-Layered Approach." AI and Ethics, vol. 4, no. 4, Nov. 2024, pp. 1085–115. Springer Link, https://doi.org/10.1007/s43681-023-00289-2

[73] "AI Continent Action Plan." European Commission. 9 April 2025. https://commission.europa.eu/topics/eu-competitiveness/ai-continent_en

nant position at the hardware layer and counts states seeking control over compute power among its key customers.[74] A sovereign AI program requires, in principle, full control of the AI stack – making it a highly contested concept. As Pablo Chavez notes, this is difficult to achieve: "In reality, what most countries working toward AI sovereignty are doing is building a Jenga-like AI stack that gives them enough control and knowledge of AI technology to understand and react to changing technology, market and geopolitical conditions but falls short of complete control."[75] In the following chapters, we offer a vision of public AI that does not seek sovereign control but instead aims to secure the ability to orchestrate resources across the AI stack in service of the public good.

Concentrations of power in the AI stack

Public AI visions, while not focused on digital sovereignty per se, must confront the question of whether developing AI systems in the public interest requires some form of "sovereignty" – that is, control over the AI stack. In other words, public AI must address the concentrations of power that exist at various layers of the stack, where key resources are held by commercial actors with dominant or near-monopolistic positions. This is especially true at the compute layer. The scale of investment needed to develop viable alternatives makes such control extremely difficult, if not impossible. As a result, the value generated by AI systems is increasingly privatized and, in some cases, monopolized. Public AI policies aim to mitigate this trend.

Market concentration is an outcome of the immense costs of training and deploying transformer-based generative AI models. Only a few companies can afford the costs to train state-of-the-art models. While DeepSeek initially appeared to mark a shift in the economics of AI training, later analysis suggested otherwise.[76] In addition to the high cost of acquiring GPUs, building a data center with sufficient networking infrastructure and covering operational expenses – such as electricity for running and cooling hardware – requires major investment. As a result, only a few of the largest AI companies (Amazon, Google, Meta, xAI and Microsoft) are able to pursue a full-stack approach, which demands massive investments in proprietary data centers.[77] Among these, Google and Amazon have a fully integrated AI stack, having developed their own chips (Google's TPU and Amazon's Inferentia and Trainium). Others still depend on Nvidia, which holds a monopolistic position in the GPU market.

The costs, however, do not end with training. Deploying large AI models is also expensive, as it requires sustained access to significant compute resources to process user queries in real time. For instance, OpenAI's ChatGPT reportedly incurred daily operating costs of up to $700,000 in 2023, due to the need to continuously run thousands of GPUs.[78]

This financial burden has pushed leading AI companies without full-stack capabilities – such as OpenAI, Anthropic and Mistral AI – to form partnerships with cloud hyperscalers like Amazon Web Services, Microsoft Azure and Google Cloud. This has resulted in a circular flow of capital between AI startups and scale-ups on the one hand and these cloud hyperscalers on the other hand. It is estimated that the three cloud hyperscalers "contributed a full two-thirds of the $27 billion raised by fledgling AI companies in 2023"[79]

[74] Angie Lee. "What is sovereign AI?" Nvidia. 28 February 2024. https://blogs.nvidia.com/blog/what-is-sovereign-ai/ Accessed 3 April 2025.

[75] Pablo Chavez. "Sovereign AI in a Hybrid World: National Strategies and Policy Responses." Lawfare, 7 November 2024. https://www.lawfaremedia.org/article/sovereign-ai-in-a-hybrid-world--national-strategies-and-policy-responses

[76] Dario Amodei. ibid.

[77] Ben Thompson. "AI Integration and Modularization." Stratechery. 29 May 2024. https://stratechery.com/2024/ai-integration-and-modularization/

[78] Aaron Mok. "It Costs OpenAI Millions of Dollars a Day to Run ChatGPT, Analyst Estimates." Business Insider. 25 April 2023. https://www.businessinsider.com/how-much-chatgpt-costs-openai-to-run-estimate-report-2023-4

[79] George Hammond. "Big Tech Outspends Venture Capital Firms in AI Investment Frenzy." Financial Times, 29 December 2023. https://www.ft.com/content/c6b47d24-b435-4f41-b197-2d826cce9532

and that the majority of capital raised by AI startups, "up to 80-90% in early rounds," was paid back to the same cloud hyperscalers.[80]

This uneven allocation of AI's resources – and of the financial returns they generate – is a defining feature of modern AI technologies. The field exhibits its characteristics of a natural monopoly, driven by the high cost of training and deploying AI systems. These costs stem from intense demand for computing power, the high price of chips, the effort required to obtain and prepare large datasets, limited access to proprietary data and the high switching costs between cloud platforms. Economies of scale in generative AI create "winner-takes-most" dynamics, which are reinforced by network effects and first-mover advantages, as early, large-scale systems benefit from user-generated data and established customer bases. This inequality is global, not limited to any single jurisdiction.[81]

These concentrations of power occur, first of all, at the compute layer. AI development efforts are highly dependent on the three dominant cloud providers – Amazon, Microsoft and Google – and on Nvidia, which currently holds an overwhelming share of the chip market.[82] At the data layer, leading AI development labs typically benefit also from their privileged access to proprietary data generated on platforms they own or control. This trend is exemplified by the merger of xAI and X, an AI company and a social media platform respectively, both owned by Elon Musk.[83]

Closely related to this is the emergence of a "compute divide."[84] Outside a small group of hyperscalers and AI labs that have partnered with them, most companies have to rent compute for AI development and deployment. These costs have resulted in a divide between the "GPU rich" and "GPU poor" companies.[85] A similar "computing divide" also exists between commercial labs and academic or nonprofit research institutions.[86] On a global scale, the uneven distribution of GPU-equipped data centers has produced a new kind of digital divide. Countries are now classified into three tiers: "Compute North" nations with advanced GPU data centers capable of developing cutting-edge AI, "Compute South" nations with less-powerful facilities suitable for deploying existing AI, and "Compute Desert" nations that lack such infrastructure entirely and must rely on foreign computing resources.[87]

Across the world, public and non-commercial computing resources are miniscule in comparison to commercial computing power. While the public sector was an early mover – developing the first supercomputers for research purposes – today its investments are outpaced by the growth of commercial compute capacity, as outlined above. In addition, public supercomputers must support a wide range of research unrelated to generative AI and are therefore neither optimized for AI training nor available for providing inference compute to deployed AI systems. As a result, even nonprofit and academic initiatives

[80] Matt Bornstein, Guido Appenzeller, and Martin Casado. ibid.

[81] Competition and Markets Authority. "AI Foundation Models: Update Paper." GOV.UK, 16 April 2024. https://www.gov.uk/government/publications/ai-foundation-models-update-paper; Anselm Küsters, Matthias Kullas. "Competition in Generative Artificial Intelligence." CEP. 12 March 2024. https://www.cep.eu/eu-topics/details/competition-in-generative-artificial-intelligence-cepinput.html; Tejas N. Narechania and Ganesh Sitaraman. "Antimonopoly Tools for Regulating Artificial Intelligence." SSRN. 25 September 2024. https://www.ssrn.com/abstract=4967701

[82] Jai Vipra and Sarah Myers West. "Computational Power and AI." AI Now Institute. September 27, 2023. https://ainowinstitute.org/publication/policy/compute-and-ai

[83] Maxwell Zeff. "Elon Musk says xAI acquired X." TechCrunch. 29 March 2025. https://techcrunch.com/2025/03/29/elon-musk-says-xai-acquired-x/

[84] Bridget Boakye, et al. "State of Compute Access: How to Bridge the New Digital Divide." Tony Blair Institute. 7 December 2023. https://institute.global/insights/tech-and-digitalisation/state-of-compute-access-how-to-bridge-the-new-digital-divide

[85] Alistair Barr. "The tech world is being dividing into 'GPU rich' and 'GPU poor.' Here are the companies in each group." Business Insider Nederland, 28 August 2023, https://www.businessinsider.nl/the-tech-world-is-being-dividing-into-gpu-rich-and-gpu-poor-here-are-the-companies-in-each-group/

[86] Tamay Besiroglu, et al. The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny? arXiv:2401.02452, arXiv, 8 Jan. 2024. arXiv.org, https://doi.org/10.48550/arXiv.2401.02452

[87] Vili Lehdonvirta, Bóxī Wú and Zoe Hawkins. (2024). "Compute North vs. Compute South: The Uneven Possibilities of Compute-based AI Governance Around the Globe." Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 7(1), 828-838. https://doi.org/10.1609/aies.v7i1.31683

now typically rely on corporate infrastructure. Public investments in compute are also constrained by fiscal, infrastructural and ecological limitations.

A global mapping of public compute initiatives, conducted by the Ada Lovelace Institute, shows that most rely on hybrid provisioning and funding models. They depend on private sector resources and expertise, while aiming at ensuring public oversight of commercial compute.[88]

Cecilia Rikap argues that the power exerted by the largest AI companies goes beyond conventional forms of market concentration, as it is not limited to ownership of technology.[89] These companies also employ a range of strategies – including leveraging venture capital to secure preferential access to knowledge and capabilities, entrenching their positions through cloud services and vertical integration, capturing AI talent and influencing research agendas – to consolidate control and privatize the value generated by others.

These concentrations of power pose significant challenges for developing public AI infrastructures, which face dependencies on monopolistic or oligopolistic actors across multiple layers of the stack. Two potential strategies have emerged in response. One seeks independence at the compute layer through massive investments in data centers and computing power on a scale comparable to commercial expenditures. However, this still entails reliance on Nvidia's GPUs due to the company's entrenched, monopolistic position. The other strategy accepts dependencies at the compute layer and instead focuses on independence at the model layer and in the development of applications built on top of public models. These pathways to public AI are explored in greater detail in chapter 5.

Characteristics of key layers of the stack

In the following sections, we outline in more detail the characteristics of key resources and infrastructures at the three core layers of the AI stack: compute, data and models. These characteristics must be considered when designing public AI policies.

Compute

In the absence of a universally accepted definition, "compute" can refer both to a performance metric – measured in terms of calculations or floating-point operations per second (FLOPs) – and to the physical hardware that performs these calculations,[90] namely semiconductors. The UK government defines compute as "computer systems where processing power, memory, data storage and network are assembled at scale to tackle computational tasks beyond the capabilities of everyday computers."[91]

A 2018 analysis by OpenAI showed that computing power used for AI training had increased 300,000 times since 2012, the beginning of the "deep learning era." A follow-up study by Epoch.ai, which analyzed 120 training runs of machine learning systems, found a fourfold annual growth rate in recent years – making it one of the fastest technological expansions in decades. Overall, training compute has grown by a staggering factor of 10 billion since 2010.[92]

Further scaling, however, faces four key constraints: energy consumption, chip manufactur-

[88] Matt Davies and Jai Vipra. "Mapping global approaches to public compute." Ada Lovelace Institute. 4 November 2024. https://www.adalovelaceinstitute.org/policy-briefing/global-public-compute/

[89] Cecilia Rikap. "Dynamics of Corporate Governance Beyond Ownership in AI." Common Wealth. 15 May 2024. https://www.common-wealth.org/publications/dynamics-of-corporate-governance-beyond-ownership-in-ai

[90] Amlan Mohanty. "Compute for India: A Measured Approach." Carnegie Endowment for International Peace. 17 may 2024. https://carnegieendowment.org/posts/2024/05/compute-for-india-a-measured-approach?lang=en

[91] Department of Science, Innovation and Technology. "Independent Review of The Future of Compute: Final Report and Recommendations." GOV.UK. https://www.gov.uk/government/publications/future-of-compute-review/the-future-of-compute-report-of-the-review-of-independent-panel-of-experts Accessed 23 Apr. 2025.

[92] Jaime Sevilla, et al. "Compute Trends Across Three Eras of Machine Learning." Epoch AI. 16 Feb. 2022. https://epoch.ai/blog/compute-trends

ing, data availability and speed limits inherent to AI training.[93]

To better understand what compute entails – and where bottlenecks arise in the development of generative AI – it is helpful to break this layer into three key component: advanced chips (primarily GPUs), and the two additional elements needed to make them usable at scale: specialized software that enables efficient use of those chips, and data centers where GPUs are networked into large-scale compute systems.

Advanced chips

Chips, or semiconductors, are arguably among the most important technological hardware in use today. They underpin all digital technologies and serve as the backbone of most economic activities. The enormous computational demands of training and deploying AI models – often involving trillions of calculations – depend on modern chips' ability to coordinate the work of billions of transistors etched into each unit.

There are two distinct categories of chips, each with its own supply chains, production requirements and strategic dependencies:

- Memory chips store and enable access to data, which is essential for high-performance AI workloads and a prerequisite for any computational task.

- Logic (or processing) chips – including CPUs, GPUs and specialized chips like TPUs – carry out computations.

The rapid advancements in AI have been driven largely by improvements in specialized logic chips. As traditional CPUs proved too slow for AI training, GPUs – originally developed for graphics rendering – have been repurposed and optimized for this purpose.[94]

Software frameworks to run chips

Recognizing the role of specialized software is essential – it acts as the bridge between hardware and infrastructure and helps explain much of the current concentration of power in the generative AI ecosystem.

Effective use of GPUs for training and deploying AI models requires specialized software frameworks and tools. Two well-known examples are Nvidia's Compute Unified Device Architecture (CUDA) and AMD's Radeon Open Compute (ROCm). CUDA, in particular, has become the industry standard for GPU-accelerated computing and quickly gained a first-mover advantage. Introduced in 2007, it enables developers to harness the parallel processing power of GPUs for general-purpose tasks critical to AI training and deployment.[95]

CUDA's importance became especially clear with the development of AlexNet in 2012, when the framework enabled the training of the neural network and reduced computation time from weeks to hours.[96] Today, leading deep learning frameworks like PyTorch and TensorFlow – open sourced by Meta and Google, respectively – are deeply integrated with CUDA, making it difficult for competing platforms to gain traction. While media coverage often highlights Nvidia's GPUs, CUDA represents an equally powerful competitive "moat" for the company.

In contrast, ROCm, developed by AMD, is an open source alternative designed to offer similar functionality. Although ROCm supports various programming models and provides an open platform, it has struggled to match CUDA's widespread adoption. Nvidia's extensive investments in its developer ecosystem have helped cement CUDA as the de facto standard

[93] Jamie Sevilla. "Can AI Scaling Continue Through 2030?" Epoch AI. 20 Aug. 2024. https://epoch.ai/blog/can-ai-scaling-continue-through-2030

[94] More recently, other custom AI accelerators such as Google's tensor processing units (TPUs) or Amazon's Trainium chips have been developed or are in development. However, as the dominant logic chips in AI are GPUs primarily supplied by Nvidia, other accelerators are not the focus of this paper.

[95] Fatima Hameed Khan, et al. "Advancements in Microprocessor Architecture for Ubiquitous AI—An Overview on History, Evolution, and Upcoming Challenges in AI Implementation." Micromachines 2021, 12(6), 665; https://doi.org/10.3390/mi12060665

[96] Alex Krizhevsky, et al. "ImageNet Classification with Deep Convolutional Neural Networks." NeurIPS Proceedings. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf

preferred by most AI researchers and companies.[97] The popularity of this proprietary network creates a major dependency for AI development.

