The Compute Divide as an Access-Countercapacity Gap

AI Infrastructure, Epistemic Dependence, and Public-Interest Governance

Author: Benjamin Metzig
Affiliation: Wissenschaftswelle.de
Version: Version 1.0
Date: 15 June 2026
Publication type: Exploratory research report / working paper
Language: English

DOI: 10.5281/zenodo.20700430

Contents

Abstract

Advanced artificial intelligence is often discussed through model capability, productivity, safety, or regulation. This exploratory report examines a more infrastructural problem: the compute divide. It argues that the compute divide is best understood as the gap between AI access and AI countercapacity: the difference between being able to use AI systems and being able to independently scrutinize, reproduce, regulate, or contest them. Drawing on selected academic, policy, technical, and institutional sources, the report maps infrastructure concentration across compute, cloud platforms, advanced chips, data centers, energy, model access, transparency, and evaluation environments. The strongest current evidence concerns concentrated AI infrastructure, cloud and chip dependencies, rapid data-center electricity growth, transparency deficits in foundation models, and uneven global AI readiness. The strongest present-day governance risk is a preparedness gap: AI adoption is advancing faster than many public-interest actors' capacity to independently inspect or challenge high-impact systems. Open-weight models, smaller models, algorithmic efficiency, public compute, structured access, competition policy, and regulatory expertise are real counterforces, but they mitigate different layers of the divide unevenly. The report concludes that public compute, audit rights, transparency duties, regulatory capacity, competition remedies, and democratic contestation mechanisms should be understood as attempts to build public-interest countercapacity against an emerging inequality in the means of AI-enabled knowledge production.

Keywords

artificial intelligence; compute divide; countercapacity; AI infrastructure; foundation models; cloud concentration; regulatory capacity; public compute; digital inequality; epistemic dependence

Suggested Citation

Metzig, B. (2026). *The Compute Divide as an Access-Countercapacity Gap: AI Infrastructure, Epistemic Dependence, and Public-Interest Governance* (Version 1.0). Wissenschaftswelle.de. https://doi.org/10.5281/zenodo.20700430

Executive Summary

This exploratory report develops the compute divide as a sociological and governance concept. Its central argument is that advanced AI societies may become stratified not only by who can use AI systems, but by who has countercapacity: the infrastructure, rights, expertise, money, institutional authority, and practical access needed to independently scrutinize and shape them. Consumer access to AI tools can expand while countercapacity remains concentrated.

The report uses "compute" in a broad infrastructural sense. Compute includes accelerator chips, memory, networking, cloud services, data-center capacity, electricity, cooling, capital, engineering expertise, model access, evaluation environments, and institutional ability to operate these resources. This definition is necessary because advanced AI is not produced by chips alone. It is produced by an infrastructure stack that combines hardware, software, platforms, energy, capital, supply chains, standards, and organizations.

The strongest empirical foundation concerns AI infrastructure concentration. Evidence from Stanford HAI, OECD, the World Bank, competition authorities, transparency benchmarks, energy agencies, and public compute initiatives supports the claim that several layers of the AI infrastructure stack are unevenly distributed. These include advanced semiconductors, cloud platforms, data centers, electricity capacity, talent, and access to frontier or frontier-adjacent model ecosystems.

The strongest conceptual contribution is to extend digital inequality theory into the infrastructure layer of knowledge production. Earlier digital divide research emphasized access, skills, use, and outcomes. The compute divide adds a deeper question: who controls the infrastructures through which AI-enabled knowledge is produced, validated, operationalized, and contested?

The strongest governance finding is the preparedness gap. Public authorities, universities, civil society, smaller firms, and less-resourced countries may have formal responsibilities or public expectations to regulate, audit, reproduce, or challenge AI systems while lacking practical access, compute, documentation, expertise, funding, and evaluation environments comparable to the developers and infrastructure providers they seek to scrutinize.

The report treats open-weight models, smaller models, algorithmic efficiency, public compute, structured access, AI safety institutes, standards, and competition policy as real counterforces. The revised thesis is conditional and institutional: compute concentration matters most when it creates a gap between AI access and AI countercapacity. The central policy question is therefore not only whether more actors can use AI, but whether public-interest actors can independently verify, challenge, and shape the AI systems on which they increasingly depend.

How To Read This Report

This report should be read as an exploratory synthesis, not as a systematic review or a prediction. Its strongest empirical claims concern measurable infrastructure concentration, energy and data-center growth, transparency gaps, and public compute responses. Its broader claims about epistemic dependence, democratic contestation, and technological stratification are mechanism-based interpretations. The evidence labels, boundaries, and appendix matrix are therefore part of the argument rather than decorative cautions.

1. Introduction

1.1 Opening Problem

Artificial intelligence is becoming a general-purpose layer of social, economic, scientific, and administrative life. AI systems are increasingly embedded in search, software development, education, public administration, scientific research, media production, enterprise workflows, and security-relevant analysis. Yet the ability to use AI should not be confused with the ability to produce, evaluate, reproduce, govern, or contest the systems that mediate these domains.

The central problem of this report is the possible emergence of a compute divide: an unequal distribution of the material, economic, institutional, and epistemic capacities required to participate meaningfully in advanced AI societies. Many individuals and organizations can now access AI tools through consumer interfaces, APIs, open-weight releases, or enterprise platforms. Far fewer actors can train frontier-scale models, secure large amounts of accelerator compute, operate global inference infrastructure, inspect training data and training compute, reproduce model claims, evaluate systemic risks, or enforce regulatory requirements with technical independence.

This matters because AI systems increasingly function as components of knowledge infrastructure. They help define what can be searched, summarized, classified, predicted, recommended, evaluated, generated, and administratively acted upon. This claim is not a settled empirical law about all AI systems. It is a conceptual diagnosis supported by evidence from AI adoption, platform integration, public-sector use, and model-evaluation debates. If the infrastructure required to shape and scrutinize these systems is concentrated, then knowledge power may also become concentrated.

Evidence status: Conceptual diagnosis / well-supported assumption.

1.2 Why The Compute Divide Is More Than A Technical Issue

The term "compute" can sound narrow, as if the problem were simply access to graphics processing units or accelerator chips. This report uses compute in a broader infrastructural sense. Compute includes processors, accelerators, memory, networking, cloud access, data-center capacity, electricity, cooling, platform access, and the organizational capability to train, evaluate, deploy, and monitor AI systems.

In this sense, compute is not merely a technical input. It is part of an institutional system through which AI-enabled knowledge is produced and made socially consequential. The sociological question is therefore not only "who has GPUs?" but "who possesses the means of AI-enabled knowledge production?"

Earlier digital inequality debates often focused on access to computers, internet connections, skills, autonomy, and meaningful use. The compute divide shifts the focus from digital use to infrastructural control. The key issue is not only who can use a model, but who can build it, test it, reproduce it, govern it, and challenge its role in social decision-making.

Evidence status: Conceptual claim with supporting literature.

1.3 Central Thesis

This report advances the following thesis:

> The compute divide is the gap between AI access and AI countercapacity. It emerges when many actors can use advanced AI systems, but far fewer have the infrastructure, rights, resources, and expertise needed to scrutinize and shape those systems independently.

This thesis shifts the center of analysis from mere access to countercapacity. A society can have broad AI access and still lack robust public-interest countercapacity. The strongest claim is therefore not that AI necessarily removes public control, but that advanced AI can create a layered capacity hierarchy: some actors build and shape

systems, while others use them, depend on them, regulate them with incomplete access, or contest them from a weaker infrastructural and epistemic position.

1.4 Structure Of The Report

Section 2 defines the research aim and scope. Section 3 explains the exploratory methodology and evidence categories. Section 4 develops the conceptual framework: digital divide, AI divide, compute divide, countercapacity, technological stratification, and means of knowledge production. Section 5 summarizes the current state of knowledge. Section 6 analyzes mechanisms and drivers. Section 7 develops four scenarios using a two-axis scenario matrix. Section 8 assesses and prioritizes risks. Section 9 examines counterarguments and limitations. Section 10 discusses implications for science, regulation, transparency, democracy, markets, global development, and local infrastructure. Section 11 concludes.

---

2. Research Aim And Scope

2.1 Research Question

The main research question is:

> What mechanisms, actors, and risks can be identified when the compute divide is understood as an access-countercapacity gap in advanced AI societies?

Supporting questions are:

1. How does the compute divide differ from the digital divide and the broader AI divide?
2. Which layers of AI infrastructure are most relevant for social and institutional power?
3. How do compute, cloud, chips, energy, capital, talent, model access, and evaluation capacity interact?
4. How might infrastructure concentration affect scientific independence, regulatory capacity, democratic oversight, global development, and local communities?
5. Which countervailing developments may reduce or redirect the compute divide?
6. Which risks are already current, which are emerging, and which should be treated only as conditional high-risk scenarios?

2.2 Research Aim

The report has six aims:

1. To make the compute divide visible as a distinct access-countercapacity gap with potential stratifying consequences.
2. To map the infrastructure stack underlying advanced AI development and governance.
3. To distinguish AI use access from production, deployment, evaluation, reproduction, governance, and contestation capacity.
4. To synthesize evidence from AI infrastructure, competition policy, transparency benchmarks, digital inequality, energy systems, science studies, and AI governance.
5. To develop plausible scenarios using a two-axis matrix of infrastructure concentration and public-interest countercapacity.
6. To prepare a risk and impact assessment that can guide further research and policy debate.

2.3 Scope

This report focuses on the compute divide as an access-countercapacity gap and as a possible mechanism of technological stratification. It examines how unequal access to large-scale compute, cloud infrastructure, semiconductors, energy, capital, model access, and evaluation capacity may shape the distribution of power in advanced AI societies.

The primary time horizon is 2020 to 2030. The report is present-day and near-future oriented. Medium-term scenarios are used only as analytical tools, not as forecasts.

The geographic focus is international, with particular attention to the United States, the European Union, China where relevant through public sources, high-income AI infrastructure centers, smaller states, and low- and middle-income countries. The report does not claim to map all national AI systems equally.

The technological focus includes foundation models, frontier or frontier-adjacent AI systems, large language models, multimodal systems, AI agents where relevant, AI-as-a-service platforms, cloud infrastructure, accelerator hardware, data centers, public compute, open-weight models, and evaluation environments.

2.4 Exclusions

This report does not:

- provide operational instructions for building or misusing advanced AI systems;
- investigate classified surveillance, intelligence, or military systems beyond publicly available sources;
- make predictions about the long-term dominance of specific firms or states;
- claim that compute alone determines AI capability;
- treat all AI applications as frontier-scale AI;
- treat open-source or open-weight systems as either inherently safe or inherently dangerous;
- present a complete systematic literature review;
- present original empirical measurements unless explicitly stated.

2.5 Key Definitions

Artificial intelligence: Computational systems that perform tasks commonly associated with human cognitive abilities, including language processing, pattern recognition, prediction, classification, decision support, generation, planning, and problem-solving.

Advanced AI systems: AI systems with broad capabilities, high performance across multiple tasks, significant deployment potential, and substantial requirements for compute, data, engineering expertise, evaluation, and governance.

Foundation models: Large, general-purpose AI models trained on broad datasets and adaptable to many downstream tasks through prompting, fine-tuning, tool use, or integration into applications.

Frontier AI systems: Among the most capable AI systems available at a given time, with distinctive governance challenges because their capabilities, risks, and evaluation needs may exceed those of widely accessible systems.

Compute: The computational resources used to train, fine-tune, evaluate, and operate AI systems, including processors, accelerators, memory, networking, data-center capacity, cloud access, and related infrastructure.

Compute divide: The unequal distribution of access to the computational and infrastructural resources required to shape and scrutinize advanced AI systems. In the sharpened formulation used throughout this report, the compute divide is the gap between AI access and AI countercapacity.

AI infrastructure stack: The layered set of resources that make advanced AI possible, including chips, memory, networking, software frameworks, cloud platforms, data centers, electricity, cooling, data pipelines, model access, evaluation tools, standards, expertise, capital, and institutions.

AI access: The ability to use AI systems through interfaces, APIs, open-weight models, cloud services, enterprise platforms, or local applications.

AI countercapacity: The practical ability to independently scrutinize, reproduce, evaluate, contest, regulate, or reshape AI systems. Countercapacity combines compute, model access, documentation, data access where lawful, technical expertise, legal rights, institutional authority, funding, and evaluation environments.

Cloud infrastructure: Networked computing infrastructure offered as a service, including storage, compute, databases, development tools, AI services, and deployment environments. In AI, cloud infrastructure matters because many organizations rent or access compute through hyperscale platforms rather than owning the full hardware stack.