Data centers

Data centers are complex facilities that provide computing, storage and network infrastructures. Thanks to data centers, GPUs become integrated into a computing resource that is usable at scale. This includes not only GPUs but networking interfaces and other hardware elements. The storage infrastructure holds the vast datasets required for AI training, while the network infrastructure connects system components within the data center, transfers data between compute nodes, and links the facility to end users and global cloud networks.[98]

The cost of building and operating data centers is immense, encompassing hardware, energy and cooling systems, network infrastructure and personnel. State-of-the-art data centers are powered by tens of thousands of high-end GPUs, primarily supplied by Nvidia. For example, the AI training cluster built by xAI in 2024 to train the large language model Grok – one of the largest of its kind – began with 100,000 Nvidia H100 GPUs and expanded to 200,000 by the end of the year.[99] The estimated cost of such an investment is several billion dollars, with the company reportedly raising $6 billion to support it.[100]

Data

Data is a critical resource for AI systems. The development of modern AI became possible only when vast amounts of data became available in digital form on the internet. In this context, the term data is used as a catch-all phrase encompassing all types of information, content and data sources. Rules and norms around access and use have shifted in recent years, as AI training emerged as a disruptive new use of data at massive scale. This has reopened long-standing debates about copyright and sparked new concerns about fair data use and the risk of exploitation by a small group of dominant commercial AI companies.

Under the current AI scaling paradigm, building more capable models requires ever-larger datasets. Initially, the development process appeared open-ended, as developers tapped into as much accessible data as possible. Today, however, data is increasingly viewed as a finite resource. Researchers from Epoch.ai predict that a “peak human data” moment may occur between 2026 and 2032, when further gains within the current paradigm may no longer be possible due to data scarcity.[101] Research conducted by the Data Provenance Initiative also shows that, in response to data use for AI development, various actors are taking steps to reduce availability of content that they publish on the web.[102]

The development of AI systems faces a paradox when it comes to data: it is at once abundant and scarce. On one hand, the fact that the entire web’s content became a foundational resource for training all of the dominant commercial generative models, whose creators accessed it for free, proves the abundance of data. On the other hand, much of the available data remains proprietary. For the largest commercial AI enterprises, access to restricted, proprietary data – often amassed through consumer-facing digital

[97] Serhii Nakonechnij. “ROCm vs CUDA Practical Comparison.” Scimus. 12 August 2024 https://thescimus.com/blog/rocm-vs-cuda-a-practical-comparison-for-ai-developers/

[98] Aadya Gupta and Adarsh Ranjan. “A primer on compute. Carnegie Endowment for International Peace.” 30 April 2024. https://carnegieendowment.org/posts/2024/04/a-primer-on-compute?lang=en

[99] Mark Mantel. “xAI Has Apparently Completed the World’s Fastest Supercomputer.” Heise Online. 4 September 2024. https://www.heise.de/en/news/xAI-has-apparently-completed-the-world-s-fastest-supercomputer-9857540.html

[100] Wayne Williams. “Elon Musk raises USD 6 billion for xAI’s Memphis data center; will purchase 100,000 NVIDIA chips to boost Tesla’s full self-driving FSD capabilities.” 28 November 2024. https://www.techradar.com/pro/elon-musk-raises-usd6-billion-for-xais-memphis-data-center-will-purchase-100-000-nvidia-chips-to-boost-teslas-full-self-driving-fsd-capabilities Accessed 3 April 2025.

[101] Jamie Sevilla, et al. “Can AI scaling continue through 2030?.” Epoch AI. 20 August 2024. Available at: https://epoch.ai/blog/can-ai-scaling-continue-through-2030

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

platforms controlled by the same companies – serves as a key competitive advantage.[103] For others, the lack of access to this high-quality data presents a significant competitive disadvantage.

Data sources for AI training

When discussing data and datasets in the context of AI training, it is important to recognize that data is not a homogeneous concept. Generative AI models rely on diverse data sources for training, which can be categorized by accessibility, licensing, structure and sensitivity.

Accessibility is the most important category, and distinctions should be made between private (proprietary), public and openly shared data. Private data typically includes user-generated content collected by companies that own dominant online platforms and are now building generative AI models (e.g., Meta, Google, Microsoft, AWS). Incumbent AI companies like OpenAI and Anthropic can also use data generated by their own chatbots. Public data includes content from the open internet, either scraped directly by AI companies or aggregated into datasets such as Common Crawl and its derivatives. Finally, openly licensed data – such as that from Wikimedia – is valued by developers for both its quality and the legal certainty it offers for training use.

Data comes in many forms, each governed by different legal frameworks and requiring tailored governance. The use of data for AI training often raises copyright issues,[104] leading to a growing number of high-profile infringement lawsuits brought by creators, publishers and rights holders against leading AI firms. These legal challenges question the extent to which copyrighted works can be used for AI training without explicit permission or licensing, especially under the fair use doctrine.

There is also growing evidence that some AI labs have used data without permission, possibly in violation of the law – as demonstrated by multiple court cases involving major AI companies. For example, the Books3 dataset, which included 183,000 books sourced from pirate websites, was used to train early-generation models released in 2022.[105] While its use was eventually discontinued under pressure from rights holders, Meta was reported to have trained models on LibGen – a similar pirate repository – as late as 2024.[106] In 2025, Anna's Archive, an aggregator of pirated books and research articles, announced it had granted AI companies, including DeepSeek,[107] access to its database. These examples show that AI training often operates in a legal gray area – and sometimes outside the boundaries of the law – when it comes to the use of data.

Several copyright frameworks have been introduced to regulate generative AI training, most notably the European Union's exception for text and data mining, which includes opt-out provisions for commercial training under the 2019 Digital Single Market Directive. However, there remains a lack of clarity around their interpretation and enforcement, and these frameworks are increasingly contested, as shown by ongoing policy debates about opt-outs in the United Kingdom.[108]

Other types of training data may consist of personal data (subject to personal data protection laws) and

[103] Cade Metz, et al. "How Tech Giants Cut Corners to Harvest Data for A.I.." New York Times. 8 April 2024. https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html

[104] Daniel J. Gervais. "The Heart of the Matter: Copyright, AI Training, and LLMs." SSRN. 1 November 2024. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4963711 Accessed 3 April 2025; Matthew Sag. "Fairness and Fair Use in Generative AI." SSRN. 20 December 2023. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4654875 Accessed 3 April 2025.

[105] Alex Reisner. "These 183,000 Books Are Fueling the Biggest Fight in Publishing and Tech." The Atlantic, 25 September 2023. https://www.theatlantic.com/technology/archive/2023/09/books3-database-generative-ai-training-copyright-infringement/675363/

[106] Alex Riesner. "The Unbelievable Scale of AI's Pirated-Books Problem." The Atlantic. 20 March 2025. https://www.theatlantic.com/technology/archive/2025/03/libgen-meta-openai/682093/

[107] Anna's Archive. "Copyright Reform Is Necessary for National Security." Anna's Archive (blog). 31 January 2025. https://annas-archive.li/blog/ai-copyright.html Accessed 3 April 2025.

[108] Joseph Bambridge and Dan Bloom. "UK Plans Fresh Round of Talks to Take Sting out of AI Copyright Proposals." POLITICO, 3 April 2025, https://www.politico.eu/article/uk-plans-fresh-round-talks-lawmakers-ai-copyright-proposals/

non-personal data industrial, anonymized, statistical and administrative datasets. While general principles can be applied to all types of data, there is no "one size fits all" approach to data sharing. Copyright and privacy or personal data rights are the two most important factors determining how data can be shared and accessed – and how "open" it can be.

The ongoing public debates about AI training datasets often focus exclusively on pretraining data. Copyright and data governance laws attempt to strike a balance that allows publicly available internet data to be reused in training datasets without enabling its exploitation. However, there is a tension between the view – common among AI developers – that data is a raw resource to be used freely, and the perspectives of those who create and steward that data.

It is also important to note that datasets used in post-training stages play an increasingly significant role. For example, methods based on RLHF require data on human preferences, typically sets of generative AI prompts and responses. New types of reasoning models rely on datasets designed to help models follow instructions, solve problems or evaluate results.[109] And as this type of training becomes increasingly important, public discussions will also need to address the provision and governance of these datasets, which are often constructed differently and do not raise the same legal concerns.[110] First, fine tuning datasets are usually domain- or application-specific and require governance tailored to their context. Many key domains for generative AI – such as health and finance – depend on sensitive data. Second, benchmarks for evaluating model capabilities are key tools that – unlike pretraining data – can often be developed and shared as digital public goods. Since standardized benchmarks guide generative AI development, they benefit from open access and sound governance – though in some cases, particularly for security-related benchmarks, openness may be limited.[111]

Synthetic data

Synthetic data, generated by generative AI models, is a distinct category of data increasingly used in AI development, and it presents its own governance challenges. Because it is synthetically produced, it can be created quickly and cheaply. It is not subject to intellectual property restrictions and typically avoids privacy and other rights-related concerns. For instance, synthetic health data can be used in place of real patient data to protect privacy. In general, synthetic data can be used to reflect real-world patterns while helping to safeguard privacy or reduce bias. Some models – such as Microsoft's Phi family of small language models – have been trained entirely on synthetic data.[112] Recently, a new model development paradigm, called model distillation, uses an approach similar to training with synthetic data. In this paradigm, a "teacher" generative AI model produces outputs that a "student" model is then trained to replicate, bypassing the need for access to the original pretraining data.

Some researchers are optimistic about the potential of synthetic data, particularly for protecting personal data during AI training.[113] Others warn of associated risks, most notably model collapse – a hypothesized decline in performance when models are trained on synthetic rather than real data.[114] Use of synthetic data remains contested, and the validity of the "model

[109] Nathan Lambert. "The State of Post-Training in 2025." Interconnects. 12 March 2025, https://www.interconnects.ai/p/the-state-of-post-training-2025

[110] Nathan Lambert. "Why reasoning models will generalize." Interconnects. 28 January 2025, https://www.interconnects.ai/p/why-reasoning-models-will-generalize Accessed 3 April 2025.

[111] Peter Mattson, et al. "Perspective: Unlocking ML requires an ecosystem approach." MLCommons. 10 March 2023. https://mlcommons.org/2023/03/unlocking-ml-requires-an-ecosystem-approach/

[112] Microsoft. "Phi open model family." Microsoft. https://azure.microsoft.com/en-us/products/phi/ Accessed 23 April 2025.

[113] Philippe De Wilde, et al. "Recommendations on the Use of Synthetic Data to Train AI Models." Tokyo: United Nations University, 2024. https://collections.unu.edu/eserv/UNU:9480/Use-of-Synthetic-Data-to-Train-AI-Models.pdf

[114] Ilia Shumailov, et al. "AI Models Collapse When Trained on Recursively Generated Data." Nature, vol. 631, no. 8022, July 2024, pp. 755–59. www.nature.com, https://doi.org/10.1038/s41586-024-07566-y; University of Oxford. "New Research Warns of Potential 'Collapse' of Machine Learning Models." Department of Computer Science, 25 July 2024. https://www.cs.ox.ac.uk/news/2356-full.html

collapse" hypothesis has been questioned by other researchers.[115] It is unclear whether the growing presence of generative AI content on the web will also have an impact on pretraining AI models with web content. And most probably, synthetic data can help with some challenges (like responsible AI training) but not overcome other issues, such as the lack of sufficiently varied and complex data to train more capable models.

Models

Generative AI models are statistical models trained to process a given type of input into a given type of output. At their core, models consist of:

- their architecture (e.g., neural networks, transformers or diffusion networks), and
- their parameters (i.e., the weights and biases that have been optimized through training).

In the field of generative AI, models are primarily neural networks built on the transformer architecture or its variants.[116]

In the following section, we provide an overview of different types of models and their development pathways. This section explains the distinctions between foundation models, LLMs and small models, while detailing various methods of model derivation. The ability to derive new models from existing ones, provided that the latter are shared openly, fosters an ecosystem of collaboration and resource sharing.

There is currently no consistent typology for categorizing generative AI models. Terms such as foundation models, general-purpose AI, large language models and small language models are often used to describe different types of models, typically differentiated by their parameter size.

Foundation models

As defined by Stanford University's Center for Research on Foundation Models, foundation models "are trained on broad data at scale and are adaptable to a wide range of downstream tasks."[117] These models exemplify AI development in the transformer paradigm and require vast compute and data resources. They have general capabilities that allow them to serve as the basis for developing more specialized models. Because of their adaptability, they function as general-purpose technologies with infrastructural characteristics.[118] The legal concept of general-purpose AI, introduced in the European Union's AI Act, is based on this idea, and the term LLMs typically describes the same type of models. While foundation models are not necessarily multimodal, an increasing number of foundation models are multimodal in their capabilities. For example, OpenAI's GPT-4 is a foundation model that can process both text and images.

Small models

Small models, including small language models and small vision models, are compact, efficient alternatives to LLMs that are designed to balance performance with resource requirements.[119] Their size typically ranges from a few million to a few billion parameters, in contrast to larger models, which may contain hundreds of billions. The term small model generally does not apply to task-specific models like BERT, which perform individual tasks such as summarization or categorization, even though such models still account for a large share of industrial AI applications. Like their larger counterparts, small models are usually based on transformer architectures and can be trained from scratch or derived from foundation models using techniques such as distillation, pruning and quantization. Small models offer advantages in efficiency, accessibility and

[115] Rylan Schaeffer, et al. "Position: Model Collapse Does Not Mean What You Think." arXiv:2503.03150, arXiv, 18 Mar. 2025. arXiv.org, https://doi.org/10.48550/arXiv.2503.03150

[116] Dan Hendrycks, et al. "Measuring Massive Multitask Language Understanding." arXiv:2009.03300, arXiv, 12 Jan. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2009.03300

[117] Rishi Bommasani, et al. ibid.

[118] Alison Gopnik. "What AI Still Doesn't Know How to Do." The Wall Street Journal. 15 July 2022. https://www.wsj.com/tech/ai/what-ai-still-doesnt-know-how-to-do-11657891316

[119] https://medium.com/@nageshmashette32/small-language-models-slms-305597c9edf2

customization, making them ideal for deployment on resource-constrained devices, edge computing environments and domain-specific applications. They require less computational power and memory, enabling wider adoption by a range of stakeholders while still maintaining strong performance in targeted use cases.

Open models

Open models are AI models whose architecture and trained parameters (i.e., weights and biases) are released under open source licenses.[120] Since EleutherAI's release of GPT-Neo[121] as an open alternative to OpenAI's GPT in 2021, it has become increasingly common for AI researchers and developers to release open models. For example, in 2023, 66% of foundation models were released as open models, and more than 1.5 million models are hosted on the Hugging Face Hub.[122]

However, there is currently no standard approach to open releases of AI models, and many so-called open models come with significant limitations – such as withholding training data, using restrictive licenses or prohibiting commercial use – compared to the norms established in open source software development. Often, references to "open source models" are viewed as attempts at open-washing, diluting traditional open source standards in the context of generative AI.[123]

In recent years, efforts have been made to more precisely define what constitutes an open model. This is a necessary step toward creating standardized methods for governing and sharing AI models. Frameworks such as the Model Openness Framework[124] by the Generative AI Commons and the Framework for Openness in Foundation Models[125] by the Mozilla Foundation list up to 16 components that extend beyond model architecture and parameters. These include code components (for training, evaluation and inference), data components (for training, post-training and evaluation), and documentation (such as model cards and dataset cards).