Data centers: Physical facilities that house servers, storage, networking equipment, power systems, cooling systems, and related infrastructure. Data centers make cloud and AI services possible and connect the compute divide to electricity, water, land, grid planning, and local governance.

Open-weight models: AI models whose trained parameters are publicly or widely available for download or reuse under specified terms. Open-weight models can support research, adaptation, local deployment, and competition, but they do not necessarily provide training data, training compute, safety documentation, or deployment infrastructure.

Public compute: Compute resources provided, subsidized, coordinated, or governed by public institutions or public-interest programs. Public compute can support researchers, public agencies, startups, education, safety work, and evaluation, but its effectiveness depends on scale, access rules, staffing, continuity, energy planning, and integration with oversight needs.

Evaluation environment: The technical, legal, and institutional setting in which an AI system is tested or assessed. This can include model access, data access, benchmark infrastructure, secure testing facilities, audit rights, documentation, compute, and expertise.

Technological stratification: A form of social and institutional inequality in which access to critical technologies and infrastructures shapes durable differences in power, opportunity, knowledge production, institutional autonomy, and governance capacity. In this report, technological stratification is treated as a possible consequence of the compute divide, not as an already fully proven social order.

Means of knowledge production: The resources, institutions, infrastructures, methods, and validation mechanisms through which knowledge is produced, tested, legitimized, and made socially consequential.

Regulatory capacity: The ability of public institutions to understand, monitor, evaluate, and govern AI systems through legal authority, technical expertise, access to information, testing capacity, enforcement tools, and institutional resources.

Epistemic dependence: A condition in which public-interest actors must rely on information, interfaces, evaluation environments, or access conditions controlled by the actors they seek to evaluate or regulate.

Preparedness gap: A mismatch between the societal importance of advanced AI systems and the practical capacity of public institutions, science, and civil society to understand, evaluate, reproduce, and govern them.

Scenario: A structured, evidence-informed exploration of how current developments could plausibly unfold under specified assumptions. Scenarios are not predictions.

---

3. Methodological Note

3.1 Exploratory Nature

This report is exploratory in nature. It does not claim to provide a systematic review of the entire research field, nor does it present original empirical data unless explicitly stated. Instead, it maps relevant developments, identifies emerging patterns, evaluates available evidence according to its strength, and develops plausible present-day and future-oriented scenarios. Established findings, well-supported assumptions, indicative evidence, scenario-based projections, and speculative claims are explicitly distinguished throughout the report.

This methodological stance is necessary because the compute divide is not a single mature research object. It crosses AI research, cloud infrastructure, semiconductor supply chains, energy systems, competition policy, science studies, digital inequality, public administration, and democratic accountability. Some parts of the argument are empirically robust, such as data-center electricity growth, cloud concentration in particular markets, and documented transparency gaps. Other parts are interpretive, such as the claim that these patterns may weaken independent scrutiny. Others are scenario-based, such as the risk of durable epistemic dependency.

3.2 Source Strategy

The report uses several source types:

- academic studies and preprints;
- peer-reviewed conference papers;
- international organization reports;
- government reports and regulatory documents;
- competition authority reports;
- transparency benchmarks;
- energy and data-center reports;
- digital inequality literature;
- AI governance and accountability literature;
- policy analyses from credible NGOs and research organizations;
- selected media or tertiary sources only as pointers, not as central evidence.

Primary and institutional sources are weighted more heavily than commentary. Secondary sources are used for interpretation and synthesis. Tertiary sources are not used as foundations for central claims.

Sources were selected through targeted exploratory search across AI infrastructure, digital inequality, cloud competition, transparency benchmarks, public compute, AI governance, auditability, and energy reports. Selection prioritized sources with direct relevance to the access-countercapacity mechanism, traceable institutional or academic provenance, and concrete data, page references, or policy mechanisms. Because this is not a systematic review, the source base should be treated as a structured evidentiary map rather than an exhaustive bibliography.

3.3 Evidence Labels

Label	Meaning
Established finding	Supported by robust sources, official statistics, multiple independent studies, or well-established empirical findings.
Well-supported assumption	Not conclusively proven, but plausible and supported by several sources, technical conditions, or institutional mechanisms.
Indicative evidence	Supported by early studies, case examples, limited datasets, or observable patterns.

Label	Meaning
Scenario-based projection	A plausible future-oriented pathway based on current trends and specified assumptions.
Speculative claim	Possible but weakly supported, highly uncertain, or dependent on contested assumptions.
Open question	Unclear, contradictory, insufficiently studied, or dependent on future developments.

3.4 Claim-Evidence Rule

Major claims in this report should be read through a three-part evidence rule:

1. **What the source directly supports:** the concrete empirical, legal, technical, or conceptual point established by the cited source.
2. **What mechanism the source helps explain:** how that point relates to access, countercapacity, infrastructure concentration, transparency, reproducibility, or governance.
3. **What the source does not prove:** the boundary beyond which the claim would become overgeneralized, speculative, or unsupported.

This rule is especially important for claims about democratic contestation, scientific control, global dependency, and epistemic dependence. These are not treated as already proven social outcomes. They are treated as plausible implications supported by narrower evidence chains.

3.5 Methodological Limits

Several data limitations matter for the argument:

- training compute, training data, model internals, and evaluation pipelines are often opaque;
- cloud capacity, data-center capacity, and market data can be proprietary or proxy-based;
- energy projections are scenario-based and sensitive to assumptions;
- government readiness indices are constructed indicators, not direct measures of compute ownership;
- open-weight and frontier-model performance comparisons change quickly;
- democratic audit limitations need concrete case studies before strong causal claims are made.

These limits do not invalidate the report's central thesis, but they require careful language. The report should not state that the compute divide has already produced a fully consolidated social order. It should state that advanced AI infrastructure creates measurable and plausible conditions for a new form of technological stratification.

---

4. Conceptual Framework

4.1 From Digital Divide To Compute Divide

Digital inequality research moved beyond binary access to devices and internet connections toward differentiated use, skills, autonomy, support, and outcomes. The compute divide extends this trajectory into the infrastructure layer of AI-enabled knowledge production. It asks not only who can access digital services, but who can shape the systems that produce, classify, recommend, evaluate, and operationalize knowledge.

The digital divide remains relevant, but it is no longer sufficient. AI tools may be widely accessible while the capacity to build, audit, reproduce, and govern advanced AI systems remains narrowly distributed. Thus, the compute divide is best understood as a continuation and deepening of digital inequality: from access to devices and networks, to meaningful use, to unequal control over the infrastructure of AI-mediated knowledge production.

Evidence status: Conceptual claim with supporting literature.

4.2 Compute: Narrow Technical Meaning And Broader Infrastructure Stack

The term "compute" has a narrow technical meaning and a broader analytical use in this report. In the narrow sense, compute refers to computational resources used to train, fine-tune, evaluate, and operate AI systems: processors, accelerators, memory, networking, storage, and data-center capacity. This narrow meaning matters because advanced AI development is materially constrained by hardware availability, cloud allocation, energy, cooling, and capital expenditure.

In the broader sense, the compute divide is shorthand for inequality across the AI infrastructure stack. That stack includes chips, cloud platforms, data centers, electricity, cooling, capital, technical expertise, model access, documentation, benchmarks, evaluation environments, and legal or institutional access rights. This broader use risks conceptual overextension if left implicit. The report therefore uses "compute divide" as a deliberately compact term for a wider infrastructure inequality, while distinguishing the underlying layers whenever the argument requires precision.

This distinction prevents two errors. The first error is GPU reductionism: treating the compute divide as only a question of chip counts. The second error is conceptual inflation: treating compute as a label for all AI-related power. The report's more precise claim is that computational resources sit at the center of a wider AI infrastructure stack, and that countercapacity depends on several layers of that stack at once.

4.3 The Four Capacity Layers

The compute divide can be operationalized across four capacity layers:

Capacity layer	Core question	Typical privileged actors	Typical constrained actors	Main evidence
Production	Who can build advanced AI systems?	Frontier labs, hyperscalers, major states	Universities, SMEs, smaller states	Stanford AI Index; OECD AI infrastructure; World Bank
Deployment	Who can run AI at scale?	Cloud providers, enterprise platforms, large public agencies	Smaller public agencies, smaller firms, many LMIC institutions	OECD; FTC; Ofcom/CMA; World Bank
Evaluation and reproduction	Who can test and reproduce claims?	Developers, selected partners, some safety institutes	Academia, civil society, journalists, affected communities	FMTI; NTIA; Royal Society; Besiroglu et al.

Capacity layer	Core question	Typical privileged actors	Typical constrained actors	Main evidence
Governance and contestation	Who can regulate and challenge systems?	Well-resourced regulators, courts, platform owners	Smaller regulators, affected communities, less-resourced countries	NTIA; AI Act context; Amnesty; Frontiers; Springer AI and Ethics

This framework prevents a shallow reading of the compute divide as "some actors have more GPUs." The deeper issue is layered capacity asymmetry. Broad user access can coexist with narrow production, evaluation, and governance capacity.

4.4 From Linear Chain To Feedback Model

The compute divide should not be represented as a one-way path from technical scale to social stratification. A linear chain is useful for orientation, but it risks overstating causality. A better model treats the compute divide as a feedback system in which infrastructure concentration, AI adoption, public countercapacity, open ecosystems, regulation, and competition interact.

Figure 1. Core feedback model of the compute divide

Step	Mechanism	Feedback or counterforce
1	AI adoption and capability growth increase demand for compute, cloud, chips, energy, data-center capacity, capital, and specialized expertise.	Rising demand can reinforce infrastructure investment and platform dependency.
2	Resource demand can strengthen infrastructure concentration where economies of scale, supply-chain bottlenecks, and cloud ecosystems matter.	Concentrated infrastructure can shape which AI systems are built and deployed.
3	Concentration becomes socially consequential when it creates a gap between AI access and AI countercapacity.	Broad use access can coexist with limited evaluation, reproduction, and governance capacity.
4	The countercapacity gap can produce epistemic dependence and preparedness gaps.	Public-interest actors may rely on controlled disclosures, APIs, or negotiated access.
5	These gaps can create scientific, regulatory, democratic, local, and global governance risks.	Risks can trigger policy, market, and institutional responses.
6	Counterforces include public compute, structured access, standards, competition policy, open-weight models, smaller models, and efficiency gains.	Counterforces can reduce parts of the gap, but not necessarily all layers of countercapacity.
7	Energy, local planning, and supply-chain constraints feed back into infrastructure concentration.	Physical constraints affect where and by whom AI infrastructure can be scaled.

This model makes five points explicit.

First, AI adoption and capability growth can increase demand for compute, cloud, chips, energy, data-center capacity, capital, and specialized expertise. Second, this resource demand can strengthen infrastructure concentration and platform dependency where economies of scale, supply-chain bottlenecks, and cloud ecosystems matter. Third, concentration becomes socially consequential when it creates a gap between AI access and AI countercapacity. Fourth, counterforces can interrupt or redirect the mechanism: open-weight models, smaller models, algorithmic efficiency, public compute, structured access, regulatory expertise, standards, and competition policy can all reduce some asymmetries. Fifth, the same counterforces can be partial or unstable. Open-weight models may improve use and research access while leaving frontier training, secure deployment, or regulatory testing concentrated. Public compute may help researchers while remaining far smaller than private infrastructure. Competition may grow at the application layer while cloud and chip layers stay concentrated.

The model is therefore conditional. It does not say that infrastructure concentration automatically produces social stratification. It says that technological stratification becomes more plausible when concentration persists across several layers and public-interest countercapacity fails to scale with AI adoption.

4.5 Evidence Chain For The Core Thesis

Claim	Direct evidence used	Mechanism supported	What this does not prove	Evidence label
Advanced AI depends on concentrated infrastructure resources	Stanford AI Index; OECD AI infrastructure; chip and cloud policy sources; energy/data-center reports	Resource intensification and infrastructure concentration	Compute alone determines capability	Established finding / well-supported assumption
The compute divide is an access-versus-countercapacity gap	Digital inequality theory; compute divide literature; AI governance and audit literature	User access can expand while production, reproduction, governance, and contestation capacity remain unequal	All AI use creates dependency	Conceptual claim / well-supported assumption
Transparency gaps create epistemic dependence risks	FMTI 2025; NTIA; audit ecosystem literature	Public-interest actors rely on controlled disclosures and interfaces	Actual democratic failure is already proven	Established benchmark evidence / indicative implication
Public compute and open-weight models mitigate but do not solve the divide	NAIRR; EuroHPC; Stanford open/closed benchmark evidence; open-weight policy sources	Counterforces reduce some barriers but not all layers of countercapacity	Open models are irrelevant or sufficient	Counterargument / well-supported assumption
Global AI countercapacity gaps can create rule-taker risks	World Bank; UNCTAD; ILO/UN; readiness indices	Adoption can outpace domestic or regional ability to evaluate, adapt, and negotiate AI infrastructure	All less-resourced countries are passive or permanently dependent	Established finding / scenario-based implication

4.6 From Infrastructure Inequality To Technological Stratification

The title-level term "technological stratification" requires a transmission model. Infrastructure concentration does not automatically become social stratification. It becomes stratifying only when resource asymmetries are translated into durable differences in institutional position, autonomy, agenda-setting capacity, and legitimate authority.

This report treats the compute divide as a potential mechanism of technological stratification through five transmission pathways:

Pathway	How infrastructure inequality becomes stratifying	Main affected actors	Evidence label
Resource accumulation	Actors with compute, cloud access, data-center capacity, capital, and talent can improve systems, attract users, generate revenue, and reinvest at a scale others cannot match.	Frontier labs, cloud providers, large states, startups without infrastructure access	Established finding / well-supported assumption
Gatekeeping and dependency	Infrastructure owners can shape terms of access, pricing, documentation, model availability, evaluation conditions, and procurement dependencies.	Public agencies, universities, SMEs, less-resourced countries	Well-supported assumption
Epistemic authority	Actors that build and operate systems can become privileged producers of benchmarks, documentation, safety claims, and technical interpretations.	Regulators, researchers, journalists, courts, civil society	Indicative evidence / well-supported assumption
Institutional reproduction	Training, hiring, grants, partnerships, and research agendas may increasingly depend on access to proprietary or state-scale infrastructure.	Universities, public research institutions, AI safety institutes	Indicative evidence / scenario-based implication
Legitimation and contestability	Systems become socially consequential when their outputs are trusted, procured, standardized, or legally embedded while affected actors lack practical contestation capacity.	Affected communities, workers, citizens, civil society	Well-supported assumption / indicative evidence

This model is intentionally narrower than a full theory of class or status stratification. It does not claim that compute concentration alone creates a complete social hierarchy. It claims that unequal AI countercapacity can become a stratifying mechanism for knowledge-producing and governing institutions. The strongest current evidence supports infrastructure concentration and countercapacity asymmetry; the broader social reproduction effects remain a research agenda.

4.7 Means Of AI-Enabled Knowledge Production

The phrase "means of knowledge production" is used analytically rather than polemically. It refers to the infrastructures, institutions, datasets, methods, evaluation environments, and interpretive authority through which knowledge is produced and made actionable. In AI societies, these means include model training pipelines, cloud platforms, benchmarks, APIs, documentation, deployment environments, monitoring systems, and feedback loops.

If advanced AI systems increasingly mediate research, search, decision support, education, administration, and information access, then control over AI infrastructure becomes a form of knowledge power. The compute divide is therefore not only an economic inequality. It is also an epistemic inequality.

4.8 The Preparedness Gap

The preparedness gap is the mismatch between AI adoption and public-interest capacity. It arises when advanced AI systems diffuse into social, economic, and administrative life faster than regulators, universities, courts, civil society, and affected communities can independently inspect or contest them.

The gap has several dimensions:

- legal authority without technical access;
- technical expertise without compute access;
- transparency requirements without verification environments;
- public accountability expectations without institutional staffing;
- academic critique without reproduction capacity;
- civil-society contestation without documentation, standing, or resources;
- national AI strategies without domestic infrastructure.

This concept is central to the report because it turns the compute divide from an abstract distributional problem into a concrete governance problem.

---

5. Current State Of Knowledge

5.1 Main Current-State Finding

Current evidence supports a strong but bounded claim: advanced AI development and governance are increasingly tied to a concentrated infrastructure stack. This stack includes accelerator chips, high-bandwidth memory, cloud platforms, data centers, electricity, cooling, capital, specialized engineering labor, model access, and evaluation environments.

The concentration is not uniform across all layers. Some layers are highly concentrated in specific firms, countries, or supply-chain nodes. Others are more distributed but still require capital and institutional capacity. The important point is that advanced AI is not only a software phenomenon. It is a physical, economic, and institutional infrastructure.

This section is therefore organized as evidence-to-mechanism analysis. Each evidence cluster is used to support a specific mechanism in the compute-divide model: production gap, platform dependency, hardware bottleneck, energy and locality, transparency and epistemic dependence, and global countercapacity gap.

5.2 Quantitative Evidence Dashboard

The following dashboard summarizes verified data points used in the report. It is intentionally selective: the purpose is not to list every source, but to make the main quantitative anchors visible and to state their limits.

Theme	Verified data point / fact	Source and location	Evidence status	Recommended use	Caution
Industry production gap	Industry produced 91.2% of notable AI models; in 2025, Epoch identified 93 industry notable models and 2 academic notable models.	Stanford AI Index 2026, p. 19, Fig. 1.1.4	Established finding	Industry/academia production asymmetry	Not a measure of all AI research
Compute capacity growth	Global AI compute capacity grew 3.3x per year since 2022 and reached 17.1 million H100-equivalents.	Stanford AI Index 2026, p. 15	Established finding	Infrastructure scaling baseline	Does not prove compute alone drives capability
Data-center geography	The United States hosted 5,427 data centers in 2025, more than ten times any other country; Germany had 529, the United Kingdom 523, and China 449.	Stanford AI Index 2026, p. 10, p. 15, p. 33	Established finding	Geographic concentration	Counts do not show facility size or accelerator capacity
Chip fabrication bottleneck	TSMC fabricates virtually every leading AI chip.	Stanford AI Index 2026, p. 10, p. 15, p. 32	Established finding	Supply-chain concentration	Does not imply one firm alone controls AI
Open vs closed model performance	As of March 2026, the top closed model led the top open model by 3.3%; six of the top ten Arena models were closed.	Stanford AI Index 2026, p. 72; exact score comparison p. 76-77; observation date March 2026	Established finding / counterargument	Calibrate open-weight claims	Arena is one benchmark family, not all capability
Global data-center electricity	Global data-center electricity consumption was about 415 TWh in 2024 and rises to around 945 TWh by 2030 in the IEA Base Case.	IEA Energy and AI, p. 49, p. 63; annex table p. 260	Established finding / scenario projection	Global energy baseline	Data-center electricity is not identical with AI electricity

Theme	Verified data point / fact	Source and location	Evidence status	Recommended use	Caution
Data-center growth rate	IEA Base Case estimates about 15% annual growth in data-center electricity consumption from 2024 to 2030.	IEA Energy and AI, p. 63	Scenario-based projection	Shows rapid infrastructure growth	Projection depends on assumptions
Accelerated servers	Accelerated-server electricity consumption is projected to grow 30% annually in the IEA Base Case; conventional server electricity grows 9% annually.	IEA Energy and AI, p. 63	Scenario-based projection	AI-relevant server-load proxy	Accelerated servers are an imperfect proxy for AI workloads
U.S. data-center electricity	U.S. data centers consumed 176 TWh in 2023, 4.4% of U.S. electricity; scenario range for 2028 is roughly 325-580 TWh.	LBNL / DOE 2024 report, p. 6-7, p. 52	Established finding / scenario projection	U.S. baseline	Projection range is wide and U.S.-specific
U.S. water and emissions	U.S. data centers' indirect water footprint in 2023 was nearly 800 billion liters; associated emissions were about 61 billion kg CO2e.	LBNL / DOE 2024 report, p. 57	Established finding / modeled estimate	Environmental implication	Depends on grid mix and indirect accounting
Ireland electricity case	Irish data centers used 6,969 GWh in 2024, representing 22% of metered electricity, up from 5% in 2015.	CSO Ireland, 2025 official release	Established finding	Non-U.S. national case	Ireland is a distinctive data-center hub
Ireland growth	Quarterly Irish data-center consumption rose from 290 GWh in Q1 2015 to 1,829 GWh in Q4 2024, a 531% increase.	CSO Ireland, 2025 official release	Established finding	Local-infrastructure case box / chart	Metered electricity only
Foundation-model transparency	FMTI reports that the average transparency score fell from 58 in 2024 to 40 in 2025; overall 2025 average was 40.69/100.	FMTI 2025, p. 1-3, p. 18	Benchmark evidence	Auditability and epistemic dependence	Benchmark design matters
Training compute transparency	Training compute transparency remains especially weak; IBM and Writer are the only two companies above 50% in the compute domain.	FMTI 2025, p. 21	Benchmark evidence	Compute-transparency gap	Does not prove misuse of undisclosed compute
AI infrastructure concentration	OECD identifies concentrated AI infrastructure segments, including ASML in advanced lithography, TSMC in advanced AI chip fabrication, Nvidia in GPUs for AI compute, AWS/Google/Microsoft in cloud, and major EDA providers.	OECD AI Infrastructure Competition, p. 24, Table 1	Established finding with OECD caveat	Multi-layer concentration	OECD notes this is not a formal market-definition exercise
Cloud market concentration	AWS and Microsoft had a combined 70-80% share of UK public cloud infrastructure services in 2022; Google had 5-10%.	Ofcom cloud market referral, 2023	Established finding	Cloud lock-in case	UK-specific
Cloud switching barriers	Egress fees, technical barriers to interoperability/portability, and committed spend discounts can restrict switching or multi-cloud use.	Ofcom cloud market referral, 2023; OECD p. 29	Established finding	Lock-in mechanism	Does not quantify all customer harms
Government AI readiness	Government AI Readiness Index 2025 assesses 195 countries using 69 indicators, 14 dimensions, and six pillars; AI Infrastructure includes compute capacity, technical infrastructure, and data.	Oxford Insights 2025, p. 5, p. 10	Indicative evidence	Preparedness context	Constructed index, not compute ownership
Global compute inequality	HICs hosted 86% of TOP500 HPC systems and 97% of TOP500 capacity as of June 2025.	World Bank 2025, p. 61	Established finding	Global countercapacity gap	TOP500 omits some private AI clusters

Theme	Verified data point / fact	Source and location	Evidence status	Recommended use	Caution
Global server and data-center proxies	In 2024, 50% of global secure internet servers were in the United States, 41% in other HICs, and 9% in the rest of the world; HICs accounted for 77% of global co-location data-center capacity as of June 2025.	World Bank 2025, p. 16, p. 61	Established finding	Global infrastructure inequality	Proxies are not direct AI accelerator ownership
Public compute mitigation	NAIRR reports 14 federal agencies, 28 industry/nonprofit partners, about $100 million in private in-kind contributions, and more than 600 projects across all 50 U.S. states, DC, and Puerto Rico.	NAIRR 2-Year Progress Update, p. 5, p. 8	Policy response / counterargument	Public-compute mitigation	Does not prove sufficiency relative to private frontier compute
European public compute	EuroHPC AI Factories offer access modes including Fast Lane up to 50,000 GPU hours and Large Scale above 50,000 GPU hours; the European Commission reports 19 AI Factories and 13 antennas operational.	EuroHPC access page; European Commission AI Factories page	Policy response / counterargument	Public-compute mitigation	Allocation, queueing, and long-term impact remain open

5.3 Industry, Academia, And The Production Gap

One of the clearest compute-divide concerns is the gap between industry and academia in compute-intensive AI research. Work on the compute divide in machine learning argues that industrial and academic AI labs differ sharply in compute access and that this coincides with reduced academic representation in compute-intensive areas such as foundation models. Universities remain important for methods, theory, critique, evaluation, education, and many forms of AI research. However, for frontier-scale model development and independent reproduction, compute and capital constraints can shift academia toward a more dependent role.

The concern is not that public research disappears. The concern is that some of its traditional functions become harder: independent replication, adversarial scrutiny, open benchmarking, methodological transparency, and training of researchers outside proprietary environments.

Mechanism supported: Production and reproduction countercapacity can lag behind use access. This supports the claim that the compute divide affects who can independently produce and verify frontier-scale AI knowledge.

Boundary: This evidence does not show that universities are losing all AI relevance. It applies most strongly to compute-intensive frontier or foundation-model research.

Evidence status: Indicative evidence / well-supported assumption.