A handful of research labs and nonprofit initiatives – such as the Barcelona Supercomputing Center, the Allen Institute for AI and EleutherAI – aim to set a higher bar for open source AI by releasing all model components openly, including trained parameters, code, data and documentation, all under free and open source licenses.

While most leading AI companies keep their models closed and accessible only via commercial APIs, Meta has adopted a model openness strategy. However, its models fall short of commonly accepted open source standards due to limitations imposed by its custom licenses.[126] Other AI companies, including Mistral, DeepSeek or Cohere, have also released open models. In recent months, DeepSeek is seen as the strongest example of an AI lab that combines commercial goals with an open source mission.

Open models enable model derivation, enabling the creation of smaller and more specialized models that offer benefits beyond those associated with scale. This was demonstrated by DeepSeek, whose researchers distilled the reasoning capabilities of the DeepSeek-R1 model (671 billion parameters) into smaller models ranging from 1.5 billion to 70 billion parameters. Within a week of R1's release on the Hugging Face platform, more than 500 derivative

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

[121] Sid Black, et al. "GPT-NeoX-20B: An Open-Source Autoregressive Language Model." arXiv:2204.06745, arXiv, 14 Apr. 2022. arXiv.org, https://doi.org/10.48550/arXiv.2204.06745

[122] Yolanda Gil and Raymond Perrault. "Artificial Intelligence Index Report 2025." Stanford University Human-Centered Artificial Intelligence. 7 April 2025. https://hai-production.s3.amazonaws.com/files/hai_ai_index_report_2025.pdf

[123] David Gray Widder, et al. "Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI." 4543807, 17 Aug. 2023. Social Science Research Network, https://doi.org/10.2139/ssrn.4543807

[124] Matt White, et al. ibid.

[125] Adrien Basdevant, et al. ibid.

[126] Stefano Maffuli. "Meta's LLaMa License Is Not Open Source." Open Source Initiative, 20 July 2023, https://opensource.org/blog/metas-llama-2-license-is-not-open-source

versions had been shared, most of them quantized.[127] Among these derivative models, worth noting is an open source derivative created by Perplexity.ai, which "un-censored" the original Chinese model.[128]

It is useful to view the various AI development projects as part of a broader ecosystem rooted in the sharing of knowledge, components or entire models. Elements of this collaborative approach can be found across the AI landscape, including in major commercial labs. At the same time, a more narrowly defined ecosystem has emerged, composed of teams committed to open source AI and aligned approaches to trustworthy or responsible AI.[129] Distributed development efforts and AI collaboratives like the Big Science project are among the strongest examples of ecosystem building.[130] A mature AI development ecosystem also includes interoperability and data-sharing standards, benchmarks, best practices for responsible AI development and governance frameworks.[131]

[127] Florent Daudens. "Yes, DeepSeek R1's Release Is Impressive. But the Real Story Is What…." LinkedIn. https://www.linkedin.com/posts/fdaudens_yes-deepseek-r1s-release-is-impressive-activity-7289681233427484672-I9MY Accessed 24 Apr. 2025.

[128] AI Team. "Open Sourcing R1 1776." Perplexity (blog). 18 February 2025. https://www.perplexity.ai/hub/blog/open-sourcing-r1-1776

[129] Mark Surman. "Introducing Mozilla.ai: Investing in trustworthy AI." Mozilla (blog). 22 March 2023. https://blog.mozilla.org/en/mozilla/introducing-mozilla-ai-investing-in-trustworthy-ai/

[130] Jennifer Ding, et al. "Towards Openness Beyond Open Access: User Journeys through 3 Open AI Collaboratives." arXiv:2301.08488, arXiv, 20 Jan. 2023. arXiv.org, https://doi.org/10.48550/arXiv.2301.08488

[131] Peter Mattson, et al. ibid.

4 | The public AI framework

In this section, we propose a definition of public AI that builds on the broader concept of a public digital infrastructure. We also draw on three previous papers that define public AI and outline policies that can support its development.

A standard for the publicness of digital infrastructure must go beyond vague notions of the public interest and instead rely on a clear understanding of how public value is created. The definition of a public digital infrastructure achieves this by identifying three key characteristics: attributes, functions and forms of control.

By combining the AI stack framework described in the previous section with this definition of public digital infrastructure, we propose an approach that defines public AI by considering how public attributes, public functions and public control can apply to the different layers of the AI stack. In addition, we offer a gradient of public AI releases that accounts for the fact that public AI can have dependencies on non-public resources, especially at the chip and compute layers.

The concept of public digital infrastructure

Public digital infrastructures are digital infrastructures designed to maximize public value by combining public attributes with public functions and various forms of public control.[132] The concept is related to that of digital public infrastructure (DPI), but focuses on the provision of alternatives to key digital platforms and communication services.[133] These infrastructures are built and governed with the goal of advancing the public interest and maximizing public value. They stand in contrast to extractive solutions that concentrate power in the hands of a few at the expense of the broader population.

This definition, proposed by Open Future, builds on previous work done by researchers from the UCL Institute for Innovation and Public Purpose, aims to more precisely define the public nature of such infrastructures by describing how infrastructures can generate public value. The goal of this "complex unpacking of what 'public' means […] is to shift the focus of the debate from the technical aspects of infrastructure (i.e., making things digital) to its social relevance (i.e., making things public)."[134]

The IIPP report focuses on the first two characteristics of public digital infrastructure. Public attributes refer to the accessibility, openness or interoperability of infrastructure. These features aim to ensure universal and unrestricted access, often through open licensing or interoperability mechanisms.

Public functions of infrastructure means the infrastructure contributes to public goals, rather than

[132] Jan Krewer and Zuzanna Warso. "Digital Commons as Providers of Public Digital Infrastructures." Open Future, 13 November 2024. https://openfuture.eu/publication/digital-commons-as-providers-of-public-digital-infrastructures

[133] For an explanation of the two concepts, see: Jan Krewer. "Signs of Progress: Digital Public Infrastructure Is Gaining Traction." Open Future, 13 March 2024. https://openfuture.eu/blog/signs-of-progress-digital-public-infrastructure-is-gaining-traction

[134] Zuzanna Warso. "Toward Public Digital Infrastructure: From Hype to Public Value." AI Now. 15 October 2024. https://ainowinstitute.org/publication/xii-toward-public-digital-infrastructure-from-hype-to-public-value

Table 1 | Defining publicness: Attributes, functions and control of public infrastructures.

Public interest		Public control		
"For the public"		"Of and by the public"		
Public attributes	**Public functions**	**Public control**	**Public funding**	**Public production**
Infrastructure is publicly accessible, open or interoperable.	Infrastructure contributes to attaining public interest goals and creating public goods.	Infrastructure is governed or overseen by the public.	Infrastructure is funded by the public.	Infrastructure is produced by the public.

Source: Own table. Adapted from: https://openfuture.eu/wp-content/uploads/2024/11/241113_Digital-Commons-as-Providers-of-Public-Digital-Infrastructures.pdf

merely serving as an alternative provider of market-based goods or services. These goals can include enabling civic participation, fostering community and social relationships, stimulating economic activity, improving quality of life or securing essential capabilities. Public infrastructure often creates public goods – resources with social rather than purely market value. Brett Frischmann cites research as an example of such a public good.[135] Public functions of infrastructure often entail filing supply gaps left by market actors.

The underlying concept of the common good, as framed by Mariana Mazzucato, involves both the pursuit of shared objectives and care for shared processes and relationships. It is a perspective that emphasizes the importance of governance in the process of generating social value and positions the state as both a public entrepreneur and a market shaper. A common good perspective underscores the state's role in setting direction and coordinating collective action. Through effective governance, states can ensure co-creation and participation, promote collective learning, secure access, transparency and accountability – all of which are essential to advancing the public interest in digital infrastructure.[136]

Neither public attributes nor public functions alone are sufficient to define publicness. A focus on attributes can be agnostic with regard to how infrastructure is used, and the outcomes of such uses. For example, open source AI solutions can be used in ways that pursue private rather than public goals. Conversely, a functional focus can overlook accessibility. In other words, public interest goals can also be achieved through closed, private infrastructures.

Public digital infrastructure also needs to meet the criterion of public control.[137] This can take various forms, including public oversight, public funding or even public production and provision of infrastructure. Such infrastructure need not be state-owned or produced by public institutions. What matters is the presence of public control, which can also be understood as governance for the common good. This is the minimum necessary condition for digital infrastructure to meet the standard of publicness.

Further on, these three characteristics of public digital infrastructure will be applied to the AI stack to provide an overall definition of public AI. Public control is the most complex characteristic, where instead of a binary choice there are multiple approaches that entail forms of public production, funding or control of infrastructures. Public actors

[135] Brett M. Frischmann. "Infrastructure: The Social Value of Shared Resources." Oxford Academic, 24 May 2012. https://doi.org/10.1093/acprof:oso/9780199895656.001.0001

[136] Mariana Mazzucato. "Governing the Economics of the Common Good: From Correcting Market Failures to Shaping Collective Goals." Journal of Economic Policy Reform, vol. 27, no. 1, Jan. 2024, pp. 1–24. DOI.org (Crossref), https://doi.org/10.1080/17487870.2023.2280969

[137] Open Futures' report on Public Digital Infrastructures describes this characteristic in terms of public ownership. In this report, we rephrase this as public control, which encompasses also forms of public ownership but is not limited to them. See: Jan Krewer and Zuzanna Warso, ibid.

do not necessarily need to fully produce or own such infrastructure – what matters is their ability to orchestrate other actors in support of public digital infrastructures that meet the remaining characteristics.

Public, private and civic actors in public digital infrastructure

Ownership of public digital infrastructures – and the respective roles of public, private and civic actors – is a key issue. In the context of generative AI, this is largely a question of who owns or controls computing power, the key dependency for building public AI. Proper forms of public control can ensure sustainability, while some forms of private control risk creating a situation in which rewards are privatized and risks are socialized.

Governments and public institutions must therefore play a central role in the development and governance of public digital infrastructure, including the public AI stack. As noted by the World Bank, governments should have "a primary role and responsibility in deciding whether and how digital public infrastructure is provided in the interests of the broader society and economy."[138] Deployment of such infrastructures is therefore a collective effort involving various actors, but it is the state that plays a key role in orchestrating collective action and ensuring proper outcomes.

This view of government as an orchestrator of outcomes and public value goes beyond the traditional public/private ownership divide. The idea of orchestrating actions of various actors entails "government direction, centrally defined public purpose, and large-scale planning [to be] combined – in still-emergent ways – with market mechanisms, private actors and public input."[139] In doing so, the state not only guides collective action but also protects public infrastructure from being co-opted for private gain. At the same time, strategies used by commercial actors to gain control over the AI stack can be repurposed in service of the mission-driven approach that should characterize public AI policies.

Understanding of the state's role calls for a shift from viewing government as a passive actor or a mere fixer of market failures to recognizing it as an orchestrator capable of coordinating diverse contributors. The deployment of PDIs should be guided by mission-oriented strategies and a market-shaping approach to policy.[140] Generating public value through digital infrastructures is not merely meant to fix the market or fill market gaps – it is a goal in itself. Importantly, public value can be created by various actors, including those in the private sector. The state's ability to steer this co-creation process is more important than its direct production capacity.

Proposals for public AI

Over the past year, several organizations – including the Public AI Network, Mozilla and the Vanderbilt Policy Generator – have introduced frameworks for a public AI agenda. Each proposal outlines a set of conditions intended to maximize public value and safeguard the common good through the development and deployment of AI. In our analysis of these proposals, we identify a set of shared characteristics that define public AI.

Public AI Network

The Public AI Network's policy paper adopts a framing that aligns closely with Mariana Mazzucato's

[138] Vyjayanti T. Desai, et al. "How Digital Public Infrastructure Supports Empowerment, Inclusion, and Resilience." World Bank Blogs, 15 March 2023. https://blogs.worldbank.org/en/digital-development/how-digital-public-infrastructure-supports-empowerment-inclusion-and-resilience

[139] Stephen J. Collier, James Christopher Mizes, and Antina von Schnitzler, "Preface: Public Infrastructures / Infrastructural Publics," Limn, https://limn.it/articles/preface-public-infrastructures-infrastructural-publics/

[140] Mariana Mazzucato. "From Market Fixing to Market-Creating: A New Framework for Innovation Policy." Industry and Innovation, vol. 23, no. 2, Feb. 2016, pp. 140–56. DOI.org (Crossref), https://doi.org/10.1080/13662716.2016.1146124

mission-driven approach to public intervention. It argues that public AI initiatives are essential to safeguarding the common good – a goal unlikely to be achieved if "the next generation of infrastructure … is under the control of a few publicly unaccountable Big Tech firms."[141] Public AI is thus defined as a set of alternatives that "make the advancement of the common good their central goal." To meet this definition, the paper outlines a set of "minimum viable requirements":

- **Public access:** providing everyone with affordable, direct access to AI tools
- **Public accountability:** empowering citizens to shape technological development
- **Permanent public goods:** establishing sustainable foundations for AI development

The concept of public access here entails offering public options for core AI technologies that are otherwise delivered by the market, thereby providing alternatives to commercial offerings that are prone to becoming natural monopolies. This includes access to essential tools for AI development, such as code libraries, training data and compute resources. Public AI also aims to guarantee access to newly created public goods. Public accountability is understood both in terms of compliance with trustworthy AI principles and in fostering public participation in AI development. This includes mechanisms for oversight and a clearly articulated public purpose, centered on societal needs and capabilities deemed valuable by the public.

Finally, the requirement of permanent accessibility is meant to ensure that public AI provides a stable and reliable foundation that is not constrained by private interests. The Public AI Network emphasizes that this does not necessarily imply direct public ownership. Instead, it advocates for strategies that allow public AI "to be sustainably developed and independently maintained as a public good, guaranteeing public control in perpetuity."

Mozilla Foundation

The Mozilla Foundation describes public AI efforts as aimed at "reducing the friction for everyone to build and use AI in a trustworthy manner." Its analysis suggests that the market will prioritize only a narrow set of profitable applications and therefore not build "everything our society needs from AI."[142] In response, Mozilla proposes a public AI agenda centered on building a "robust ecosystem of initiatives" around three core goals:

- **Public goods:** the creation of open, accessible public goods and shared resources at all levels of the AI technology stack;
- **Public orientation:** centering the needs of people and communities, particularly those most underserved by market-led development;
- **Public use:** prioritizing AI applications in the public interest, especially those neglected by commercial incentives or those that are considered inappropriate for private development.

Mozilla frames public AI initiatives in contrast to private AI development, advocating for competitive alternatives to the proprietary models currently dominating the field. At the same time, public AI is not intended to replace private companies, but to coexist with them by offering "a different way of building technology for different needs." Ultimately, the report argues for a public AI ecosystem that is pluralistic and involves public, civic and commercial actors.

The report also identifies three significant risks that could lead public AI solutions to replicate the harms seen in today's AI ecosystem. These include: first, the development of an alternative ecosystem focused solely on creating public goods without a clear public orientation; second, the risk of financial unsustainability without adequate funding; and third, the po-

[141] Public AI network. "Public AI: Infrastructure for the common good." 10 August 2024. https://publicai.network/whitepaper

[142] Nik Marda, Jasmine Sun and Mark Surman. "Public AI. Making AI work for everyone, by everyone." Mozilla. September 2024. https://assets.mofoprod.net/network/documents/Public_AI_Mozilla.pdf

tential overdependence of public AI on governments and public funding.