5.4 Cloud Platforms And Infrastructure Dependency

Cloud infrastructure is a central layer of the compute divide because many organizations access compute through cloud platforms rather than direct hardware ownership. Cloud can reduce entry barriers by providing on-demand access to compute. At the same time, large-scale cloud platforms can create lock-in, switching costs, data gravity, preferential access, bundling, and dependency on proprietary tools.

Competition authorities and policy organizations have raised concerns about cloud concentration and the role of partnerships, vertical integration, and platform ecosystems in AI infrastructure. The FTC's work on AI partnerships, OECD analyses of AI infrastructure, and cloud market investigations such as Ofcom/CMA in the UK all point to the importance of market structure. The compute divide is therefore not only a question of absolute scarcity. It is also a question of who controls the gateways through which scarcity is allocated.

Mechanism supported: Cloud can simultaneously lower entry barriers and deepen dependency. This is a feedback mechanism, not a one-way harm claim: cloud access democratizes some AI use while platform concentration can limit switching, bargaining power, independent evaluation, and public procurement autonomy.

Boundary: UK cloud market evidence and competition-authority findings should not be generalized mechanically to every jurisdiction or cloud layer.

Evidence status: Established finding / well-supported assumption.

5.5 Chips, Supply Chains, And Hardware Bottlenecks

Advanced AI depends on specialized hardware, high-bandwidth memory, semiconductor manufacturing, packaging, lithography, networking, and systems integration. AI chips are not interchangeable commodities. Performance, availability, export controls, supply chains, and cloud allocation all shape who can train and deploy advanced systems.

This hardware layer gives the compute divide a geopolitical dimension. Countries and firms that control chip design, fabrication, cloud deployment, or export access can influence the distribution of advanced AI capacity. Conversely, countries without access to advanced chips and cloud infrastructure may be able to adopt AI tools without controlling the underlying production stack.

Mechanism supported: Hardware bottlenecks affect production capacity and strategic bargaining power. They help explain why AI access through APIs or open weights is not the same as independent frontier production capacity.

Boundary: Hardware concentration does not imply that every useful AI system requires frontier chips, nor that algorithmic or architectural shifts cannot reduce dependence.

Evidence status: Established finding / well-supported assumption.

5.6 Energy, Data Centers, And Local Infrastructure

The compute divide is also an energy and land-use issue. Data centers require electricity, cooling, grid interconnections, water in some configurations, land, and local permitting. AI workloads can raise rack density and cooling demands. Energy reports from IEA and LBNL indicate rapid growth in data-center electricity demand, while official national cases such as Ireland show that data centers can become a large share of metered electricity.

Ireland illustrates why data centers should be treated as local infrastructure, not only as abstract cloud capacity. Official CSO data report that Irish data centers used 6,969 GWh in 2024, representing 22% of metered electricity, up from 5% in 2015. Other data-center hubs face different mixes of power, water, land-use, permitting, and sustainability constraints, but Ireland is sufficient for the narrower point made here: AI and cloud infrastructure are materially embedded in local systems.

These cases do not prove that data centers always harm communities. Effects are case-specific. But they show that AI infrastructure can shift costs, grid pressures, planning conflicts, and environmental tradeoffs onto particular places.

Mechanism supported: Compute capacity is physically embedded. Energy and data-center constraints can turn AI infrastructure into a local distributional issue and can constrain who can scale deployment or training.

Boundary: Electricity share and capacity figures do not by themselves prove social harm. Local impact depends on energy mix, grid capacity, water use, permitting, taxation, community benefits, and environmental governance.

Evidence status: Established finding / scenario-based projection.

5.7 Transparency, Model Access, And Evaluation Capacity

Foundation-model transparency is uneven. The Foundation Model Transparency Index reports declining average transparency scores in its 2025 update and continued opacity around training data, training compute, downstream use, and impact. This matters because transparency is not a decorative value. It is a precondition for independent evaluation, public accountability, and scientific reproduction.

If public-interest actors cannot access training information, model documentation, evaluation details, usage patterns, or meaningful testing environments, then their oversight capacity depends on selective disclosure. This is the core of epistemic dependence. A regulator, researcher, journalist, or civil society organization may know that a system matters socially without being able to independently inspect the conditions under which it was built and deployed.

Mechanism supported: Transparency gaps convert infrastructure concentration into epistemic dependence. Without adequate access, public-interest actors may be limited to black-box testing, public claims, developer reports, or negotiated access.

Boundary: Transparency benchmarks do not prove regulatory failure by themselves. They show information asymmetry and auditability constraints.

Evidence status: Established finding / indicative evidence.

5.8 Global Development And AI Readiness

The compute divide has a global development dimension. International sources such as the World Bank, UNCTAD, ILO/UN, and Government AI Readiness Index work point to uneven distributions of digital infrastructure, data capacity, skills, governance readiness, and compute access. While readiness indices are constructed indicators rather than direct measures of compute ownership, they provide useful context: AI capacity is not globally even.

The risk is that some countries become AI rule-takers. They may adopt systems, standards, APIs, cloud services, and governance templates developed elsewhere, while lacking domestic or regional capacity to shape, inspect, adapt, negotiate, or contest them. However, this should not be framed as a passive deficit story. Many actors in the Global South and in smaller economies may pursue sovereignty through adaptation rather than frontier training: fine-tuning, open-weight deployment, regional compute sharing, domain-specific models, public procurement rules, language-localization ecosystems, and sectoral governance capacity.

Two brief examples show why this should be framed as an agency question, not only as a deficit question. IndiaAI's compute-capacity program seeks to make subsidized AI compute available to researchers, startups, MSMEs, students, public-sector agencies, and other approved users through a public-private model. This is not proof of frontier parity, but it is evidence of a public strategy to convert compute scarcity into shared national countercapacity. At the regional level, the African Union's 2024 Continental AI Strategy frames AI around development, responsible governance, local capacity, data governance, skills, and Africa-centered institutional coordination. This does not show that African states have solved infrastructure dependency, but it demonstrates an explicit regional attempt to build strategic agency rather than merely import external AI systems.

Mechanism supported: Global AI inequality is best understood as a countercapacity gap, not simply absence of domestic frontier models.

Boundary: This report is an international scoping report, not a full global mapping. It uses well-documented infrastructure, policy, and market cases rather than claiming exhaustive country coverage.

Evidence status: Established finding / well-supported assumption.

5.9 Case Boxes

Case Box 1: FMTI 2025 As A Transparency Countercapacity Case

The Foundation Model Transparency Index is useful because it turns a broad governance concern into a measurable auditability problem. Its 2025 update reports a decline in average transparency scores from 58 in 2024 to 40 in 2025 and highlights continued opacity around training data, training compute, usage, and impact. The direct evidence is not "governance failure" in general. It is narrower: leading foundation-model developers disclose uneven information about the conditions under which models are built and used.

Mechanism: Limited transparency weakens epistemic countercapacity. Regulators, researchers, journalists, and civil society may be able to observe model outputs or read developer claims, but they face constraints when trying to evaluate training conditions, compute use, data governance, downstream deployment, or societal impact.

Boundary: FMTI is a benchmark with its own design choices. It should support claims about transparency and auditability, not stand alone as proof of democratic failure.

Case Box 2: Ireland As A Local Infrastructure Case

Ireland illustrates why AI infrastructure should not be treated as immaterial cloud capacity. Official CSO data report that data centers used 6,969 GWh of metered electricity in 2024, representing 22% of metered electricity consumption, up from 5% in 2015. This makes Ireland a strong case for the physical and local dimension of AI and cloud infrastructure.

Mechanism: Data-center growth can shift AI infrastructure costs into local electricity systems, grid planning, land-use decisions, water and cooling constraints, taxation debates, and community consultation.

Boundary: Ireland is a hub case, not a universal template. High electricity share does not by itself prove net harm. Local effects depend on energy mix, grid investment, tax arrangements, employment effects, planning rules, and community benefit.

Case Box 3: NAIRR And EuroHPC As Partial Public Countercapacity

Public compute initiatives such as the U.S. National Artificial Intelligence Research Resource and European AI Factories show that governments increasingly understand compute as public-interest infrastructure. Their importance is not that they immediately match private frontier infrastructure. Their importance is that they can provide researchers, public agencies, startups, and public-interest actors with compute access, support, and evaluation environments that would otherwise be unavailable.

Mechanism: Public compute can narrow the countercapacity gap by supporting reproducibility, benchmarking, safety research, public-sector prototyping, and domain-specific public-interest models.

Boundary: Public compute effectiveness is not yet established. Scale, access rules, queueing, staffing, continuity, model access, energy constraints, and connection to regulators and procurement bodies determine whether these programs become genuine countercapacity or only symbolic infrastructure.

---

6. Exploratory Analysis

6.1 Core Mechanism: From Adoption To Countercapacity Gap

The compute divide becomes socially important when AI adoption grows faster than countercapacity. An organization may adopt AI systems for productivity, research, administration, or service delivery. But adoption does not automatically create evaluation capacity. A public agency can procure AI without being able to audit it. A university can use AI tools without being able to reproduce frontier training claims. A smaller state can adopt cloud AI services without controlling compute allocation or model governance.

This creates a structural asymmetry:

- AI developers and infrastructure providers possess more direct knowledge of training, deployment, scaling, and system behavior.
- Public-interest actors often depend on documentation, APIs, regulatory submissions, benchmark results, or negotiated access.
- Affected communities may experience AI-mediated decisions without access to the infrastructure or expertise required to contest them.

The result can arise from ordinary institutional mismatch: fast technological deployment on one side, slower public capacity-building on the other. The same mechanism can be softened by open-weight models, public compute, structured access, stronger procurement rules, standards, and competition remedies.

6.2 Technical Drivers

Several technical drivers intensify the compute divide:

- larger and more capable foundation models;
- multimodal systems that process text, image, audio, video, code, and structured data;
- AI agents that integrate models with tools, search, planning, and workflow automation;
- long-context systems that increase inference costs and enable larger-scale analysis;
- synthetic data and evaluation loops that require iterative compute;
- growing importance of post-training, reinforcement learning, red teaming, and evaluation pipelines;
- secure deployment and monitoring requirements;
- high-performance inference infrastructure for widely used AI services.

Not all technical progress increases compute demand. Algorithmic efficiency, smaller models, distillation, quantization, retrieval, sparsity, on-device inference, specialized chips, and architectural changes can reduce costs for particular tasks. This technological volatility is a real risk to any simple infrastructure-stratification thesis. The compute divide should therefore be formulated around countercapacity rather than GPU scarcity alone: even if useful AI becomes cheaper, independent evaluation, secure deployment, reproduction of frontier claims, regulatory testing, and institutional bargaining power may still remain unevenly distributed.

6.3 Economic Drivers

Economic drivers include capital intensity, cloud business models, economies of scale, data-center financing, chip procurement, energy contracts, and platform lock-in. Frontier AI development requires large upfront investment and sustained operating expenditure. This favors firms and states with access to capital, cloud scale, supply chains, and strategic partnerships.

At the same time, AI infrastructure can reinforce platform power. Cloud providers can bundle compute, storage, data services, AI APIs, development tools, compliance offerings, and enterprise relationships. Customers may face

switching costs, technical dependency, procurement complexity, and data migration risks. These dynamics do not prove monopoly by themselves, but they justify treating cloud concentration as part of the compute divide.

6.4 Institutional Drivers

Institutional drivers include regulatory capacity, public procurement, standards, public compute programs, national AI strategies, and research funding. Institutions can either mitigate or deepen the compute divide.

Public compute initiatives such as NAIRR and EuroHPC AI Factories are important because they recognize compute as a public-interest capacity. They provide access, expertise, and infrastructure that may help researchers, public-sector actors, and smaller innovators. Their effectiveness remains an open question because scale, funding continuity, queueing, support staff, energy constraints, model access, and integration with regulators all matter. A public compute program that is too small, too temporary, or too detached from audit and procurement needs may improve research access without closing the countercapacity gap.

Regulatory institutions also matter. Legal obligations without testing capacity may produce formal compliance without meaningful scrutiny. Technical standards without data access may produce checklists without reproducibility. Audit rights without institutional resources may remain symbolic.

6.5 Social And Political Drivers

Social and political drivers include security logics, national competitiveness, productivity pressure, geopolitical rivalry, and public-sector modernization. Governments may seek sovereign AI capacity for strategic reasons. Firms may race to integrate AI into products and services. Public agencies may adopt AI under pressure to improve efficiency. Universities may adopt AI tools to stay competitive.

These drivers can increase AI adoption before oversight capacity matures. The preparedness gap is partly a timing problem: AI systems are deployed into institutions whose audit, procurement, legal, and technical capacities were not designed for rapidly evolving foundation models.

6.6 Epistemic Dependence

Epistemic dependence is one of the report's central analytical concepts. It describes a condition in which actors depend on knowledge produced, filtered, or mediated by the systems or organizations they seek to evaluate. In the compute-divide context, this can occur when:

- model developers control documentation;
- API providers control access conditions;
- cloud platforms control evaluation environments;
- training data and compute remain undisclosed;
- benchmarks are selected or shaped by developers;
- researchers lack reproduction resources;
- regulators lack independent testing environments;
- civil society lacks legal and technical access.

Epistemic dependence is not identical to ignorance or ordinary information asymmetry. Public actors can still know a great deal. The stronger point is methodological: when the systems used to produce, filter, or operationalize knowledge cannot be independently reproduced or inspected, public knowledge becomes partly dependent on controlled interfaces and selective disclosures. In science, regulation, and journalism, this changes the conditions under which claims can be verified.

The concept should also be distinguished from regulatory capture. Regulatory capture describes situations in which regulated industries shape regulators' behavior, incentives, information flows, or policy agendas in their own interest. Epistemic dependence can contribute to capture, but it is not the same thing. A regulator may be

institutionally independent and acting in good faith while still lacking the infrastructure, access rights, or technical environments needed to verify developer claims independently. Conversely, a captured regulator may have formal access but fail to use it. Epistemic dependence is therefore a capacity and verification problem; regulatory capture is an institutional power and incentive problem.

A narrower formulation is useful for the final report:

> Epistemic dependence exists when an actor has responsibility or reason to evaluate an AI system, but the conditions of evaluation are materially, legally, or technically controlled by the system developer, infrastructure provider, or another actor whose claims are under evaluation.

This definition links the concept to science studies and AI governance without overstating it. It does not imply that all dependence on expertise is illegitimate. Complex societies always rely on expert systems, delegation, and specialized knowledge. The concern here is more specific: the loss of independent verification capacity in domains where AI systems become socially consequential.

The concept is compatible with science-studies work on the social organization of knowledge and co-production of science and social order. Its narrower contribution in this report is to connect those concerns to AI infrastructure: epistemic dependence becomes acute when verification itself requires access to compute, models, data, benchmarks, or deployment environments controlled by the actors under scrutiny.

6.7 Scientific Reproducibility And The Academic Control Function

Science depends on reproducibility, critique, peer review, open methods, and independent verification. Advanced AI complicates this model when key systems cannot be reproduced outside industrial or state-scale infrastructure.

Frontier AI can narrow the range of questions academia can answer independently. Researchers may be able to evaluate outputs but not reconstruct training conditions. They may be able to benchmark APIs but not inspect internal data. They may critique model behavior but not replicate development pipelines.

The careful formulation is:

> Frontier-scale AI can shift universities from independent producers and replicators toward downstream users, evaluators, critics, and partial auditors with constrained access.

Academic contribution can persist through theory, evaluation, interpretability, safety research, model efficiency, social science, law, and public-interest critique. The risk is narrower and more precise: frontier-scale AI may reduce the range of claims academia can verify without external infrastructure access.

6.8 Democratic Contestation

Democratic oversight is not only a matter of law. It requires practical contestation capacity. Civil society organizations, journalists, affected communities, lawyers, and public-interest researchers can investigate algorithmic harms, conduct some forms of black-box testing, litigate, and advocate for accountability. However, meaningful contestation of advanced or high-impact AI systems often requires access to information, technical expertise, legal standing, model or data access, shared vocabularies, and institutional legitimacy that are unevenly distributed.

Evidence from algorithmic governance beyond foundation models suggests that contestation is not merely a legal right or technical design property. It depends on access to information, shared vocabularies, legitimacy for civic intervention, and practical investigative capacity. This is adjacent evidence rather than direct proof of foundation-model audit failure, but it strengthens the concern that democratic audit capacity is itself unequally distributed.

The audit ecosystem literature makes this point more concrete. Raji et al. (2020) frame accountability as an end-to-end audit problem rather than a one-time transparency disclosure, while Raji et al. (2022) emphasize that third-party oversight requires institutional design, access conditions, incentives, and protections for outside auditors. In the terms of this report, those works support the claim that auditability is a countercapacity problem:

disclosure matters, but independent verification requires a functioning ecosystem around the system being audited.

Evidence status: Well-supported assumption / indicative evidence.

6.9 Summary Of Current Analysis

The compute divide is strongest as a layered capacity argument:

1. AI adoption is spreading.
2. Advanced AI infrastructure remains concentrated.
3. Evaluation and reproduction capacity is more limited than use access.
4. Regulatory and democratic oversight require more than legal authority.
5. Public compute and open-weight systems are meaningful counterforces but do not eliminate the deeper infrastructure asymmetry.

---

7. Scenario Development

Scenarios are not predictions. They are analytical tools for clarifying assumptions, risks, counterforces, and governance needs. This report uses a two-axis scenario matrix rather than a timeline or pessimism scale.

The two axes are:

1. **AI infrastructure concentration:** Are compute, cloud, model access, data-center capacity, and evaluation environments concentrated or more distributed?
2. **Public-interest countercapacity:** Do regulators, universities, civil society, smaller firms, and less-resourced countries have meaningful capacity to evaluate, reproduce, negotiate, and contest high-impact AI systems?

	Lower public-interest countercapacity	Higher public-interest countercapacity
Higher infrastructure concentration	Scenario 1: Private Epistemic Dependency	Scenario 2: Contestable Concentration
Lower infrastructure concentration	Scenario 3: Diffuse Capability, Weak Safeguards	Scenario 4: Plural Countercapacity Ecosystem

7.1 Scenario 1: Private Epistemic Dependency

Scenario type: High-risk concentration scenario
Evidence status: Scenario-based projection / speculative claim

In this scenario, AI infrastructure remains highly concentrated while public-interest countercapacity remains weak. Many actors can use AI systems, but the conditions for independent evaluation, reproduction, procurement scrutiny, and democratic contestation remain controlled by a small number of developers, cloud providers, and infrastructure owners.

Mechanism: High concentration plus weak countercapacity produces dependence on private documentation, API access, vendor-provided assurance, and negotiated audit conditions.

Possible harms:

- regulatory review becomes overly dependent on developer disclosures;
- universities can benchmark outputs but not reproduce frontier claims;
- public agencies rely on vendor-provided compliance evidence;
- civil-society contestation is limited to visible outputs and litigation after harm;
- smaller states become rule-takers for externally governed AI infrastructures.

What would reduce this risk: enforceable structured access, public evaluation infrastructure, procurement rules, competition remedies, independent AI safety institutes, and open-weight or smaller-model ecosystems.

7.2 Scenario 2: Contestable Concentration

Scenario type: Regulated-concentration scenario
Evidence status: Policy response / well-supported assumption

In this scenario, AI infrastructure remains concentrated, but public-interest actors gain meaningful countercapacity. Regulators have technical staff, secure testing environments, information rights, and access powers. Universities and public researchers receive targeted compute and model access for evaluation and reproducibility. Civil society and journalists have pathways for evidence access and contestation. Competition authorities scrutinize cloud-AI partnerships, switching costs, and preferential access.

Mechanism: Concentration persists, but its social consequences are moderated because public-interest actors can inspect, challenge, and negotiate the systems they depend on.

Possible benefits:

- stronger regulatory enforcement without requiring full public frontier parity;
- more credible public procurement;
- reduced reliance on private assurance;
- better scientific verification of high-impact claims;
- more meaningful democratic contestation.

Constraints: Public budgets remain smaller than private investment; expertise is hard to retain; audit rights may conflict with privacy, security, or trade-secret claims; international coordination is slow.

7.3 Scenario 3: Diffuse Capability, Weak Safeguards

Scenario type: Open-diffusion risk scenario
Evidence status: Counterargument / scenario-based projection

In this scenario, algorithmic efficiency, open-weight models, smaller domain models, distillation, quantization, retrieval, and on-device deployment substantially reduce some access barriers. Infrastructure concentration matters less for many socially useful tasks. However, public-interest countercapacity remains weak. Capabilities diffuse faster than evaluation norms, procurement standards, civil-society expertise, or safety practices.

Mechanism: Lower concentration improves access and adaptation, but weak countercapacity means that diffusion does not automatically produce accountability.

Possible benefits:

- greater local adaptation;
- stronger language and domain-specific ecosystems;
- less dependence on closed APIs for many tasks;
- more competition at the application layer;
- more opportunities for actors in the Global South to build practical AI sovereignty through adaptation.

Residual risks:

- safety and evaluation capacity may lag behind capability diffusion;
- open weights do not provide training data, audit rights, or secure deployment infrastructure by themselves;
- public agencies may adopt locally adapted systems without meaningful evaluation standards;
- misuse risks may increase where safeguards are weak.

7.4 Scenario 4: Plural Countercapacity Ecosystem

Scenario type: Desirable mitigation scenario
Evidence status: Policy response / open question

In this scenario, infrastructure becomes more plural and public-interest countercapacity grows. Open-weight models, smaller models, regional compute, public compute programs, universities, AI safety institutes, civil-society audit capacity, and competition remedies reinforce one another. No single public actor needs to match frontier private infrastructure alone, but the ecosystem as a whole improves verification, adaptation, procurement, and contestation capacity.

Mechanism: Distributed capability and stronger countercapacity combine. Access improves, and the ability to evaluate, challenge, and govern AI systems improves with it.

Possible benefits:

- robust local and regional adaptation;
- stronger scientific reproducibility for high-impact claims;
- less dependence on vendor-provided assurance;
- more credible public-sector AI procurement;
- better democratic contestability;
- more realistic AI sovereignty for smaller states through shared infrastructure and specialized capability.

Constraints: This is a demanding coordination scenario. It requires sustained funding, shared standards, public-interest expertise, energy planning, institutional trust, and legal access mechanisms. It is desirable, but not guaranteed.

---

8. Risk And Impact Assessment

8.1 Assessment Criteria

Each major risk is assessed using the following criteria:

Criterion	Meaning
Likelihood	How plausible is the risk under present or near-future conditions?
Impact	How severe would the consequences be?
Scalability	How easily can the pattern expand across actors, sectors, or countries?
Accessibility	How low are the barriers for actors to participate in or exploit the pattern?
Detectability	How visible would the risk be to affected people or institutions?
Reversibility	Can harms be repaired or institutions recover capacity?
Affected groups	Who is especially exposed?
Governance gap	Are legal, technical, or institutional safeguards insufficient?
Misuse potential	Is deliberate harmful use plausible?

The scoring is qualitative, not mathematical. It uses five levels:

Score	Meaning
Low	Limited evidence, narrow scope, strong existing safeguards, or low present relevance.
Low-medium	Some plausible concern, but limited current evidence, narrow scope, or strong counterforces.
Medium	Plausible and relevant, but evidence is partial, effects are domain-specific, or counterforces are significant.
Medium-high	Strong concern with meaningful evidence, but still limited by scope, uncertainty, or significant counterforces.
High	Strong current relevance, multiple supporting sources, broad scalability, weak safeguards, or severe potential impact.

Priority is assigned by combining current relevance, likelihood, impact, governance gap, and evidence strength:

- **Priority 1:** high current relevance or high near-future relevance, medium-high to high likelihood, high impact, and a clear governance or countercapacity gap.
- **Priority 2:** plausible and important, but either more domain-specific, less directly evidenced, more mitigable, or less currently widespread.
- **Watchlist:** potentially high-impact but more speculative, more dependent on future developments, or currently supported mainly by scenario logic.

8.2 Priority Risk Matrix

Risk	Current relevance	Likelihood	Impact	Scalability	Accessibility	Detectability	Reversibility	Governance gap	Misuse potential	Evidence strength	Priority rationale
Regulatory preparedness gap	High	High	High	High	Medium	Medium	Medium	High	Medium	Medium-high	Priority 1: AI adoption and legal obligations are already advancing; independent public evaluation capacity is slower to build.