In this regard, Mozilla's approach diverges from Mazzucato's mission-driven model of AI development by framing public AI as an ecosystem that functions independently of both corporate and governmental control. As the foundation puts it, "We need a resilient and pluralistic AI ecosystem, in which no single entity – whether Big Tech or national governments – can unilaterally decide AI's future." However, this vision does not fully address the current ecosystem's structural dependencies or propose concrete strategies for mitigating them.

Vanderbilt Policy Generator

The white paper "The National Security Case for public AI" presents its approach to public AI in the context of the threats AI systems may pose to democracy and national security – specifically, how they "may threaten the resilience of democracies around the world."[143] Public AI is framed as a dual-purpose strategy: it safeguards democratic values, privacy and other fundamental rights while also providing secure and resilient solutions for national defense and homeland security.

The Vanderbilt Policy Generator's public AI framework has two components:

- Developing a publicly-funded, publicly owned and publicly-operated AI tech stack
- Adopting public-utility-style regulations for layers of the AI tech stack

The report argues for the establishment of public options in AI – publicly provided and managed components of the AI stack and supply chain. This vision calls for substantial government investment to challenge existing monopolies, especially at the hardware layer, and to promote the creation of public data centers, public cloud services, public datasets for AI development and the recruitment of AI talent into government roles.

These public initiatives are to be complemented by stringent public-utility style regulations aimed at private AI companies with monopolistic or oligopolistic market power. Proposed regulatory measures include structural separation rules that aim to dismantle monopolistic control over multiple, interconnected layers of the AI stack; non-discrimination rules to ensure equal access for all actors; and restrictions on foreign ownership, control and investment. Such regulations are deemed necessary to maintain a competitive private industry that can provide private contractors to governments without undermining innovation, effectiveness or resilience.

The proposal offers a vision of the dynamics between public options in AI and a regulated private AI market. Government development of in-house solutions is expected to drive more competitive pricing from private contractors. At the same time, market competition would be further supported through regulatory measures. The proposal seeks to strike a balance between reliance on private companies and the need to build public sector capacity to tackle societal challenges using AI.

Defining public AI infrastructure

All three proposals can be mapped onto the public digital infrastructure framework described in the previous section. The Public AI Network paper, with its emphasis on the common good, aligns most closely with this model. The Vanderbilt Policy Generator offers the strongest argument for addressing concentrations of power within the AI stack and for establishing true public ownership of AI infrastructure.

These proposals also acknowledge that properly establishing the relationship between public and private actors – particularly regarding control and ownership of infrastructure – is a central challenge for public AI initiatives. The Mozilla paper, for instance, offers an important conceptualization of public AI as an ecosystem involving various actors.

[143] Ganesh Sitaraman and Alex Pascal. ibid.

Table 2 | Mapping different definitions of public AI onto the Public Digital Infrastructure framework.

	Public AI network	Mozilla	Vanderbilt
Public attributes	public access, permanent public goods	public goods	public option
Public functions	public accountability	public orientation	public option
Public control	public accountability	public use	public option, public utility regulation

Source: Own table.

BertelsmannStiftung

Drawing on these proposals and the concept of public digital infrastructure, public AI infrastructure can be defined in terms of the three following characteristics of the AI stack:

- **Public attributes:** Public AI provides universal and unrestricted access to components of the stack, enabled through openness and interoperability. Key components are shared as digital public goods, and solutions are built on open standards. Systems and processes are auditable and transparent. These attributes help reduce market concentration and dependency on dominant commercial actors.

- **Public functions:** Public AI delivers foundational systems and services that support broad societal and economic functions, particularly by enabling downstream activities and public benefits. It supports essential public capabilities such as knowledge sharing and civic participation. The public AI stack creates an enabling environment for innovation while safeguarding user rights and social values.

- **Public control:** Public AI involves public control, funding and/or production of the infrastructures underpinning the generative AI stack. This may take various governance forms – from direct government provision to public orchestration of other actors. The goal is to place the AI stack under democratic control, with mechanisms for collective decision-making and accountability. Public ownership should also ensure long-term sustainability, as captured by the idea of “permanent public goods.”

Both a full-stack AI infrastructure and its individual layers or components can exhibit the characteristics of public AI infrastructure. A publicly owned computing resource or an open source AI model, for example, qualifies as public AI infrastructure. However, due to interdependencies between these layers – as discussed in the previous chapter – policy efforts should support full-stack approaches.

INFOBOX | Public AI infrastructure, stack and systems

In this report, the terms public AI stack, public AI infrastructure and public AI systems are used to describe the goals and preferred outcomes of public AI policies. It is therefore important to clarify how these terms differ, as well as how they interconnect.

Infrastructures are facilities, systems or institutions that serve society-wide functions and provide foundations for downstream activities and social benefits. Frischmann defines infrastructures as "shared means for many ends," emphasizing that they should be treated as shared resources.[144]

Public policies policy should focus on building public AI infrastructures for two reasons. First, AI – as a general-purpose technology – has infrastructural characteristics due to its open-ended applications and societal impacts. In this respect, it is similar to earlier digital technologies like the internet. Second, Frischmann's concept of infrastructure as a shared resource implies that commons-based management strategies can enable broad, open access. Emphasizing the infrastructural nature of AI supports the idea of managing it as a public resource to ensure wide accessibility. AI infrastructure can also be understood as an AI stack of interconnected layers.

At the same time, AI technologies are often described as AI systems, which suggests specific instances of a technology that not only provide "means" but also operate, perform tasks and achieve "ends." An AI system is the end product or application that functions on top of the infrastructural layers beneath it. A complete AI system includes the underlying hardware and compute power, data, code and model architecture used to train the model, the trained model itself (i.e., its parameters) and the application layer built on top of the stack. In contrast, a public AI initiative might focus on providing a single type of component, for example by releasing datasets as digital public goods, or by supporting a sustainable model development ecosystem. Each of these components, on its own, should meet the characteristics of a public AI infrastructure.

The public AI policy framework assumes that deploying AI systems alone is not sufficient to achieve social goals. Instead, public AI infrastructure must be built to support the operation of AI systems in the public interest. At the same time, public AI policy can also include the deployment of specific AI systems – also referred to as AI solutions. This focus is especially important at a time when the purpose and impact of AI deployment remain unclear.

The issue is further complicated by the fact that public AI systems can serve infrastructural functions. Generative AI models, for example, are both concrete AI systems and can function as digital infrastructure. Foundation models, also referred to as general-purpose AI models, exhibit such infrastructural characteristics. When open sourced, they can serve as the basis for new or derivative models.

[144] Brett M. Frischmann, ibid.

Gradient of publicness of AI systems

In the previous section, we provided an overview of several proposals for defining public AI infrastructure, alongside a broader definition of public digital infrastructure. Today, there are few – if any – examples of public AI infrastructure that fully meet this definition. The most commonly cited reasons include the lack of publicly owned compute resources, dependence on an oligopoly of cloud and data center providers and the near-monopoly on chip production. Dependence on commercial solutions or components is not inherently problematic, as many public infrastructures rely on commercial commodities. The issue lies in the lack of public capacity to orchestrate the development of infrastructure that is independent of market dynamics – able to generate public goods and serve the public interest.

The three characteristics of public AI infrastructure – public attributes, public functions and public control – together influence the degree of publicness. While some examples of fully public AI infrastructure exist, dependencies at the compute layer often make such deployment difficult. Public computing infrastructure is typically developed through some form of public-private partnership. For this reason, it is useful to consider a gradient of publicness for AI systems.

The gradient describes varying levels of publicness based on the three characteristics – attributes, functions and control – and how these apply across the layers of the AI stack.[145] These three characteristics, taken together, determine an infrastructure's ability to support public interest goals and development trajectories independent from market actors. As such, the gradient serves as a practical tool for policymakers to identify strategies that enhance the publicness of AI systems and strengthen the public value of specific initiatives.

This gradient can be understood as a continuum, ranging from fully public AI systems to semi-public ones, and finally to commercial, closed systems with minimal public functions. As solutions move further along toward publicness, they become less dependent on dominant market players and more capable of enabling public agency and supporting independent, mission-oriented development pathways.

The gradient of publicness also illustrates what must change in AI infrastructures, systems or components to increase their alignment with public interest goals. For example, many open models demonstrate strong public attributes – such as open weights – but fail to fulfill public functions because they are not actively deployed in service of public objectives. Models shared on platforms like Hugging Face Hub may be accessible for experimentation, but are not necessarily integrated into solutions for education, healthcare or climate mitigation.

Further examples include prominent open-weight models like Llama, models by Meta, Mistral and DeepSeek. These models indirectly support public functions by enabling downstream innovation and model development. However, their overall publicness remains limited due to a lack of transparency around training data and processes, which restricts broader access and accountability.

The gradient framework is one of the core contributions of this white paper. It introduces a model that maps AI initiatives along a continuum – from fully public to fully private – based on the dimensions of attributes, functions and control. This approach provides policymakers with a diagnostic and strategic tool for evaluating where a given intervention currently stands and identifying which policy measures – whether investment, regulation or institutional design – could increase its publicness. The framework is particularly relevant when assessing choices at the compute, data and model layers of the AI stack.

Positions on the gradient depend on the extent to which the three characteristics of public AI infrastructure are met. Weak forms of publicness require at least public attributes and some public functions.

[145] The inspiration for this gradient comes from Irene Solaiman's work on a gradient of release approaches for generative AI. See: Irene Solaiman. "The Gradient of Generative AI Release: Methods and Considerations." arXiv:2302.04844, arXiv, 5 Feb. 2023. arXiv.org, https://doi.org/10.48550/arXiv.2302.04844 Accessed 3 April 2025.

Figure 5 | Gradient of publicness

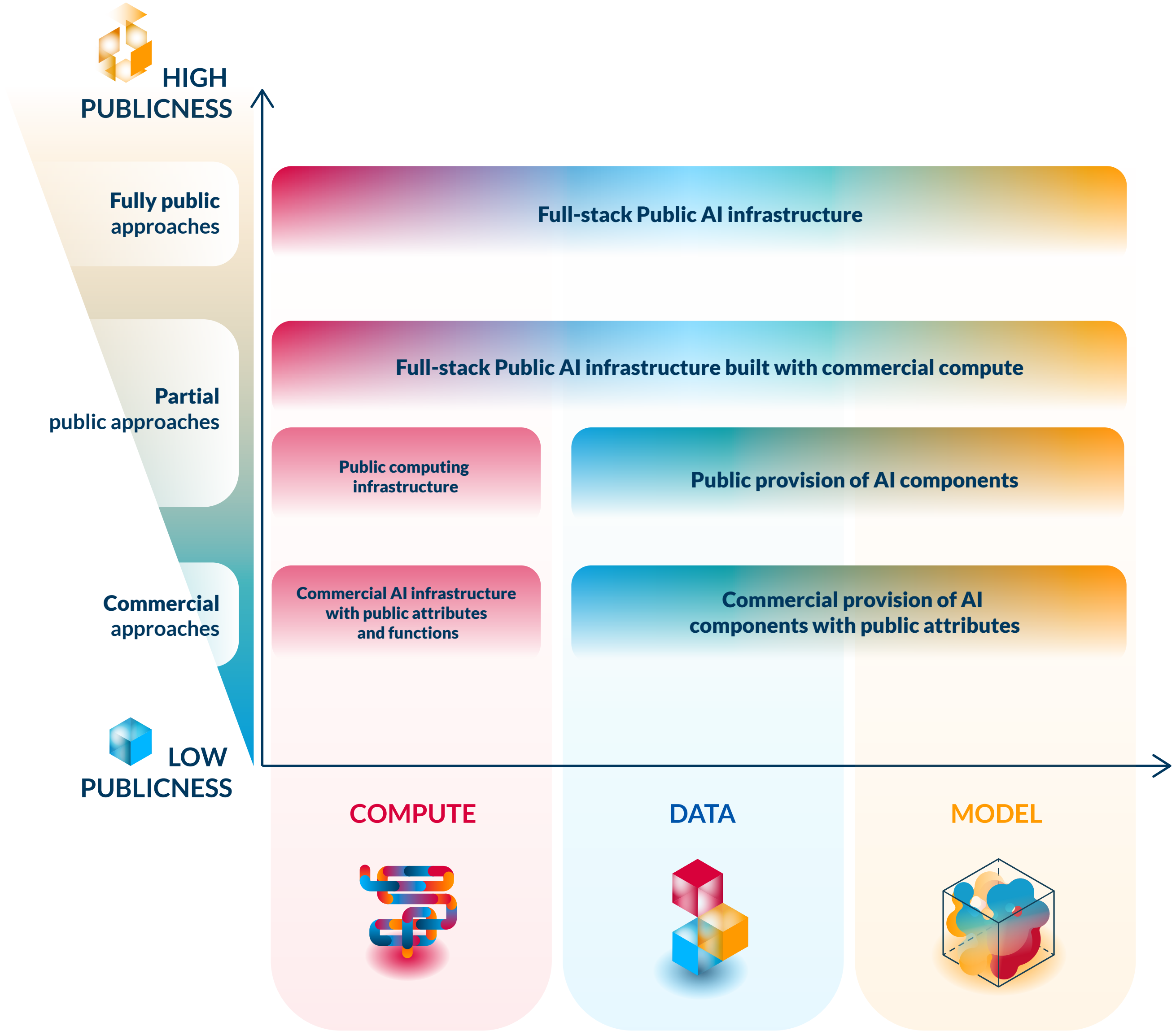

HIGH PUBLICNESS
Fully public approaches
Partial public approaches
Commercial approaches
LOW PUBLICNESS
Full-stack Public AI infrastructure
Full-stack Public AI infrastructure built with commercial compute
Public computing infrastructure
Public provision of AI components
Commercial AI infrastructure with public attributes and functions
Commercial provision of AI components with public attributes
COMPUTE
DATA
MODEL

Illustration by: Jakub Koźniewski

BertelsmannStiftung

Stronger forms require securing public control as well.

These differences can be systematically mapped onto the three foundational layers of the AI stack – compute, data and models. For each level, specific examples vary in the extent to which they meet the characteristics of public AI infrastructure.

Level 1: Commercial provision of AI components with public attributes. These are specific components – typically open models or libraries – developed and shared by commercial actors. They are publicly accessible (and often openly shared) by organizations that combine public interest goals with a commercial interest in sustaining an ecosystem around their solutions. These infrastructures typically exhibit high public attributes, low to moderate public function and low to moderate public control.

Example: PyTorch[146] is a deep learning framework that was open sourced by Meta and is currently hosted by the Linux Foundation. It is an openly shared and collectively governed AI component that plays a central role in AI development. Meta and other contributors have built an ecosystem of complementary research and innovation around PyTorch, with Meta maintaining a leading role.

Level 2: Commercial AI infrastructure with public attributes and functions. This refers to privately controlled infrastructure that ensures some level of access and has a public interest orientation. It includes mechanisms such as public access to commercial compute or platforms for sharing data and models. These infrastructures typically have high public attributes, moderate public function and low public control.