Risk	Current relevance	Likelihood	Impact	Scalability	Accessibility	Detectability	Reversibility	Governance gap	Misuse potential	Evidence strength	Priority rationale
Private infrastructure lock-in	High	High	Medium-high	High	Medium	Medium	Low-medium	High	Low-medium	High	Priority 1: cloud and AI platform concentration are current, scalable, and hard to reverse once procurement and data dependencies deepen.
Global AI countercapacity gap	High	Medium-high	High	High	Medium	Medium	Medium-low	High	Medium	Medium-high	Priority 1: global disparities in infrastructure/readiness are documented, though exact country pathways vary.
Epistemic dependence	Medium-high	Medium-high	High	Medium-high	Medium	Low-medium	Medium-low	High	Medium	Medium	Priority 1: transparency and access gaps are documented; broader social effects remain partly inferential.
Scientific reproducibility gap	Medium	Medium	Medium-high	Medium	Low-medium	Medium	Medium	Medium-high	Low	Medium	Priority 2: strongest for frontier-scale systems and less general for smaller or open systems.
Democratic audit and contestation gap	Medium	Medium	High	Medium	Medium	Low	Medium-low	High	Medium	Medium-low	Priority 2: normatively important, but direct foundation-model case evidence is still limited.
Local data-center externalities	Medium	Medium	Medium-high	Medium	Low	High	Medium	Medium	Low	Medium-high	Priority 2: energy and capacity data are strong, but local harms are case-specific.
Capability diffusion without symmetric safeguards	Emerging	Medium	High	High	Medium-high	Medium	Medium-low	Medium-high	Medium-high	Medium-low	Watchlist: plausible under open and efficient systems, but pathway depends on task, safeguards, and actor behavior.
Durable systemic epistemic dependency	Low-medium current, high future	Low-medium	High	Medium-high	Low-medium	Low	Low	High	Medium	Low-medium	Watchlist: high-impact warning scenario, but current evidence is indirect and scenario-based.

8.3 Regulatory Preparedness Gap

The regulatory preparedness gap is the most important present and near-future risk. It occurs when regulators are expected to govern AI systems but lack the technical staff, compute access, testing infrastructure, documentation, and institutional support required for independent evaluation.

This risk is likely because advanced AI systems are already being integrated into economic and public life, while public-sector capacity-building takes time. It has high impact because regulation without independent evaluation can become dependent on vendor claims, private audits, or limited documentation.

Affected groups include public agencies, regulators, courts, workers, consumers, students, patients, citizens, and communities affected by AI-mediated decisions. The governance gap is clear: laws and standards can mandate duties, but enforcement requires capacity.

Priority rationale: Priority 1 because likelihood, impact, scalability, and governance gap are all high. The evidence base is strongest for the existence of capacity asymmetry; the exact institutional consequences vary by regulator and sector.

8.4 Epistemic Dependence

Epistemic dependence is a deeper structural risk. It arises when public-interest actors must rely on the information and interfaces provided by the actors they seek to scrutinize. This can occur through API-only access, opaque training data, undisclosed compute, proprietary evaluation pipelines, restricted audit environments, or selective transparency.

The risk is difficult to detect because it can look like normal technical complexity. A system may be documented enough for procurement, but not enough for independent reproduction. A regulator may receive reports, but not enough access for verification. A researcher may test outputs, but not inspect training conditions.

Priority rationale: Priority 1 because detectability is low, reversibility is limited once institutional routines depend on controlled interfaces, and the governance gap is high. The claim should remain partly inferential: transparency gaps are directly evidenced, while systemic epistemic dependence is a broader implication.

8.5 Private Infrastructure Lock-In

Private infrastructure lock-in occurs when organizations become dependent on a small set of cloud, model, and tooling providers. Lock-in can arise from technical integration, data gravity, contractual terms, proprietary APIs, compliance tooling, pricing, and availability of specialized AI hardware.

This risk is already relevant. It does not require malicious behavior. It can emerge from ordinary economic incentives and platform dynamics. The consequence is reduced contestability and weaker public bargaining power.

Priority rationale: Priority 1 because lock-in is current, scalable, and difficult to reverse. The strongest evidence concerns cloud concentration and switching barriers; the AI-specific lock-in pathway should be tied to platform integration and procurement evidence.

8.6 Global AI Countercapacity Gap

The global AI countercapacity gap refers to the possibility that many countries adopt AI systems without developing the capacity to shape, inspect, adapt, negotiate, or govern their underlying infrastructures. This could deepen global inequality by separating AI users from AI shapers.

This risk is supported by global disparities in compute, digital infrastructure, government readiness, data capacity, and industrial capability. It should be formulated carefully. Not every country needs to train frontier models domestically, and regional cooperation, open-weight adaptation, public procurement capacity, local fine-tuning ecosystems, and public compute can mitigate dependence. The risk is not absence of domestic frontier labs alone. The risk is lack of meaningful capacity to evaluate, adapt, negotiate, and govern AI systems used domestically.

Priority rationale: Priority 1 because current disparities are well documented and the impact of weak countercapacity can be high. The pathway is not deterministic because adaptation strategies can produce meaningful local sovereignty without frontier training.

8.7 Scientific Reproducibility Gap

The scientific reproducibility gap occurs when AI research claims, model capabilities, or safety evaluations cannot be independently reproduced because the required compute, data, model access, or training details are unavailable. This risk is especially relevant for frontier-scale systems.

The consequence is not that AI research stops being science. It is that parts of frontier AI may shift toward a model of restricted verification, where external actors can test outputs but cannot reproduce underlying development conditions.

Priority rationale: Priority 2 because the risk is strongest for frontier-scale systems and safety claims. It is less severe where open models, smaller systems, or reproducible methods are available.

8.8 Democratic Audit And Contestation Gap

Democratic contestation requires more than transparency statements. It requires legal standing, access to information, technical expertise, investigative capacity, shared vocabularies, and institutional legitimacy. Civil society and journalists can investigate algorithmic systems, but advanced AI creates new asymmetries in model access, data access, and technical complexity.

This risk is important but should remain partly indicative. The strongest direct evidence comes from broader algorithmic governance and audit ecosystem literature, not exclusively from foundation-model cases.

Raji et al.'s work on internal algorithmic auditing and third-party audit ecosystems is especially relevant here because it shows that democratic audit is not reducible to publication of model cards, benchmarks, or policy statements. It requires durable access pathways, audit methods, protected external scrutiny, and institutions capable of acting on audit findings.

Priority rationale: Priority 2 because the democratic stakes are high, but the direct evidence base for foundation-model contestation is still thinner than for cloud, energy, or transparency gaps.

8.9 Local Data-Center Externalities

Data-center growth can create local planning, energy, water, land-use, grid, fiscal, and distributional tensions. Ireland shows that data centers can become a large share of national metered electricity. U.S. sources show regional load-growth and community-impact concerns. The broader point is not that every data center produces the same local effects, but that AI infrastructure requires local governance rather than only abstract cloud policy.

The careful claim is that data-center growth can produce local externalities, not that it always does. Impacts vary by energy mix, grid capacity, water use, tax policy, land use, community consultation, and local economic benefits.

Priority rationale: Priority 2 because data-center growth is well documented, but harms and benefits are local, negotiated, and dependent on energy, water, tax, planning, and community contexts.

8.10 Risk Prioritization

The report should prioritize risks as follows:

Category	Risks	Treatment
Already relevant	regulatory preparedness gap; cloud lock-in; global countercapacity gap; data-center energy pressure	Main risk assessment
Plausible and emerging	epistemic dependence; scientific reproducibility gap; democratic contestation gap	Conditional near-future analysis
Speculative / high-impact	durable systemic epistemic dependency; dependency trap for smaller states; broad risky capability diffusion without safeguards	Scenario and limitations only

---

9. Counterarguments And Limitations

9.1 Counterargument: Compute Is Becoming Cheaper

Compute costs can fall through hardware improvements, software efficiency, smaller models, quantization, distillation, sparsity, retrieval, better training recipes, and optimized inference. Many useful AI capabilities do not require frontier-scale training. This is a strong counterargument against a simplistic claim that only the largest actors can use powerful AI.

However, falling costs do not automatically eliminate the compute divide. The frontier may continue to move. Evaluation, deployment, security, and governance can remain resource-intensive even when use access becomes cheaper. Cost decline can democratize some capabilities while leaving frontier production and independent reproduction concentrated.

Revised conclusion: Efficiency mitigates the compute divide but does not dissolve it.

9.2 Counterargument: Open-Weight Models Reduce Dependence

Open-weight models are a major counterforce. They can enable research, adaptation, local deployment, competition, transparency experiments, educational use, and public-interest evaluation. They challenge the idea that AI capability is exclusively controlled by closed labs.

At the same time, open-weight access is not equivalent to full infrastructure equality. Open weights do not necessarily provide training data, training compute, safety evaluations, data governance, deployment infrastructure, fine-tuning resources, or secure large-scale operation. Benchmark evidence is also time-sensitive. Stanford's 2026 AI Index reports that the top closed model led the top open model by 3.3% on the Arena Leaderboard as of March 2026, compared with a smaller gap in August 2024. The relevant conclusion is not that open models have failed. It is that openness reduces some barriers while frontier parity remains benchmark-dependent and unstable.

Revised conclusion: Open-weight models substantially mitigate the compute divide but do not automatically equalize frontier training, reproduction, deployment, or governance capacity.

9.3 Counterargument: Public Compute Can Close The Gap

Public compute initiatives such as NAIRR and EuroHPC AI Factories are important and should be treated as serious mitigation strategies. They show that governments recognize compute as public infrastructure. They can support researchers, startups, public agencies, and public-interest evaluation.

The limitation is that the effectiveness of public compute remains open. Scale, access rules, staffing, usability, continuity, model access, energy constraints, and integration with regulatory needs all matter. Public compute can become a real countercapacity, but it is not automatically sufficient.

Revised conclusion: Public compute is a central policy response and an open empirical question.

9.4 Counterargument: Regulation Can Mandate Transparency

Regulation can mandate documentation, risk management, transparency, audits, incident reporting, and access. The EU AI Act, U.S. policy debates, NIST AI RMF, NTIA accountability work, and international standards all show growing governance attention.

The compute-divide concern is that legal mandates require enforcement capacity. A law can create duties, but public institutions need technical expertise, compute, model access, data access, audit rights, and institutional independence to verify compliance.

Revised conclusion: Regulation is necessary but insufficient without infrastructural and technical capacity.

9.5 Counterargument: Not All AI Is Frontier AI

Many socially useful AI applications use smaller models, domain-specific systems, classical machine learning, or local deployment. Treating all AI as frontier-scale AI would exaggerate the compute divide.

This is true. The report should avoid generalizing from frontier systems to all AI. Its core claim applies most strongly to advanced AI systems, foundation models, high-impact AI deployments, and institutions that require independent evaluation of systems they do not fully control.

Revised conclusion: The compute divide is strongest for frontier, foundation, and high-impact AI, not for every AI application.

9.6 Counterargument: Market Competition May Reduce Concentration

Market competition can reduce costs, increase model access, diversify providers, improve tooling, and support open ecosystems. Startups and open communities can innovate in ways that incumbents do not control.

The limitation is that infrastructure layers may remain concentrated even when application layers are competitive. Competition at the app layer does not eliminate dependence on chips, cloud, model APIs, data centers, and energy.

Revised conclusion: Competition can mitigate the compute divide, but competition must be analyzed layer by layer.

9.7 Main Limitations Of The Report

This report has several limitations:

- It is exploratory, not systematic.
- It synthesizes existing sources rather than collecting original empirical data.
- Some quantitative claims rely on institutional reports and projections.
- Some sources are preprints or policy analyses.
- AI capability benchmarks change quickly.
- Cloud and compute market data can be proprietary.
- Global comparisons are limited by data availability.
- The democratic contestation argument relies partly on adjacent algorithmic governance evidence.
- The report does not model exact causal pathways from compute concentration to social outcomes.

These limitations should be retained in the final version. They strengthen the report by preventing overclaiming.

---

10. Implications

The implication of the compute divide is not simply "spend more on public AI." Public countercapacity is expensive, politically contested, and difficult to scale. Private AI investment, cloud infrastructure, chip procurement, and data-center expansion can exceed the resources available to many public institutions. The policy problem is therefore strategic allocation: public actors need enough countercapacity to verify, contest, negotiate, and govern high-impact systems, not necessarily to duplicate every private frontier capability.

10.1 Implications For Science

The compute divide challenges the conditions of independent science. If frontier-scale AI systems require resources that only a few firms and states can access, then independent reproduction becomes harder. Researchers may still evaluate outputs, study impacts, develop smaller models, propose methods, and critique claims. But the ability to reproduce frontier training and deployment conditions may be restricted.

This has implications for peer review, scientific credibility, AI safety research, and public trust. Scientific independence requires more than publication access. It requires access to the infrastructures needed for verification.

Realistic policy implication: Public research compute should be targeted toward reproducibility, evaluation, benchmark infrastructure, and safety-relevant research where public verification matters most. It does not need to replicate every frontier training run to be valuable. Structured model access, reproducibility grants, shared evaluation environments, and independent benchmark infrastructure may be more feasible than permanent public parity with the largest private training clusters.