Example: Hugging Face[147] is a commercial entity with a mission of "democratizing good machine learning." It operates a model and dataset sharing platform that serve as the backbone of the open source AI ecosystem.

Level 3: Public computing infrastructure. These are public computing resources – such as supercomputers and data centers – funded entirely by the public sector or developed through public-private partnerships. These infrastructures typically have unclear public attributes (unless specific conditions for access are introduced), low to moderate public function and moderate to high public control.

Example: AI Factories[148] are public supercomputers in the EU that are financed through a mix of public and private funding. Their purpose is to support startups and research institutions, with generative AI development as one of their goals.[149]

Level 4: Public provision of AI components. These are individual components – such as datasets, benchmarks or evaluation tools – developed with public funding and/or hosted on public infrastructure. In some cases, such as datasets or software, there is no dependency on compute. These infrastructures typically have high public attributes, moderate to high public function and high public control.

Example: Common Voice[150] is a Mozilla initiative that provides an open platform for sharing voice data for AI training. It is widely cited as a best practice in responsible, open data sharing.

Level 5: Full-stack public AI infrastructure built with commercial compute. These are infrastructures that have public attributes and functions but depend on commercial compute both during development and deployment phases. These infrastructures typically have high public attributes, moderate to high public function, and low public control, at least with regard to computing power.

Example: OLMo[151] is an open model built by the Allen Institute for AI that sets a high bar for transparency in model, code and training data. The institute partnered with Google to train the model on its Augusta computing infrastructure and to deploy it on Vertex AI, Google's cloud platform.

Level 6: Full-stack public AI infrastructure. These infrastructures integrate data, models and compute resources that all meet the public AI standard. Systems built on such infrastructure benefit from synergies across layers and are free from commercial dependencies. These infrastructures typically have high public attributes, moderate to high public function and moderate to high public control.

Example: Alia[152] is a large language model developed by the Barcelona Supercomputing Center, a public

[146] PyTorch. https://pytorch.org/ Accessed April 27 April 2025.

[147] Hugging Face. https://huggingface.co/ Accessed 27 April 2025.

[148] European Commission. "AI Factories." Digital Strategy. https://digital-strategy.ec.europa.eu/en/policies/ai-factories Accessed 27 April 2025.

[149] AI Factories | Shaping Europe's digital future

[150] Common Voice Mozilla. https://commonvoice.mozilla.org/en Accessed 27 April 2025.

[151] OlMo Ai2. https://allenai.org/olmo Accessed 27 April 2025.

[152] Alia. https://www.alia.gob.es/eng/ Accessed 27 April 2025.

research institution in Spain, using its MareNostrum 5 supercomputer. The model sets a high standard for transparency and openness and addresses a linguistic gap by supporting Spanish and four co-official languages in Spain.

Several important points emerge from examining the gradient of publicness. First, most initiatives requiring computing power will remain dependent on commercial providers, except in rare cases where public institutions maintain their own supercomputers. Second, the provision of certain components, especially datasets and open source software for model development, involves fewer such dependencies, making it easier to build highly public AI infrastructure. Third, the ability to orchestrate a full-stack approach – integrating computing, data and software – can significantly enhance publicness by creating synergies across these layers. Full-stack initiatives are thus more critical to public AI strategies than isolated efforts focused on individual components. Conversely, public AI components tend to have greater impact when they are part of a coordinated, full-stack framework.

Goals and governance principles of public AI policies

The goal of public AI policy should not be to create competing infrastructures to commercial systems. Rather, public intervention should focus on supporting alternatives: offering new options, advancing new development paradigms and strengthening new capacities. In the next chapter, we outline governance principles that support a mission-oriented, public interest approach to AI. This also means that public policy should avoid participating in an AI race driven by commercial interests or fixating on speculative future needs.[153]

The issue of meaningful interventions at the compute layer illustrates this approach well. A report by the Ada Lovelace Institute on the role of public compute notes that "The aim of these policies should not simply be to build more and faster, but to challenge concentrated power and promote the creation of public value throughout the AI supply chain," and that "public compute investments should instead be seen as an industrial policy lever for fundamentally reshaping the dynamics of AI development and therefore the direction of travel of the entire sector."[154]

Even this framing of public policy goals might be too ambitious or unrealistic in the face of concentrated power in the AI ecosystem. Rather than aiming to fundamentally impact generative AI markets, policy should prioritize reducing dependencies and building independent capacity to generate public value. The key strategic challenge lies in identifying effective public interventions in a landscape where dominant AI firms are consolidating control over digital stacks and networks – and leveraging vast private funding to reinforce their position in the AI stack.[155]

The goals of public AI policy should be to:

- Develop AI infrastructure with components that are as public as possible by ensuring strong public attributes and functions
- Reduce dependency on dominant commercial providers of computing power and hardware through investments in public compute infrastructure, procurement policies with public interest conditions and other governance mechanisms
- Create incentives to develop AI infrastructure characterized by greater publicness through interventions that support public functions rather than relying solely on market-led development
- Identify, finance and support the development and maintenance of digital public goods critical to public AI development

[153] Zuzanna Warso, ibid.

[154] Eleanor Shearer, Matt Davies and Mathew Lawrence, ibid.

[155] Cecilia Rikap, "Antitrust Policy and Artificial Intelligence: Some Neglected Issues," Institute for New Economic Thinking, 10 June, 2024, https://www.ineteconomics.org/perspectives/blog/antitrust-policy-and-artificial-intelligence-some-neglected-issues

- Promote research and innovation that reduce reliance on proprietary AI components by advancing new paradigms for AI development

- Ensure adequate research talent and institutional capacity within the public sector to participate meaningfully in AI development

Governance of public AI

Governance of AI systems is a necessary component of any public AI agenda. By governance, we refer to the processes, structures and coordinated actions by multiple actors through which decisions related to AI are made and enforced. This concept extends beyond traditional legal frameworks to include various methods for setting and upholding norms – such as standards, codes of practice, voluntary licensing models or community-based rules.

It emphasizes the need for diverse stakeholders to work together to achieve desired outcomes while mitigating potential risks and harms. In the context of public AI, these governance mechanisms are intended to ensure that generative AI systems – and their underlying components – are either publicly owned or, at a minimum, subject to meaningful public oversight. This helps ensure that AI infrastructures meet the criteria of public attributes and public functions. Some of these mechanisms also address foundational conditions necessary for digital infrastructures to serve the public interest.

In what follows, we provide an overview of the core principles for governing public AI. Some principles focus on reinforcing the public character of AI solutions. Others are broader, representing good governance standards for AI systems more generally. To serve a public function, AI systems must meet high governance standards that guarantee accountability, transparency and sustainability.

To achieve these ends, public AI governance should be grounded in several high-level principles:

- **Directionality and purpose:** Public AI infrastructure should be built with clear intent and direction, ensuring alignment with public values and the principles outlined below. Public actors must orchestrate resources and stakeholders to ensure public AI initiatives generate public value and serve the common good.[156]

- **Commons-based governance:** Datasets, software, models and other key components of AI systems should be stewarded as commons. Such a framework encourages open access while establishing responsible use, democratic oversight and collective stewardship. Commons-based governance encompasses approaches that challenge proprietary ownership and promote shared control over resources.[157] It balances broad access to resources with governance mechanisms that protect rights, ensure quality and generate public value (including economic value).[158]

- **Open release of models and their components:** key components of AI models – including model weights, architectures and documentation – should be released under open source licenses and made universally accessible for use, study, modification and redistribution. Governance of open models should include strong transparency requirements and, where possible, openly shared training datasets.[159]

- **Open source software:** The advancement of AI has relied heavily on open source tools like scikit-learn and PyTorch, which are developed and

[156] Mariana Mazzucato, David Eaves and Beatriz Vasconcellos. "Digital public infrastructure and public value: What is 'public' about DPI?." UCL Institute for Innovation and Public Purpose. 2024. https://www.ucl.ac.uk/bartlett/publications/2024/mar/digital-public-infrastructure-and-public-value-what-public-about-dpi

[157] Alek Tarkowski and Jan Zygmuntowski. "Data Commons Primer." Open Future, 20 September 2022. https://openfuture.eu/publication/data-commons-primer

[158] Alek Tarkowski and Zuzanna Warso. "Commons-Based Data Set Governance for AI." Open Future, 21 March 2024. https://openfuture.eu/publication/commons-based-data-set-governance-for-ai

[159] Alek Tarkowski. "Data Governance in Open Source AI." Open Future, 24 January 2025. https://openfuture.eu/publication/data-governance-in-open-source-ai

maintained through community-led processes. Open source AI tools should be governed as digital public goods, in line with the Digital Public Goods Standard,[160] employing vendor-neutral and transparent community governance and development processes.

- **Conditional computing:** Public investments in the AI stack should, wherever possible, be tied to specific conditions and rules that shape how a technology is developed and used. This principle is especially relevant to public investment in computing hardware and infrastructure. Limited compute resources made available through public funding should be used efficiently and in service of public interest goals, such as openness.

- **Protecting digital rights:** Public AI systems should set a high standard for the protection of digital rights, including privacy and data protection, copyright, freedom of expression and access to information. This requires transparency, accountability, grievance mechanisms and robust frameworks for risk assessment and mitigation.

- **Sustainable AI development:** AI systems should be developed and deployed in ways that promote environmental sustainability, fair resource use and long-term societal benefit. Public procurement of AI infrastructure should include requirements for sustainable compute provision and the development of responsible supply chains.[161]

- **Reciprocity:** Public and private actors that benefit from public AI resources should ensure that downstream applications and derivative products adhere to these governance principles. This helps prevent the privatization of public value and protects against corporate capture.

When incorporating these governance principles, public AI policies must go beyond simply providing compute, data or model capacity for public interest use cases. They must also shape how AI infrastructures are developed and used.

[160] DPG Alliance. “DPG Standard.” GitHub. Accessed 27 April 2025. https://github.com/DPGAlliance/DPG-Standard

[161] Green Screen Coalition, et al. “Within Bounds: Limiting AI‘s environmental impact.” Green Screen Coalition. 5 February 2025. https://greenscreen.network/en/blog/within-bounds-limiting-ai-environmental-impact/

5 | AI strategy and three pathways to public AI

In this chapter, we outline key elements of a public AI strategy by presenting three pathways for developing public AI solutions, based on the three core layers of AI systems: data, compute and models.

The metaphor of a vertically integrated AI stack suggests that the hardware layers at the bottom form the infrastructural foundation of any AI system. We propose extending this concept to also treat datasets and models as forms of public infrastructure – critical building blocks upon which public AI solutions can be developed. In other words, there are three potential pathways to public AI, each focused on developing computing power, datasets or models as public infrastructural resources.

The aim of this chapter is to illustrate what a complete public AI strategy could look like, with interventions at the various layers of the AI stack.

Elements of a public AI strategy

As shown in the previous chapter, full-stack approaches to public AI development offer a higher degree of publicness than partial solutions. A public AI strategy should therefore orchestrate the provision of computing power and training data to support an ecosystem of public AI models. This ecosystem would include a state-of-the-art "capstone" model and a variety of smaller models. Supporting actions should include investment in research and innovation, development of a strong talent base and institutional capacity and programs that enable the deployment of public AI systems for specific solutions and uses.

In doing so, a public AI strategy should shift away from the current "AI race" among major commercial labs. In other words, it should treat AI technologies not as a potential superintelligence, but as a "normal technology."[162] This means pursuing a pragmatic development strategy in which investments in compute are tied to clearly defined goals for model development and deployment. Another pillar of this strategy should focus on fostering a more sustainable path for AI development – one centered on small models and innovations that make AI technologies more sustainable in terms of their energy consumption and environmental footprint.

In the following sections, we describe in more detail the elements of this strategy, divided into three pathways. Each pathway focuses on one layer of the AI stack: compute, data and models. In each case, the goal of public AI policy should be to secure public attributes, functions and control over AI systems and their components.

1. **Compute layer:** Public AI strategy at the compute layer should aim to reduce dependence on commercial computing power and, where necessary, develop publicly owned compute infrastructure. Supporting measures include research into more efficient AI development paradigms (within the model layer) and research to establish meaningful estimates of the computing needs of public AI initiatives.

[162] Arvind Narayanan and Sayash Kapoor. "AI as Normal Technology." Knight First Amendment Institute. 2025. https://kfai-documents.s3.amazonaws.com/documents/2b27e794d6/AI-as-Normal-Technology---Narayanan---Kapoor.pdf Accessed 3 April 2025.

Figure 6 | Elements of a public AI strategy

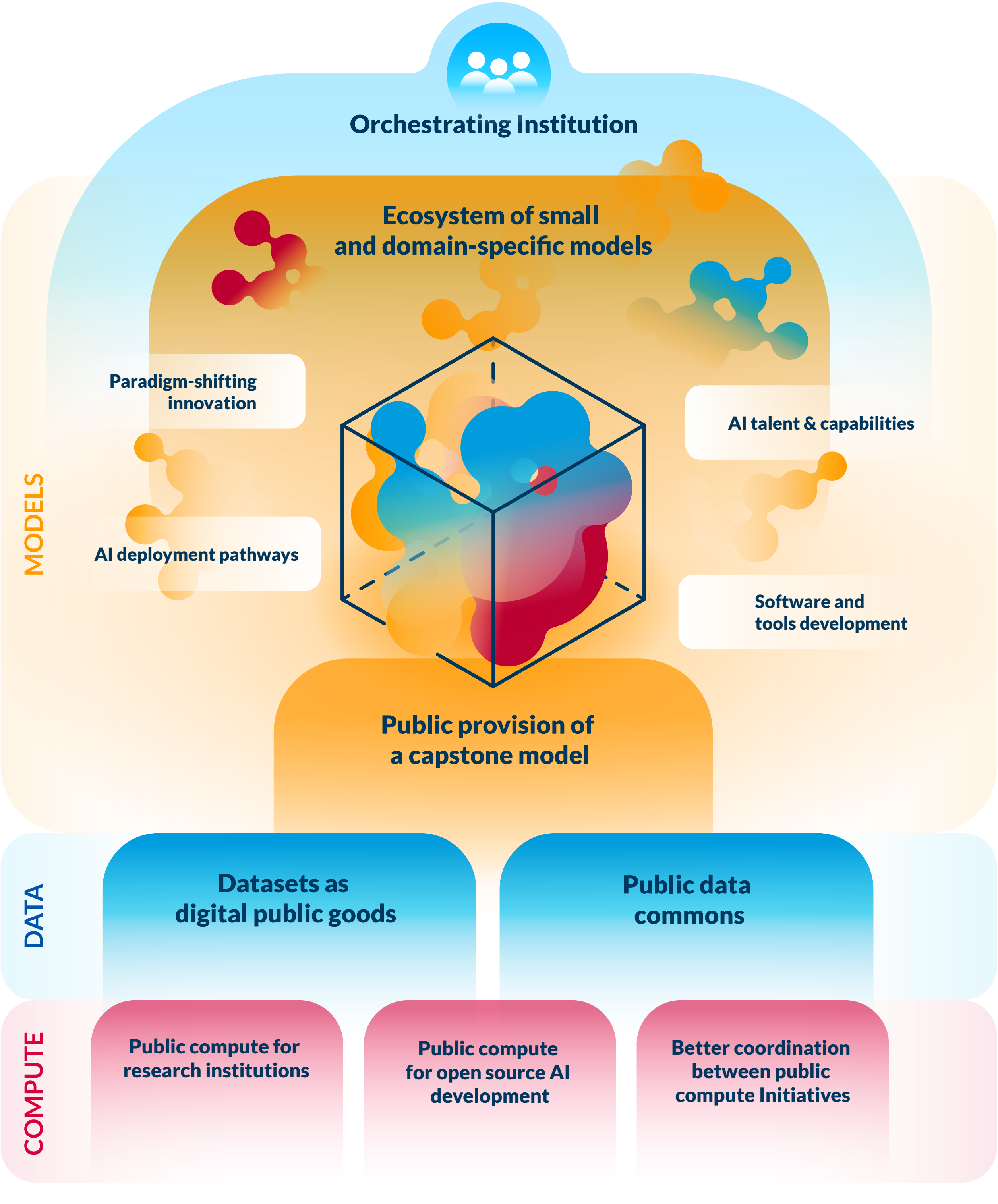

Orchestrating Institution
Ecosystem of small and domain-specific models
Paradigm-shifting innovation
AI deployment pathways
AI talent & capabilities
Software and tools development
MODELS
Public provision of a capstone model
DATA
Datasets as digital public goods
Public data commons
COMPUTE
Public compute for research institutions
Public compute for open source AI development
Better coordination between public compute Initiatives

Illustration by: Jakub Koźniewski

2. **Data layer:** At the data layer, public AI strategy should promote the development of high-quality datasets that are both publicly accessible and governed through democratic, commons-based frameworks. These governance models should ensure public attributes and functions while protecting data from inappropriate value extraction, such as free-riding.

3. **Model layer:** At the model layer, public AI strategy should build on open source generative AI ecosystems that demonstrate how technologies can be developed with strong public attributes. It should also aim to strengthen the public functions of these technologies by setting a clear development agenda focused not on technology for its own sake, but on generating public value – through targeted application development and demand creation.

Taken together, this public AI strategy is relatively complex and requires new forms of governance capable of coordinating the actions of diverse actors to achieve policy goals. It calls for a strong institutional framework able to lead and orchestrate the strategy effectively.

The public AI ecosystem and its orchestrating institution

A public AI strategy should aim to create an ecosystem of public generative AI infrastructures and systems, rather than focus solely on individual initiatives or standalone institutional capacity. Calls for centralized public AI development often overlook the distributed nature of modern AI research, which is supported by open source development norms. Even when investing in centralized capacity – such as public compute resources – a public AI strategy must also support a broader ecosystem centered on public functions and the creation of public value.

The governance mechanisms proposed in chapter 4 support this ecosystem-based approach by promoting the sharing of data, software and knowledge, and by ensuring interoperability among solutions within the ecosystem.

Helping this ecosystem thrive requires leveraging the power of public institutions, including market-shaping tools like industrial policy and public procurement, to support and strengthen it. Public institutions must also be part of this ecosystem, and public AI solutions should be built on infrastructures developed within it.[163]

A key role in this ecosystem in this ecosystem should be played by a public institution capable of orchestrating actions across a decentralized network of actors. Orchestration involves managing how public digital infrastructure is produced and how it aligns with evolving needs and values over time. It also includes the ability to adapt the ecosystem as those needs, values and conditions change.[164] In other words, proper orchestration allows an institution to control the generative capability of various infrastructures in the public AI ecosystem.

An orchestrating institution – or, more likely, a network of institutions – thus plays a critical role in the ecosystem. Several blueprints for such institutions have been proposed, drawing inspiration from organizations that have successfully delivered other public goods and coordinated complex ecosystems.

Brandon Jackson, writing for Chatham House, proposes a “BBC for AI” or a British AI Corporation (BAIC), which he describes as “a new institution that would ensure that everyone has access to powerful, responsibly built AI capabilities. Yet the BAIC should be more than just a head-to-head competitor with the private AI companies. It should be set up with an institutional design that empowers it to chart an

[163] Katja Bego. “Towards Public Digital Infrastructure: A Proposed Governance Model.” Nesta, 30 March 2022. https://www.nesta.org.uk/project-updates/towards-public-digital-infrastructure-a-proposed-governance-model/; Alek Tarkowski, et al. “Generative Interoperability.” Open Future, 11 March 2022. https://openfuture.eu/publication/generative-interoperability

[164] Antonio Cordella and Andrea Paletti. “Government as a platform, orchestration, and public value creation: the Italian case.” Government Information Quarterly, 36 (4). ISSN 0740-624X. https://doi.org/10.1016/j.giq.2019.101409

independent path, building innovative digital infrastructure in the public interest."[165]

Another proposal, from the Center for Future Generations, outlines a "CERN for AI"[166] – a more centralized model for an institution combining in-house research with a broad mandate to collaborate with academia and industry. A separate proposal by Daniel Crespo and Mateo Valero suggests that European public AI efforts should draw inspiration from Airbus and Galileo, two examples of successful coordination among industry actors to develop innovative, competitive technologies.[167] Most recently, the CurrentAI partnership, launched at the Paris AI Action Summit in February 2025, aims to orchestrate public interest AI development at the global level by coordinating the efforts of governments, philanthropies and private companies.[168]

It should be emphasized that the aim of this publication is not to prescribe any specific institutional model. Rather, it seeks to underscore the importance of thoughtful institutional design as the foundation for any effective public AI strategy.

Three pathways toward public AI infrastructure: compute, data and model

Each of these pathways offers a different approach to meeting public AI objectives and addressing dependencies on market monopolies. For each pathway, we begin with an overview of bottlenecks and opportunities, followed by a list of proposed solutions. At the end, we offer three additional recommendations for supportive measures.

The goal of these sections is not to provide detailed blueprints for every solution. Rather, we aim to outline a comprehensive strategy, coordinated across the three layers and pathways to public AI. These recommendations do not take into account the economic dimensions of deploying public AI infrastructure, which must be shaped by the specific local or regional context where such policies are developed. Nor do we attempt to provide a comprehensive list of public AI efforts; examples are included only to illustrate different solutions.

Compute pathway to public AI

Computing infrastructure forms a foundation for public AI development, yet policy approaches must strike a balance between addressing real infrastructural needs and avoiding inflated investment claims. While compute is essential to AI progress, public initiatives should prioritize targeted, strategic investments rather than attempting to replicate the massive spending patterns of dominant commercial players.

Compute: bottlenecks

Compute is an integral component of AI development and deployment, yet its landscape is defined by an extraordinarily complex supply chain and market dynamics dominated by a few powerful players. As outlined in the previous chapter, compute can be broken down into three critical components: advanced chips, software frameworks to run those chips and data centers.

At the hardware level, semiconductor production is widely considered the most complex and globally distributed supply chain in the world. It is typically divided into three main stages: chip design, front-

[165] Brandon Jackson. "The UK needs a 'British AI Corporation', modelled on the BBC." Chatham House. 10 June 2024. https://www.chathamhouse.org/2024/06/artificial-intelligence-and-challenge-global-governance/07-uk-needs-british-ai-corporation

[166] Alex Petropoulos, et al. "Building CERN for AI." Center for Future Generations. 30 January 2025. https://cfg.eu/building-cern-for-ai/

[167] Daniel Crespo and Mateo Valero. "Es la hora de 'AIbus': por qué Europa debe crear una gran empresa de AI." El Pais. 11 April 2024. https://elpais.com/tecnologia/2024-04-11/es-la-hora-de-aibus-por-que-europa-debe-crear-una-gran-empresa-de-ai.html

[168] "Nouveau partenariat pour promouvoir l'IA d'intérêt général." Élysée. 11 February 2025. https://www.elysee.fr/emmanuel-macron/2025/02/11/nouveau-partenariat-pour-promouvoir-lia-dinteret-general

end manufacturing and back-end manufacturing.[169] Chip design involves creating the architecture and layout of the semiconductor, often using specialized software. Front-end manufacturing refers to the fabrication process in advanced foundries, where billions of transistors are etched onto each chip using photolithography and other precision techniques. Back-end manufacturing involves the assembly, packaging and testing of chips before integration into devices.

Each stage of this supply chain is highly specialized and concentrated in the hands of a few firms. For instance, ASML in the Netherlands is the only company in the world capable of producing extreme ultraviolet (EUV) lithography machines – critical equipment for manufacturing advanced chips. These machines consist of over 100,000 components and cost roughly €350 million each.

Front-end manufacturing is particularly capital- and technology-intensive, leading to extreme market concentration. Only a few firms – most notably Samsung and Taiwan Semiconductor Manufacturing Company (TSMC) – can operate in this space. In 2024, TSMC held a 90% market share in advanced logic chips, which are essential for AI training and deployment.[170] That same year, TSMC invested more than $30 billion in capital expenditures and manufactured most of Nvidia's chips, along with chips for numerous Chinese companies, despite geopolitical tensions.[171]

Specialized software frameworks also create significant lock-in. Nvidia's proprietary CUDA platform, designed to run exclusively on its GPUs, dominates the market and has created a ‚walled garden' ecosystem. The deep integration of CUDA with Nvidia hardware, its robust ecosystem of machine learning libraries and frameworks and its superior multi-GPU scaling capabilities have entrenched the company's position and raised substantial barriers for competitors.

Data centers – the final layer – integrate chips and software into usable compute systems. These facilities are also marked by high concentration due to their massive cost and operational complexity. In January 2025, OpenAI, Oracle and SoftBank announced the Stargate project, with a planned investment of up to $500 billion in data center infrastructure. While the feasibility of this investment remains uncertain, the scale illustrates the intensifying global race for compute leadership.[172] In 2024, companies such as Microsoft, Meta, Amazon and Apple spent approximately $218 billion on physical infrastructure. As a result, state-of-the-art data centers remain the domain of large, well-capitalized corporations or publicly backed initiatives.[173]

In short, these three components – chips, software and data centers – highlight deep dependencies on a small number of dominant players and reflect the extraordinary capital intensity of the compute market.

Public initiatives aimed at securing compute capacity often focus narrowly on sovereignty rather than public value. These efforts treat compute as a national resource and view expanded capacity as an end in itself. Investments, typically undertaken with commercial partners, are rarely tied to public value conditions and often serve to bolster national commercial actors. The notion of sovereign AI, when built on commercial compute infrastructure, risks reinforcing the market dominance of existing players.

[169] Jan-Peter Kleinhans and Julia Christina Hess. "Governments' role in the global semiconductor value chain #2." Stiftung Neue Verantwortung. 6 July 2022. https://www.interface-eu.org/publications/eca-mapping

[170] "TSMC's Advanced Processes Remain Resilient Amid Challenges." Trendforce. 8 April 2024. https://www.trendforce.com/news/2024/04/08/news-tsmcs-advanced-processes-remain-resilient-amid-challenges/

[171] Chris Miller. "Chip War. The Fight for the World's Most Critical Technology." Simon & Schuster. 4 October 2022. https://www.simonandschuster.com/books/Chip-War/Chris-Miller/9781982172008

[172] Steve Holland. "Trump announces private-sector $500 billion investment in AI infrastructure." Reuters. 22 January 2025. https://www.reuters.com/technology/artificial-intelligence/trump-announce-private-sector-ai-infrastructure-investment-cbs-reports-2025-01-21/

[173] Michael Flaherty. "Tech dollars flood into AI data centers." Axios. 26 December 2024. https://www.axios.com/2024/12/20/big-tech-capex-ai

While there is growing consensus among policymakers that public compute provision is essential, the high costs and fast pace of technological change make such interventions challenging. Unlike sovereign AI strategies, the goal should not merely be to expand national capacity, but to ensure that these investments generate public benefit.

Compute: opportunities

The high market concentration and capital intensity of computing infrastructure create significant challenges for developing public compute initiatives. As noted in the Computing Commons report by the Ada Lovelace Institute, “policymakers need to be realistic about what can be achieved through public compute projects alone,” as dependencies in semiconductor supply chains mean that “for most jurisdictions the goal of ‘onshoring’ production will likely be a near impossibility in the short to medium term.”

The report also highlights “a lack of genuinely independent alternatives to AI infrastructures operated by the largest technology firms, which are overwhelmingly headquartered in the USA and China.”[174] At the same time, compute costs continue to rise sharply, with spending on AI training runs increasing by a factor of 2.5 annually in recent years.[175] Even as efficiency and optimization improve, total expenditures remain massive, with some leading AI companies announcing plans to invest hundreds of billions of dollars in the coming years.

Due to the high capital requirements and rapid pace of technological change, it is more realistic to pursue public-private partnerships, where governments serve as one partner rather than the sole provider. Aurora GPT – an initiative to build a science-focused foundation model in the United States – and the European AI Factories initiative are both examples of such partnerships, resulting in infrastructures with a high degree of publicness.

In addition, public compute initiatives face two major shortcomings. First, they often lack effective allocation mechanisms and operate without a clear vision for which AI components and infrastructures should be prioritized. Not all projects require large-scale compute, and a better understanding of actual computing needs is necessary to inform a sound public compute strategy. Without clear criteria and governance, resources risk being misallocated or underused. Any large-scale initiative – such as a potential “CERN for AI” – should begin with a systematic assessment of demand, identifying which institutions, researchers and projects require which levels of compute. This demand-driven approach should guide allocation policies, access rules and future scaling. A mission-driven strategy should also link compute use to clear public goals.

Second, compute initiatives are often framed primarily as sovereign assets, with a focus on national control and expanding capacity as ends in themselves. This perspective risks prioritizing geopolitical narratives about AI sovereignty over more meaningful questions about who has access to computing power and for what purpose.

Given these structural dependencies, most compute-based pathways toward public AI will likely fall in the middle of the publicness gradient.

Policy recommendations for the compute pathway include:

Public compute for open source AI development

The first recommendation is to ensure that fully open source AI projects – with open, accessible training data – have access to sufficient compute resources. A core interest of any public AI agenda should be to guarantee that at least one open model exists with capabilities comparable to the state-of-the-art, alongside the institutional capacity to develop and work with such models (this is discussed further in the model pathway section). Public supercomputers

[174] Matt Davies and Jai Vipra. “Computing Commons.” Ada Lovelace Institute. 7 February 2025. https://www.adalovelaceinstitute.org/report/computing-commons/

[175] Be Cottier et al. “How Much Does It Cost to Train Frontier AI Models?” Epoch. 25 January 2024. https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models

and publicly supported data center initiatives play an essential role in making this possible.

A notable example is the collaboration between the Allen Institute for AI and the LUMI supercomputer in – part of the EU's EuroHPC AI Factories initiative. In 2024, the Allen Institute released OLMo, a fully open source language model, including its pre-training dataset, trained using LUMI's computing power.[176] The EuroHPC initiative represents a form of semi-public infrastructure, funded through a combination of EU funds, national budgets and private contributions.[177]

Another example is the Aurora GPT, a U.S. project led by Argonne National Lab in partnership with Intel, which relies on a public-private collaboration using the Aurora supercomputer and Intel-provided GPUs to develop a foundation model for science.[178] In Europe, the Barcelona Supercomputing Center recently released Alia, a Spanish-language model described as open, transparent and public.[179]

These examples demonstrate the potential of public compute to support open source AI efforts. However, sustaining progress at the frontier – and enabling wide deployment – will require continuous expansion and upgrades to public computing capacity. As proprietary models continue to advance rapidly, there is a growing risk that open models will fall too far behind. To avoid this widening gap, expanded access to public compute for open source AI development is essential.