10.2 Implications For Regulation

AI regulation must be understood as a capacity problem. Legal frameworks need technical enforcement capacity. Regulators require staff, compute access, secure testing environments, model documentation, data access where lawful, procurement expertise, and the ability to commission independent audits.

Without this capacity, regulation risks becoming documentation review rather than technical scrutiny. The preparedness gap is therefore one of the central near-term policy concerns.

Realistic policy implication: Regulators cannot compete with leading AI firms by salary, hardware scale, or internal engineering depth alone. They need leverage. This can include legally enforceable information rights, protected researcher access, secure testing facilities, shared expert pools, mandatory incident reporting, public procurement conditions, and the ability to commission independent technical audits. Regulation should therefore combine legal authority with practical access and institutional bargaining power.

10.3 Implications For Transparency, Auditability, And Verification

Transparency is a central governance issue in the compute divide, but it should not be treated as a simple disclosure checklist. Transparency matters because it can reduce epistemic dependence: it gives regulators, researchers, affected communities, journalists, courts, and civil society a basis for understanding how AI systems are built, evaluated, deployed, monitored, and contested. However, transparency alone is not the same as countercapacity. A public report can inform external actors without enabling them to verify the claims it contains.

The relevant distinction is between **disclosure**, **auditability**, and **verification**. Disclosure means that developers or infrastructure providers publish information, such as model cards, system cards, risk reports, data summaries, benchmark results, incident reports, or energy and compute estimates. Auditability means that qualified external actors have enough access, documentation, tooling, legal authority, and procedural protection to

examine claims in a structured way. Verification means that some claims can be independently tested, reproduced, challenged, or compared against external evidence.

This distinction is important because the most consequential gaps identified by transparency benchmarks concern upstream and downstream information: training data, training compute, compute providers, model-development costs, evaluation methods, usage data, post-deployment monitoring, and societal impact. These are precisely the areas where outside actors need information to evaluate whether a system is safe, lawful, fair, secure, environmentally credible, or socially appropriate. If transparency remains selective, non-standardized, or non-verifiable, it can become a form of controlled visibility rather than public accountability.

Transparency policy should therefore be layered. Basic public transparency can inform users, researchers, journalists, and affected communities. Regulator-facing transparency can include more detailed information under confidentiality protections. Researcher access can support independent evaluation, red-teaming, reproducibility, and impact studies. Procurement transparency can help public agencies avoid vendor lock-in and assess lifecycle costs. Infrastructure transparency can include energy demand, water use, data-center siting, cloud dependencies, and compute allocation where disclosure is feasible and lawful.

The same approach should recognize limits. Some information may be restricted for privacy, security, intellectual property, or misuse-prevention reasons. The answer is not necessarily full public disclosure of every model detail. The more realistic goal is accountable opacity: where information cannot be fully public, there should be trusted access pathways, independent audit mechanisms, legal safeguards, and institutional checks that prevent secrecy from becoming unchallengeable dependency.

Realistic policy implication: Transparency obligations should be designed as part of a verification ecosystem. Useful measures include standardized reporting on training data and compute, external researcher access, protected audit channels, incident and post-deployment reporting, public-sector procurement disclosure, regulator access to non-public documentation, independent benchmark infrastructure, and transparency about energy and data-center impacts. The quality test is not whether information exists somewhere, but whether public-interest actors can use it to verify, challenge, and govern high-impact AI systems.

10.4 Implications For Democracy

Democratic oversight requires contestability. Citizens, journalists, civil society organizations, affected communities, and public-interest researchers need practical pathways to challenge AI systems that affect them. If contestation depends on access to technical expertise, model documentation, evaluation environments, or legal standing that only a few actors possess, then formal accountability may be weaker than it appears.

Democratic audit capacity should therefore be built as an institutional ecosystem rather than as a single transparency rule. Structured access, public-interest audit rights, whistleblower protections, litigation support, journalist and civil-society technical capacity, and enforceable explanation duties are complementary. They also require funding and legal standing. A right to contest a system is weak if affected actors cannot obtain evidence, expertise, or a venue in which contestation matters.

10.5 Implications For Markets And Competition

AI infrastructure concentration can affect market contestability. Cloud platforms, chip access, model ecosystems, data services, and enterprise integrations can create dependencies that shape who can compete. Competition policy therefore matters for the compute divide.

However, market competition should be assessed across layers. A competitive app market may still depend on concentrated cloud and chip infrastructure. Conversely, open model competition may reduce some forms of dependence while leaving deployment and evaluation infrastructure concentrated.

Realistic policy implication: Competition policy may be more scalable than trying to match private compute budgets with public spending alone. Relevant tools include scrutiny of exclusive cloud-AI partnerships,

interoperability and portability duties, limits on self-preferencing, procurement rules against avoidable lock-in, merger review, cloud switching remedies, transparency around preferential compute access, and, where justified, structural separation debates between infrastructure control and downstream AI services. These tools should be evaluated carefully, because blunt intervention can also reduce investment or security. But the report should not treat public compute as the only countermeasure.

10.6 Implications For Global Development

The global compute divide may create new forms of dependency. Countries can adopt AI tools while lacking domestic or regional capacity to evaluate, adapt, govern, or negotiate the infrastructures behind them. This could reinforce unequal positions in global digital development.

The policy question is not whether every country must build frontier models. It is whether countries and regions have enough countercapacity to make informed decisions, protect rights, adapt systems to local needs, negotiate procurement, and participate in international AI governance. For many countries, meaningful sovereignty may come from open-weight adaptation, regional compute access, language and domain specialization, procurement competence, and regulatory capacity rather than domestic frontier training.

IndiaAI and the African Union's Continental AI Strategy illustrate two different agency pathways. IndiaAI emphasizes shared compute access and a national public-private compute pool; the AU strategy emphasizes regional coordination, responsible governance, skills, data governance, and Africa-centered development priorities. Both should be treated as partial countercapacity strategies. They reduce the danger of portraying less-resourced regions as passive, while also showing why implementation capacity, infrastructure scale, financing, and institutional continuity remain central.

Realistic policy implication: International development strategies should support regional AI infrastructure, shared public compute, local language resources, data governance, regulatory training, procurement capacity, and South-South or regional cooperation. The goal should be practical countercapacity rather than symbolic national frontier-model races.

10.7 Implications For Local Communities

Data centers are physical infrastructures. They affect electricity systems, water planning, land use, local tax arrangements, jobs, community consultation, and environmental tradeoffs. The benefits of AI infrastructure may be distributed globally, while costs may be concentrated locally.

AI infrastructure planning should therefore include local accountability and distributional analysis. Data-center permitting and energy planning should include transparency about electricity demand, water use, grid impacts, community benefits, environmental tradeoffs, tax arrangements, and long-term local obligations. Where local communities host infrastructure that supports global AI services, benefit-sharing and grid-planning questions become part of AI governance.

10.8 Implications For Public Compute

Public compute is one of the most important mitigations. It can provide researchers, public agencies, startups, and civil society with resources that reduce dependence on private infrastructure. But public compute must be treated as more than hardware access. It requires governance, staffing, user support, security, energy planning, evaluation environments, model access, and integration with public-interest goals.

Public compute has limits. It may be smaller than private frontier infrastructure, politically vulnerable, oversubscribed, difficult to staff, and constrained by energy or procurement cycles. Its value should therefore be judged by strategic function rather than by parity with hyperscalers. The most realistic functions are independent evaluation, reproducibility, safety research, public-sector prototyping, open benchmarking, domain-specific

public-interest models, and support for researchers or smaller actors who would otherwise have no meaningful access.

Realistic policy implication: Public compute should be designed as a countercapacity system, not only as a pool of machines. It should be linked to regulators, universities, public procurement bodies, civil-society audit programs, and competition authorities.

10.9 Implications For Public-Private Capacity Asymmetry

Budget asymmetry is not a side issue. It is one of the central political constraints of the compute divide. A single frontier training effort, long-term cloud contract, or large data-center buildout can exceed the resources available to many public research or regulatory programs. This makes naive parity unrealistic.

Public strategy should therefore distinguish three levels of ambition:

Level	Goal	Realistic function
Access support	Help researchers, public agencies, startups, and civil society use compute	Reduces exclusion from basic experimentation and adaptation
Evaluation countercapacity	Enable independent testing, reproducibility, benchmarking, and safety evaluation	Directly addresses preparedness and epistemic dependence
Strategic infrastructure	Build or coordinate large-scale public, regional, or sovereign AI infrastructure	Reduces dependency but requires sustained funding, energy planning, and governance

The middle layer, evaluation countercapacity, may be the highest-return public investment because it does not require full frontier parity but directly improves governance, science, and democratic oversight.

10.10 Key Indicators To Monitor

Future versions of this report should monitor:

- share of notable models by actor type;
- training compute requirements;
- cloud market concentration;
- cloud switching barriers;
- advanced chip and foundry concentration;
- data-center electricity demand;
- regional data-center capacity;
- transparency scores for training data and training compute;
- external researcher access and protected audit pathways;
- incident reporting and post-deployment monitoring practices;
- regulator access to non-public documentation and secure testing environments;
- academic versus industry foundation-model publication patterns;
- public compute program scale and access conditions;
- regulator staffing, compute access, and technical evaluation capacity;
- cross-country AI infrastructure and readiness indicators.

---

11. Conclusion

The compute divide is an emerging form of technological stratification, but its sharpest formulation is the gap between AI access and AI countercapacity. It is not simply unequal access to AI tools. It is unequal practical capacity to subject advanced AI systems to independent public-interest scrutiny.

The evidence base is strongest for the infrastructural foundation of the claim: advanced AI depends on concentrated resources including compute, cloud platforms, advanced chips, data centers, electricity, capital, and specialized expertise. It is also strong for transparency gaps, cloud dependency, energy growth, and global disparities in readiness and infrastructure. The more conditional implications concern epistemic dependence, weakened scientific reproduction, regulatory preparedness gaps, democratic contestation gaps, and durable global countercapacity gaps.

The report's central conclusion is conditional rather than deterministic. Open-weight models, smaller models, efficiency gains, public compute, structured access, transparency duties, auditability mechanisms, competition policy, regulatory expertise, and civil-society capacity can all mitigate the compute divide. However, these counterforces must be assessed layer by layer. A countermeasure that improves use access may not improve independent verification. A transparency report may not create an audit ecosystem. A competitive application market may still depend on concentrated cloud and chip infrastructure.

A central policy problem is the gap between AI adoption and public-interest oversight capacity. The most realistic response is not a single demand for public frontier parity, but a three-level strategy: access support for experimentation and adaptation, evaluation countercapacity for independent testing and audit, and strategic infrastructure where sustained public or regional coordination is justified. This is how public policy can respond to budget asymmetry without pretending that all actors can or should duplicate private frontier infrastructure.

If advanced AI systems become part of the infrastructure of knowledge production, then democratic societies must ask who shapes them, who verifies them, who bears their costs, and what institutional routes exist for challenge. The compute divide names this problem. The next task is to build enough public, scientific, regulatory, and democratic countercapacity to prevent AI access from becoming a substitute for AI accountability.