Public compute for research institutions

A second strategic pillar of a public AI strategy is expanding access to compute for academic institutions and public research organizations. Many universities face serious constraints due to the high cost and limited availability of GPUs, often relying on expensive commercial cloud services. Investing in public supercomputing and data center initiatives is essential to ensure that cutting-edge AI research does not remain confined to closed, private labs but can also take place in universities and research institutes.

Providing compute access to the academic sector is also a way to attract and retain top AI talent within public research institutions. Without adequate resources, researchers and students may be drawn to well-funded corporate labs, limiting the development of fully open source models and public interest applications.

Improved coordination between public compute initiatives

Many existing public compute initiatives tend to frame infrastructure primarily as a sovereign asset, emphasizing national control and domestic capacity. However, this sovereignty-first approach risks fragmenting the broader mission of public AI. A truly public AI agenda should not reinforce narrow national strategies but instead promote cross-border cooperation – particularly among democratic nations – to maximize the impact of public investment. Without coordinated strategies, governments risk duplicating efforts, underutilizing resources and falling short of the scale needed to support open source AI ecosystems and create viable public alternatives to proprietary models.

One area where improved coordination is urgent is the AI Factories initiative in Europe, which established data centers in strategic locations, such as large supercomputing hubs. This initiative would benefit from deeper collaboration across national borders to reach its full potential. One proposal that embodies this approach is the "Airbus for AI"[180] model, which advocates pooling resources and building shared capacity to produce high-performing, open AI models. Supporting such collaborative

[176] Allen Institute for AI, "Hello OLMo: A Truly Open LLM." Allen Institute for AI Blog. January 9 January 2024. https://allenai.org/blog/hello-olmo-a-truly-open-llm-43f7e7359222

[177] European High-Performance Computing Joint Undertaking (EuroHPC JU). "Discover EuroHPC JU." https://eurohpc-ju.europa.eu/about/discover-eurohpc-ju_en Accessed 27 April 2025.

[178] Rusty Flint. "AuroraGPT: Argonne Lab and Intel." Quantum Zeitgeist. 14 March 2024. https://quantumzeitgeist.com/auroragpt-argonne-lab-and-intel/

[179] Alia. https://www.alia.gob.es/eng/ Accessed 27 April 2025.

[180] Daniel Crespo and Mateo Valero. ibid.

frameworks is essential to transform scattered infrastructure into a cohesive, strategic and effective public AI ecosystem.

Data pathway to public AI

Current approaches to data sources for AI development oscillate between proprietary control and unrestrained extraction from public sources. Unlike the compute pathway, establishing a data commons is not primarily about investments in technology or hardware – it requires better governance of various types of data. The data pathway offers opportunities to create genuine digital public goods with governance mechanisms that protect against value extraction and ensure equitable access. This requires overcoming bottlenecks related to proprietary control, declining consent for data use and insufficient attention to data quality.

Data: bottlenecks

While much of AI development is fueled by a culture of open sharing, data practices are largely shaped by either proprietary control or unregulated use of publicly available sources. Few AI development teams make meaningful efforts to share high-quality, useful datasets. At the same time, they seek competitive advantages through proprietary sources – such as user-generated and personal data from social networks and online platforms, or data obtained through exclusive agreements. Public web content continues to be crawled and scraped, with attempts to filter and improve quality. Yet there are signs that the social contract underpinning the open web is eroding, as content owners increasingly withdraw consent for their domains to be crawled. These trends contribute to a negative feedback loop, leading to what Stefaan Verhulst has described as a “data winter” – a decline in the willingness to see data as a resource that can serve the common good.[181]

Governance and ethical concerns related to data use for AI training represent another major bottleneck. The training of commercial models on the entirety of the public internet – often under unclear legal conditions and with minimal transparency – reflects a lack of proper oversight and results in the extraction of value from global knowledge and cultural commons.[182] Evidence gathered by the Data Provenance Initiative shows that consent for web crawling is steadily decreasing, especially among domains whose content is used in AI training.[183] And recent data from the Wikimedia Foundation shows that rapidly growing automated traffic from web crawlers of AI companies is becoming a financial burden.[184]

Data quality presents a further challenge. Unlike compute-related dependencies, poor data quality may not halt AI development but does undermine the usefulness and integrity of generative AI solutions. Training models on publicly available content often reflects not only poor governance but also a lack of attention to data quality. Although this issue has been raised by some experts,[185] recurring examples demonstrate that governance practices remain insufficient.

As a result, the widespread use of large datasets built from scraped web content risks reinforcing biases – such as the overrepresentation of well-resourced written languages and dominant cultural narratives – exacerbating existing inequalities. In some cases, it has also led to the scaling of harmful content, including explicit imagery.[186] On a global scale,

[181] Stefaan Verhulst. “Are We Entering a Data Winter?” Policy Labs. 21 March 2024. https://policylabs.frontiersin.org/content/commentary-are-we-entering-a-data-winter

[182] Paul Keller. “AI, the Commons and the Limits of Copyright.” Open Future. 7 March 2024. https://openfuture.eu/blog/ai-the-commons-and-the-limits-of-copyright/

[183] Shayne Longpre et al. “Compulsory Licensing for Artificial Intelligence Models,” arXiv preprint. 24 July 2024. https://arxiv.org/abs/2407.14933

[184] Birgit Mueller et al. “How Crawlers Impact the Operations of the Wikimedia Projects,” Diff – Wikimedia Foundation Blog, 1 April 1 2025. https://diff.wikimedia.org/2025/04/01/how-crawlers-impact-the-operations-of-the-wikimedia-projects/

[185] Will Orr and Kate Crawford. “Is AI Computation a Public Good?” SocArXiv preprint. 2024. https://osf.io/preprints/socarxiv/8c9uh_v1

[186] Abeba Birhane et al. “Multimodal datasets: misogyny, pornography, and malignant stereotypes.” arXiv preprint. 5 October 2021. https://arxiv.org/pdf/2110.01963 Accessed 3 April 2025.

inadequate data governance is increasingly seen as enabling new forms of data colonialism and extractive practices.[187]

Data: opportunities

The data pathway to public AI development aims to create a pool of datasets and content collections that function as digital public goods. While data is not typically seen as public infrastructure for AI development, it can in fact possess public attributes, serve public functions and be subject to public control.[188] For this to occur, there must be a dual obligation: to expand access and to better protect various data sources. In the case of data, the key challenges are less about upstream dependencies and more about downstream risks – specifically, the risk that public data is extracted for private gain, reinforcing inequality.[189] This happens when private actors capture the economic value generated by data without giving back to the people and institutions that created or maintained it as a public good.

This means data-sharing efforts must shift focus – from simply increasing the volume of available data to improving its quality and implementing governance mechanisms that ensure equitable, sustainable sharing protected from value extraction. As a result, both data transparency and novel gated access models (to protect, for example, personal data rights) are becoming central governance issues. In addition, ensuring access to commercial datasets – at least for research purposes – is a vital reciprocal measure to secure private data for public interest uses.

Model developers have always relied on high-quality open datasets. Wikipedia is a prime example of a structured, high-quality dataset, and books remain a foundational data source[190] – even when their legal status is ambiguous, as seen with the Books3 dataset. Simply increasing the volume of training data is neither the only strategy nor the most effective one. Developers are already focused on improving the quality of web-scraped data, as shown by projects like the FineWeb datasets, which filter and clean Common Crawl data.

High-quality data sources can support generative AI development at various stages: pretraining, post-training or adaptation, inference and the creation of synthetic data.[191] Newer approaches to dataset development rely less on pretraining and focus more on the post-training phase, which requires data that cannot be easily “found in the wild” or scraped from public sources. This includes domain-specific data for fine-tuning specialized models, as well as dialogues and task-specific examples used in instruction tuning. In model distillation, for example, developers rely not on new human-generated data but on synthetic data generated by a “teacher” model. Public interventions must therefore address both the governance of publicly available data and the development of specialized datasets and tools – covered further in the section on additional measures, as part of the software and tool ecosystem for public AI.

Data is not a homogenous concept, and its use is governed by multiple legal frameworks that protect data rights, including copyright and personal data regulations. As such, data sharing exists on a spectrum – from fully open to gated models – each of which can be understood as a form of commons-based governance. These approaches are underpinned by key principles: sharing as much data as possible while maintaining necessary restrictions; ensuring transparency about data sources; respecting data subjects’ choices; protecting shared resources; maintaining

[187] James Muldoon and Boxi A. Wu. “Artificial Intelligence in the Colonial Matrix of Power.” Philosophy & Technology 36, no. 80 (2023). https://doi.org/10.1007/s13347-023-00687-8 Accessed 3 April 2025.

[188] Digital Public Goods Alliance, et al. “Exploring Data as and in Service of the Public Good.” https://www.digitalpublicgoods.net/PublicGoodDataReport.pdf Accessed 23 April 2025.

[189] Paul Keller and Alek Tarkowski. “The Paradox of Open.” Open Future. https://paradox.openfuture.eu/ Accessed 23 April 2025.

[190] Alek Tarkowski et al. “Towards a Books Data Commons for AI Training.”Open Future. 8 April 2024. https://openfuture.eu/publication/towards-a-books-data-commons-for-ai-training/

[191] Hannah Chafetz, Sampriti Saxena and Stefaan G. Verhulst. “A Fourth Wave of Open Data? Exploring the Spectrum of Scenarios for Open Data and Generative AI.” The GovLab. May 2024. https://www.genai.opendatapolicylab.org/

dataset quality; and establishing trusted institutions to steward them.

Policy recommendations for the data pathway include:

Datasets as digital public goods

Open data – such as Wikimedia content – has been a foundational resource in the development of generative AI models. Using openly licensed data and content for AI training offers the advantage of legal certainty when developing models. Research by GovLab suggests that "the intersection of open data–specifically open data from government or research institutions –and generative AI can not only improve the quality of the generative AI output but also help expand generative AI use cases and democratize open data access."[192] Open data therefore has the potential to support public functions of generative AI in solutions that address global challenges like climate change or healthcare.[193]

Using open datasets for AI training also offers the possibility of moving beyond current norms of model openness, in which model weights are shared but training data remains closed and nontransparent. Open data enables the development of fully open AI models.[194] Ongoing efforts to build such models are undertaken by organizations like EleutherAI, Spawning and the Allen Institute for AI.

Public AI policies can build on more than a decade of experience developing open data infrastructure. However, the approach must shift from simply releasing as much data as possible to intentionally creating high-quality, purpose-built datasets for AI training. Beyond just training foundation models, many initiatives already provide open data for computational research and demonstrate the value of open access. Notable examples include the Human Genome Project, CERN's Open Data portal and NASA's Earthdata platform.

Public data commons

A public data commons is a data governance framework that aims to secure public interest goals through commons-based management of data.[195] These commons complement open data approaches and are particularly well suited for cases involving sensitive data, where rights must be protected, or where economic factors tied to dataset creation and maintenance must be considered.

Public data commons should be governed by three core principles:

- Stewarding access through clear sharing frameworks and permission interfaces
- Ensuring collective governance through defined communities, trusted institutions and democratic control
- Generating public value through mission-oriented goals and public interest-oriented licensing models.

To establish data commons for AI training, edicated public institutions are needed to act as trusted intermediaries. These institutions must also possess the technical capabilities to build hosting platforms for modern training datasets. Public data commons serve a gatekeeping role, supporting various data types and implementing flexible governance – from open access to gated models that preserve individual and collective data rights. Work on data commons is often motivated by the need to protect community-owned data from exploitation. Notable examples include the Māori language datasets and AI tools developed by Te Hiku Media, as well as African language datasets curated by Common Voice and the African Languages Project.

[192] Digital Public Goods Alliance, ibid.

[193] Hannah Chafetz, Sampriti Saxena and Stefaan G. Verhulst. ibid.

[194] Stefan Baack, et al. "Towards Best Practices for Open Datasets for LLM Training Proceedings from the Dataset Convening." Mozilla. 13 January 2025. https://foundation.mozilla.org/en/research/library/towards-best-practices-for-open-datasets-for-llm-training/

[195] Alek Tarkowski and Zuzanna Warso. "Commons-based data set governance for AI." Open Future. 21 March 2024. https://openfuture.eu/publication/commons-based-data-set-governance-for-ai/ Accessed 3 April 2025.

Model pathway to public AI

Bottlenecks related to model development largely stem from previously described challenges at the compute and data layers. Under the transformer paradigm, limited computing power and restricted access to training data constrain the ability to build capable models. At the same time, this layer is characterized by strong norms of open sharing – of models and of related tools and components. However, bottlenecks within this layer primarily relate to the lack of highly capable open models that could serve as a "capstone" for a broader ecosystem of open model development.

In contrast, there are many smaller models that meet high standards of public control and public attributes, yet they often remain primarily research or engineering artifacts. Building state-of-the-art open models – such as in the case of DeepSeek – still requires significant compute resources, which remain largely accessible only to well-funded commercial actors. While notable exceptions exist – such as the fully open source OLMo models released by the Allen Institute for AI – these remain rare. As a result, there are even fewer downstream initiatives focused on developing tailored, public interest applications built on top of open models.

Models: bottlenecks

Public AI policy goals at the model layer should center on ensuring the active development of open AI models that can be deployed for public interest uses. Under the dominant transformer paradigm, model development faces serious constraints related to compute and to the lack of sufficient, high-quality training data.

Building foundation models from scratch is not only extremely expensive, but also subject to rapid obsolescence as state-of-the-art capabilities are advancing quickly. This makes it difficult to justify large-scale public investments aimed at directly competing with commercial AI labs. Even when public compute resources are made available – as in the French government's support of the BLOOM model through the Jean-Z supercomputer – deployment remains a major hurdle.

At the model layer, openness and access vary widely. There is a gradient of release strategies, ranging from fully open source models to API-based access to fully closed models.[196] The dominant AI labs that build state-of-the-art models adopt various strategies, and although some have moved toward greater openness in the last year, none have released models that meet the definition of open source AI. At best, some share open weights but fail to disclose training data or provide transparency around training processes. Among the more open actors, DeepSeek is the only lab that consistently releases open weights for all models. Mistral and Alibaba Labs release some models in this way, and Meta has shared several models under restrictive licenses. Other companies have released only specific models – typically smaller or task-specific ones – such as Google's Gemma and BERT, OpenAI's Whisper or Microsoft's Phi.

As explained previously, model development – especially among labs with limited access to compute – often involves creating derivatives from openly shared models or architectures. This leads to dependencies on major commercial players such as Meta, Mistral, Alibaba Labs or DeepSeek. In each case, incomplete transparency around training data and methods hinders further research and development.

Over the last two years, multiple open small models have been released, often developed to address linguistic or regional gaps left by major AI labs. Examples include the SEA-LION models created by AI Singapore; Aya, a family of multilingual models from Cohere; and Bielik, a Polish LLM built by the grassroots Spichlerz initiative. However, these developers typically lack the resources needed to sustain long-term development or deployment of their models, limiting the impact of these alternatives.

While models themselves are a key component facilitating further AI development, open access to ad-

[196] Irene Solaiman. ibid.

ditional AI model components is just as important. As noted in the Model Openness Framework[197] and the Framework for Openness in Foundation Models by the Columbia Convening on Openness and AI,[198] openness in AI goes beyond the architecture and parameters of individual models. It also includes the code and datasets used to train, finetune or evaluate models.[199, 200] Their development faces similar constraints, typical for many open source projects, also beyond AI development.

Models: opportunities

These constraints suggest that public AI initiatives need a different approach than competing directly with commercial AI development. Public AI strategy should not aim to engage in an expensive race to build the largest and most capable models – a strategy that is neither realistic nor sustainable without major increases in public compute capacity. Instead, the focus should be on fostering an ecosystem of competitive open AI models and the components needed to build them.

The rapid pace of technological development means that the resources required to develop or deploy models can shift quickly. With new AI paradigms, the demands for successful model development or deployment may change at any moment. For instance, recent advances in model distillation have made it possible to create small, efficient models at relatively low cost – models that can effectively compete with earlier generations of large models.

Model-based pathways to public AI must be grounded in a clear understanding of the technological advances that enable more affordable yet capable models, along with targeted investments that support the entire open source AI ecosystem and public interest innovation. These pathways should support both the development of a state-of-the-art “capstone model” and the creation of derivative small models that are more sustainable and suited to specific needs.

Policy recommendations for the model pathway include:

Provision of a capstone model

Most so-called open models today fall short of genuine open source standards – often omitting training data, critical documentation or transparency around the training process. This undermines scientific reproducibility and makes it difficult to audit or assess model bias. As a result, critical infrastructure is increasingly shaped by private actors without democratic oversight. While this is particularly true at the frontier, not all cutting-edge models are developed by private firms – for example, the Allen Institute for AI’s OLMo 2 demonstrates that open and transparent alternatives are possible.

A robust public AI strategy should foster the development and long-term sustainability of high-performance, openly available models, ensuring a rich ecosystem of public alternatives to proprietary systems. Among these, governments should prioritize the creation of at least one “capstone model” – a permanently open, democratically governed model that aspires to remain at or near the frontier of AI capabilities. This model would serve not only as a flagship public asset but also as a foundation for broad-based research, innovation and deployment. While multiple such models may emerge, the capstone model would serve as a strategic anchor. Because open source is a global endeavor, adherence to open source standards should take precedence over the model’s country of origin.

However, due to the speed of innovation in AI, building a state-of-the-art model carries the risk of rapid obsolescence, as new breakthroughs by leading commercial labs may quickly outpace public efforts. As discussed below, complementary options include investments in small and domain-specific models as well as in capabilities (i.e., human capital) in AI research and development.

[197] Matt White, et al. ibid.

[198] Adrien Basdevant, et al. ibid.

[199] Matt White, et al. ibid.

[200] Adrien Basdevant, et al. ibid.

To function as a sustainable and reliable foundation for a broader ecosystem, the capstone model must be provisioned as a permanent public good. A public AI strategy must secure funding not only for model development but also for long-term deployment, particularly to cover inference compute costs.

Development of small and domain-specific models

Limited public compute resources can be strategically directed toward targeted interventions that deliver public value or shape market dynamics. Small language models that address specific linguistic or cultural gaps exemplify public AI initiatives designed not to compete directly in commercial markets but to create alternatives that generate public value.

For instance, the Southeast Asian Languages in One Network (SEA-LION), developed by AI Singapore, focuses on building domain-specific models tailored to Southeast Asian languages and cultural contexts – including Burmese, Chinese, English, Filipino, Indonesian, Khmer, Lao, Malay, Tamil, Thai and Vietnamese – thus addressing needs overlooked by global commercial AI development. Similarly, AINA is a project led by the Catalan government that supports the development of AI models in the Catalan language to contribute to cultural preservation.

There is also a need to conduct research into alternative development approaches beyond the transformer architecture and its scaling laws. Innovation should focus on creating less resource-intensive AI technologies. Investing in alternative model architectures is not merely a technical curiosity – it is a strategic necessity for ensuring the sustainability of AI development. By reducing the computational burden, these alternative models could lower energy demands and operational costs, making advanced AI capabilities more accessible to public institutions and research organizations.

Sustainable and open AI development ecosystem

Public AI strategies should include targeted investments in various digital public goods as key software components of AI development. Despite their critical role, many widely used software tools – from Python libraries for data preparation like pandas to machine learning libraries like scikit-learn and deep learning frameworks like PyTorch – struggle with sustainability, even as they deliver substantial public and economic value.

Governments can play a key role by funding the maintenance, security and advancement of these tools. Some have already begun: Germany's Sovereign Tech Fund supports core Python libraries, and France's national AI strategy committed €32 million to the further development of scikit-learn and the broader data science commons.[201]

There is also a growing need to invest in open-access AI safety research and open source tooling that facilitate safe and responsible development of public AI systems. Projects like Inspect (UK AI Safety Institute), Compl-AI (ETH Zurich), and ROOST (launched at the 2025 AI Action Summit) are examples of publicly oriented tools that help assess and improve AI safety, compliance and alignment.

Finally, open benchmarks are essential for measuring model capabilities and societal impact. While current benchmarks often focus on technical performance, new benchmarks should evaluate how well AI systems serve public goals, particularly within regulated industries where the stakes for public safety or general interest are high.

[201] Ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation. "Stratégie nationale pour l'intelligence artificiele – 2e phase." 8 November 2021. https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2021-11/dossier-de-presse---strat-gie-nationale-pour-l-intelligence-artificielle-2e-phase-14920.pdf Accessed 23 April 2025.

Additional measures

Aside from policies that build public AI infrastructures and capacities through the three pathways, several other measures are necessary for public AI policies to succeed. These are mainly aimed at supporting an ecosystem with sufficient talent and capacity to innovate, tools and resources to sustain collaboration, and the ability to build impactful solutions.

Invest in AI talent and capabilities

Investing in AI capabilities, knowledge and skills across the workforce and future generations is a cornerstone of any effective public AI strategy. These efforts not only enable the development of public generative AI models but also ensure that a wide range of stakeholders – research institutions, SMEs, public agencies and non-digital sectors – can adopt, adapt and influence AI technologies in ways that align with local needs.

Far from being secondary to model development, capability building is one of the most strategic roles governments can play, especially for smaller countries. While these countries may not lead in training state-of-the-art models, they can leverage open foundation models and build locally relevant AI applications – provided they invest in a strong base of local developers, researchers and institutions. Unlike capital-intensive model training, capability building offers sustainable, adaptable foundations for innovation that can evolve with the technology.

This requires public investment in skills, research and institutional support – from grants for public interest research agendas to funding the creation of dedicated institutes with clear mandates. For example, the UK's AI Safety Institute, launched in 2023, combines public funding, access to national compute resources and applied safety research to support responsible AI development. Similarly, Taiwan's AI Academy, funded by the private sector, addresses a talent gap in the workforce through training programs and industry-academic collaboration, supported by access to compute. These models demonstrate how cross-sector investments in human capital can anchor a resilient and inclusive public AI ecosystem.

Support paradigm-shifting innovation

Much of this report centers on proposing pathways to develop public generative AI, taking into account the realities of the current development paradigm, based on the transformer architecture and related scaling laws. These are the root cause of dependencies at the compute layer that make full public AI hard to attain. Public AI strategy should therefore also focus on supporting paradigm-shifting innovation. Efforts to design new AI model architectures and make AI solutions more energy-efficient are a vital element of a public AI strategy, as they could change the overall conditions for AI development by shifting dependencies in the AI stack. Changes in AI development paradigms and the underlying economics related to hardware and infrastructure could eventually enable the provision of public compute and cloud infrastructures. Such interventions would allow solutions that are today semi-public to become fully public, as dependencies on commercial hyperscalers and compute would decrease. Investments in public compute capacities, such as supercomputing centers, should be coupled with a research agenda on new, more sustainable paradigms of AI development.

Invest in software and tools for the AI ecosystem

Software has previously been defined as a distinct transversal layer that cuts across the other layers of the AI stack and plays a key role in all of them. Software is necessary to manage computing hardware, build and work with massive datasets and train or deploy generative AI models. For this reason, public AI policy should include funding for open source software development. This could take the form of a public AI infrastructure fund,[202] modeled on best practices like the German Sovereign Technology

[202] Paul Keller. "European Public Digital Infrastructure Fund." Open Future. 27 February 2023. https://openfuture.eu/publication/european-public-digital-infrastructure-fund/

Fund, which has been investing in Open Digital Infrastructure for AI;[203] the provision of targeted funding through existing public funding bodies, such as the aforementioned Fund or the UK Research and Innovation (UKRI);[204] or direct funding for open source software through AI strategies, such as France's funding for scikit-learn and the broader data science commons in its national AI strategy.[205]

Unlike hardware-focused computing investments, software development can deliver outsized impacts with relatively modest funding. Such software development supports the open source AI ecosystem, ensures collaboration and supports capacity development.

This recommendation is not limited to software but also includes other tools such as evaluation frameworks, benchmarks and development environments. It can also entail the development of novel tools, such as frameworks for democratic inputs to post-training.[206] In each case, public funding of open source software development would lower barriers to entry and ensure that no new bottlenecks are created due to proprietary control over key building blocks (as in the case of CUDA). Special attention should be given to supporting underserved components of the stack, including data processing pipelines, model evaluation suites and tools supporting public accountability in deployment. Support should extend beyond initial development to include the maintenance of key software and tools, treated as digital public goods.

Build AI deployment pathways

In principle, public infrastructures generate spillover effects – also known as positive externalities – through their use by a wide range of actors. At the same time, applications and solutions built on top of these infrastructures represent a more direct realization of public AI's value, as they can address real-world problems and deliver tangible social benefits. These applications also provide essential feedback that can inform and improve the underlying infrastructure.

Because AI is a general-purpose technology, public AI infrastructure is abstract and broadly applicable across many, if not all, spheres of life. To fulfill the potential of investments in this infrastructure, specific solutions need to be developed. Without them, there is a risk that the capacities and value embedded in public AI – such as datasets – will be captured by commercial actors and repurposed for private gain. Applications built on top of the public AI stack can address problems that are underserved by commercial AI development. The focus should therefore be on advancing public interest goals – areas where private industry lacks incentives to invest, or where there is a risk of value capture by commercial actors benefiting from first-mover advantage and network effects.[207]

Recent efforts to build public interest applications include GovLab's New Commons Challenge, which promotes responsible reuse of data for AI-driven local decision-making and humanitarian response; Gooey.ai's Workflow Accelerator, which helps organizations develop AI assistants for farmers, nurses, technicians and other frontline workers; and EarthRanger, an AI-powered wildlife conservation platform stewarded by the Allen Institute for AI.

[203] Adriana Groh. "AI Sovereignty Starts with Open Infrastructure." Sovereign Tech Agency. 27 February 2025. https://www.sovereign.tech/news/ai-sovereignty-open-infrastructure/

[204] Tom Milton, Cailean Osborne, Matt Pickering. "A UK Open-Source Fund to Support Software Innovation and Maintenance." Centre for British Progress. 17 April 2024. https://britishprogress.org/uk-day-one/a-uk-open-source-fund-to-support-software-innovati

[205] "Stratégie nationale pour l'intelligence artificiele – 2e phase." 8 November 2021. https://www.enseignementsup-recherche.gouv.fr/sites/default/files/2021-11/dossier-de-presse---strat-gie-nationale-pour-l-intelligence-artificielle-2e-phase-14920.pdf Accessed 23 April 2025.

[206] "A Roadmap to Democratic AI." The Collective Intelligence Project. March 2024. https://www.cip.org/research/ai-roadmap

[207] Nik Marda, Jasmine Sun and Mark Surman, ibid.

Coda: mission-driven public AI policy

The characteristics of public AI outlined in chapter 3 remain constant regardless of an AI system's architecture or capabilities. Public AI should be open and accessible, create public value and remain under public control. In this sense, the vision of public AI is intentionally technologically agnostic.

However, public AI policy must also offer clarity on the kinds of technologies needed to serve the public interest. This requires the policy debate to engage – at least to some degree – with fundamental questions about the types of AI systems being developed and deployed.

Specifically, public AI policy needs to reckon with the idea of Artificial General Intelligence (AGI), a vision promoted by many leading commercial AI labs. The fuzzy and controversial term is typically used to describe AI systems that equal or surpass human intelligence. OpenAI, for instance, defines AGI as "highly autonomous systems that outperform humans at most economically valuable work."[208] Policymakers should approach the concept of AGI with caution, as it often feeds into hype cycles and fosters a sense of technological determinism.[209]

An alternative is to treat AI as cultural technologies[210] – or "normal AI:"[211] technologies that that may fundamentally shape our societies without necessarily exhibiting superhuman capabilities.

The public AI strategy proposed in this report includes the development of a state-of-the-art foundation model. While avoiding the hype driving much of commercial AI development, public AI policy should be grounded in a careful analysis of both the demand for and supply of AI capabilities. In this case, provisioning a robust, state-of-the-art model as an open source technology would fill a critical market gap, as commercial models typically offer, at best, open weights without transparent documentation on training data or development processes.

As outlined in the model pathway above, we recommend supporting both the creation of a public foundation model and the development of various small models. This two-pronged approach fosters an open source AI ecosystem that benefits from a central, capable model while also investing in more specialized, sustainable solutions. The exact balance between these two directions should be guided by a more detailed analysis of both the economics of AI development and the needs of the public.

Currently, AI strategies often lack specificity on either the societal needs AI should address or the technical capabilities required to meet them. Instead, public investments in AI are frequently motivated by a generalized belief that industrial policy must support disruptive innovation as a remedy for economic stagnation. Too often, these investments follow the demands of industry, rather than focusing on the real, everyday needs of citizens.[212]

Proposals for public investment in computing power illustrate this problem. These often lack a clear analysis or justification for the types of AI systems to be developed or the computing resources required. As a result, public AI policy risks replicating the same unsustainable "more is better" investment logic that drives much of the commercial AI sector. A shift toward building AI as public digital infrastructure – guided by the principles outlined in this report – offers a way to avoid these pitfalls and align AI development with the public good.

[208] Lauren Leffer. "In the Race to Artificial General Intelligence, Where's the Finish Line?." Scientific American. 25 June 2024. https://www.scientificamerican.com/article/what-does-artificial-general-intelligence-actually-mean/

[209] Zuzanna Warso. "The Digital Innovation We Need. Three lessons on EU Research and Innovation funding." Open Future. 12 November 2024. https://openfuture.eu/publication/the-digital-transformation-we-need/

[210] Henry Farrell, et al. "Large AI models are cultural and social technologies." Science. 13 March 2025, Vol 387, Issue 6739, pp. 1153-1156. https://www.science.org/stoken/author-tokens/ST-2495/full

[211] Arvind Narayanan and Sayash Kapoor. ibid.

[212] Zuzanna Warso and Meret Baumgartner. "Putting money where your mouth is? Insights into EU R&I funding for digital technologies." Critical Infrastructure Lab. 2025. https://openfuture.eu/publication/putting-money-where-your-mouth-is/

Address | Contact

Bertelsmann Stiftung
Carl-Bertelsmann-Straße 256
33311 Gütersloh
Phone +49 5241 81-0
www.bertelsmann-stiftung.de

Dr. Felix Sieker
Project Manager
Digitization and the Common Good
Phone +49 30 275788-156
felix.sieker@bertelsmann-stiftung.de

Supported by

OPEN
_FUTURE

Commissioned by

BertelsmannStiftung