References

African Union. (2024). *Continental Artificial Intelligence Strategy: Harnessing AI for Africa's Development and Prosperity*. https://au.int/en/documents/20240809/continental-artificial-intelligence-strategy

Amnesty International. (2025). *Algorithmic Accountability Toolkit*. Amnesty International. https://www.amnesty.org/en/latest/research/2025/12/algorithmic-accountability-toolkit/

Besiroglu, T., Bergerson, S. A., Michael, A., Heim, L., Luo, X., & Thompson, N. (2024). *The Compute Divide in Machine Learning: A Threat to Academic Contribution and Scrutiny?* arXiv. https://doi.org/10.48550/arXiv.2401.02452

Bourdieu, P. (1986). The forms of capital. In J. G. Richardson (Ed.), *Handbook of Theory and Research for the Sociology of Education* (pp. 241-258). Greenwood. https://www.bloomsbury.com/us/handbook-of-theory-and-research-for-the-sociology-of-education-9780313235290/

Carpenter, D., & Moss, D. A. (Eds.). (2014). *Preventing Regulatory Capture: Special Interest Influence and How to Limit It*. Cambridge University Press. https://doi.org/10.1017/CBO9781139565875

Central Statistics Office Ireland. (2025). *Data Centres Metered Electricity Consumption 2024*. https://www.cso.ie/en/releasesandpublications/ep/p-dcmec/datacentresmeteredelectricityconsumption2024/

Competition and Markets Authority. (2025). *Cloud Services Market Investigation*. https://www.gov.uk/cma-cases/cloud-services-market-investigation

Cihon, P. (2019). *Standards for AI Governance: International Standards to Enable Global Coordination in AI Research and Development*. Future of Humanity Institute, University of Oxford. https://www.governance.ai/research-paper/standards-for-ai-governance-international-standards-to-enable-global-coordination-in-ai-research-development

Center for Security and Emerging Technology. (2020). *AI Chips: What They Are and Why They Matter*. https://cset.georgetown.edu/publication/ai-chips-what-they-are-and-why-they-matter/

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Governance of AI Program, Future of Humanity Institute, University of Oxford. https://www.governance.ai/research-paper/agenda

DiMaggio, P., Hargittai, E., Celeste, C., & Shafer, S. (2003). *From Unequal Access to Differentiated Use: A Literature Review and Agenda for Research on Digital Inequality*. Princeton University. https://ideas.repec.org/p/pri/cpanda/29.html

Eneva, S. A., & Mora-Gamez, F. (2025). Civil society and the (im)possibilities of algorithmic contestation in migration governance. *Frontiers in Sociology, 10*, 1641898. https://doi.org/10.3389/fsoc.2025.1641898

Epoch AI. (2026). *Trends in Artificial Intelligence*. https://epoch.ai/trends

European Commission. (2025). *The General-Purpose AI Code of Practice*. https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai

European Commission. (2026). *AI Office*. https://digital-strategy.ec.europa.eu/en/policies/ai-office

European Commission / DG CNECT. (2026). *AI Factories*. https://digital-strategy.ec.europa.eu/en/policies/ai-factories

EuroHPC Joint Undertaking. (2026). *AI Factories Access Modes*. https://www.eurohpc-ju.europa.eu/ai-factories/ai-factories-access-modes_en

Federal Trade Commission. (2025). *FTC Staff Report on AI Partnerships & Investments 6(b) Study*. https://www.ftc.gov/reports/ftc-staff-report-ai-partnerships-investments-6b-study

Hartmann, D., Laranjeira de Pereira, J. R., Streitboerger, C., & Berendt, B. (2024). Addressing the regulatory gap: Moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. *AI and Ethics*. https://doi.org/10.1007/s43681-024-00595-3

International Energy Agency. (2025). *Energy and AI*. https://www.iea.org/reports/energy-and-ai

ILO and United Nations. (2024). *Mind the AI Divide: Shaping a Global Perspective on the Future of Work*. International Labour Organization and United Nations. https://www.ilo.org/publications/mind-ai-divide-shaping-global-perspective-future-work

IndiaAI. (2026). *IndiaAI Compute Capacity*. https://indiaai.gov.in/hub/indiaai-compute-capacity

Jasanoff, S. (Ed.). (2004). *States of Knowledge: The Co-Production of Science and Social Order*. Routledge. https://www.routledge.com/States-of-Knowledge-The-Co-production-of-Science-and-the-Social-Order/Jasanoff/p/book/9780415403290

Lawrence Berkeley National Laboratory and U.S. Department of Energy. (2024). *2024 United States Data Center Energy Usage Report*. https://eta-publications.lbl.gov/sites/default/files/2024-12/lbnl-2024-united-states-data-center-energy-usage-report.pdf

Lehdonvirta, V., Wu, B., & Hawkins, Z. (2024). Compute North vs. Compute South: The uneven possibilities of compute-based AI governance around the globe. *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. https://doi.org/10.1609/aies.v7i1.31683

Luitse, D. (2024). Platform power in AI: The evolution of cloud infrastructures in the political economy of artificial intelligence. *Internet Policy Review, 13*(2). https://doi.org/10.14763/2024.2.1768

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/S/bo28451565.html

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

National Science Foundation. (2026). *NAIRR Pilot 2-Year Progress Update*. https://nsf-gov-resources.nsf.gov/files/NAIRR-2-Year-Progress-Update.pdf

National Telecommunications and Information Administration. (2024). *AI Accountability Policy Report*. https://www.ntia.gov/sites/default/files/publications/ntia-ai-report-final.pdf

National Telecommunications and Information Administration. (2024). *Dual-Use Foundation Models With Widely Available Model Weights*. https://www.ntia.gov/sites/default/files/publications/ntia-ai-open-model-report.pdf

OECD. (2025). *Competition in Artificial Intelligence Infrastructure*. https://doi.org/10.1787/623d1874-en

OECD. (2025). *Competition in the Provision of Cloud Computing Services*. https://one.oecd.org/document/DAF/COMP%282025%298/en/pdf

OECD. (2025). *Governing with Artificial Intelligence*. Organisation for Economic Co-operation and Development. https://www.oecd.org/en/publications/2025/06/governing-with-artificial-intelligence_398fa287.html

Ofcom. (2023). *Cloud Services Market Study: Final Report*. https://www.ofcom.org.uk/internet-based-services/cloud-services/cloud-services-market-study/

Ofcom. (2023). *Ofcom Refers UK Cloud Market to CMA for Investigation*. https://www.ofcom.org.uk/internet-based-services/cloud-services/ofcom-refers-uk-cloud-market-to-cma-for-investigation

Oxford Insights. (2025). *Government AI Readiness Index 2025*. https://oxfordinsights.com/ai-readiness/government-ai-readiness-index-2025/

Plantin, J.-C., Lagoze, C., Edwards, P. N., & Sandvig, C. (2018). Infrastructure studies meet platform studies in the age of Google and Facebook. *New Media & Society, 20*(1), 293-310. https://doi.org/10.1177/1461444816661553

Pilz, K. F., Heim, L., & Brown, N. (2025). Increased compute efficiency and the diffusion of AI capabilities. *Proceedings of the AAAI Conference on Artificial Intelligence, 39*(26), 27582-27590. https://doi.org/10.1609/aaai.v39i26.34971

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 33-44). ACM. https://doi.org/10.1145/3351095.3372873

Raji, I. D., Xu, P., Honigsberg, C., & Ho, D. E. (2022). Outsider oversight: Designing a third party audit ecosystem for AI governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 557-571). ACM. https://doi.org/10.1145/3514094.3534181

Royal Society. (2024). *Science in the Age of AI: How Artificial Intelligence Is Changing the Nature and Method of Scientific Research*. The Royal Society. https://royalsociety.org/news-resources/projects/science-in-the-age-of-ai/

Shapin, S. (1994). *A Social History of Truth: Civility and Science in Seventeenth-Century England*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/S/bo3626633.html

Srnicek, N. (2017). *Platform Capitalism*. Polity. https://www.politybooks.com/bookdetail?book_slug=platform-capitalism--9781509504862

Stanford Institute for Human-Centered Artificial Intelligence. (2026). *The 2026 AI Index Report*. Stanford HAI. https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf

Stigler, G. J. (1971). The theory of economic regulation. *The Bell Journal of Economics and Management Science, 2*(1), 3-21. https://www.jstor.org/stable/3003160

UNCTAD. (2025). *Technology and Innovation Report 2025: Inclusive Artificial Intelligence for Development*. United Nations Conference on Trade and Development. https://unctad.org/publication/technology-and-innovation-report-2025

UNU-INWEH. (2026). *The Environmental Cost of Artificial Intelligence: Carbon, Water, and Land Footprints*. United Nations University Institute for Water, Environment and Health. https://unu.edu/inweh/collection/environmental-cost-of-AIs-Enrgy-Use-Carbon-water-and-land-footprints

van Dijk, J. (2020). *The Digital Divide*. Polity. https://www.politybooks.com/bookdetail?book_slug=the-digital-divide--9781509534449

van Dijck, J., Poell, T., & de Waal, M. (2018). *The Platform Society: Public Values in a Connective World*. Oxford University Press. https://doi.org/10.1093/oso/9780190889760.001.0001

Wan, A., Klyman, K., Kapoor, S., Maslej, N., Longpre, S., Xiong, B., Liang, P., & Bommasani, R. (2025). *The 2025 Foundation Model Transparency Index*. arXiv. https://doi.org/10.48550/arXiv.2512.10169

Warschauer, M. (2003). *Technology and Social Inclusion: Rethinking the Digital Divide*. MIT Press. https://mitpress.mit.edu/9780262731737/technology-and-social-inclusion/

World Bank. (2025). *Digital Progress and Trends Report 2025: Strengthening AI Foundations*. https://doi.org/10.1596/978-1-4648-2264-3

World Bank. (2025). *Beyond the AI Divide: Building an AI-Ready Global Workforce*. World Bank. https://openknowledge.worldbank.org/server/api/core/bitstreams/82fbc048-b723-4818-bb91-9a4c8855daf1/content

Appendix A: Core Evidence Chain Matrix

This appendix condenses the report's main claims into evidence chains. The purpose is to make the evidentiary logic transparent: each claim is tied to source anchors, a mechanism, an evidence label, and an explicit boundary.

Main claim	Core source anchors	Mechanism supported	What this does not prove	Evidence label
Advanced AI depends on a concentrated infrastructure stack.	Stanford AI Index 2026, p. 15, p. 32-33; OECD AI infrastructure, p. 8, p. 24, p. 29; IEA, p. 49, p. 63; LBNL / DOE, p. 6-7, p. 52; World Bank, p. 61	Frontier-scale AI relies on compute, cloud, chips, data centers, energy, capital, supply chains, and specialized expertise that are unevenly distributed.	Not every AI application requires frontier-scale infrastructure.	Established finding / well-supported assumption
The compute divide is an access-countercapacity gap.	World Bank, p. 11, p. 16, p. 61; FMTI 2025, p. 1, p. 19, p. 21; Stanford AI Index 2026, p. 19, p. 72; NAIRR, p. 5, p. 8; EuroHPC access materials	AI use can spread while the capacity to evaluate, reproduce, regulate, negotiate, and contest AI systems remains concentrated.	The term "countercapacity" is this report's analytical synthesis, not a direct metric in all sources.	Conceptual claim / well-supported assumption
Transparency gaps can create epistemic dependence.	FMTI 2025, p. 1, p. 19, p. 21; Stanford AI Index 2026, p. 13, p. 15; Ofcom cloud referral; CMA cloud investigation	External actors may depend on developer disclosures, APIs, documentation, and controlled environments when training data, compute, and evaluation details are opaque.	Does not prove intentional deception, total opacity, or democratic failure by itself.	Well-supported assumption / indicative evidence
Universities and public research face constrained frontier-scale control functions.	Stanford AI Index 2026, p. 19; FMTI 2025, p. 21; NAIRR, p. 5, p. 8-9; EuroHPC AI for Science access; IEA/LBNL energy data	Frontier-scale reproduction and independent verification can exceed many academic resource envelopes, even when academic critique and smaller-scale research remain strong.	Does not mean universities are irrelevant or unable to contribute to AI research.	Indicative evidence / scenario-based projection
Regulators face a preparedness gap when mandates outpace evaluation capacity.	FMTI 2025, p. 21; OECD AI infrastructure, p. 8, p. 24, p. 29; CMA cloud investigation; Oxford Insights, p. 5, p. 10, p. 28; NAIRR/EuroHPC	Legal duties require practical access, staff, compute, testing environments, documentation, and enforcement capacity.	Does not prove every regulator is underprepared or that regulation is futile.	Well-supported assumption
Cloud concentration and switching barriers can reduce contestability.	OECD AI infrastructure, p. 18, p. 24, p. 29; Ofcom cloud referral; CMA cloud investigation; World Bank, p. 61, p. 66	AI development and deployment often depend on cloud platforms whose market structure, contracts, data gravity, and tooling can create lock-in.	Does not prove all cloud partnerships or cloud services are anticompetitive.	Established finding / well-supported assumption
The compute divide has a global development dimension.	World Bank, p. 11, p. 15-16, p. 61, p. 67; Oxford Insights, p. 5, p. 10, p. 28-29, p. 59; Stanford AI Index 2026, p. 33; OECD, p. 24	Compute, servers, HPC capacity, data-center capacity, AI capital, and government readiness are unevenly distributed across countries and income groups.	Does not imply lower-income countries are passive or that domestic frontier training is the only route to agency.	Established finding / well-supported assumption
Energy and local infrastructure are part of the compute divide.	IEA, p. 49, p. 56-57, p. 63, p. 260; LBNL / DOE, p. 6-7, p. 52, p. 57; CSO Ireland official data; OECD, p. 17-18	Compute is materially embedded in electricity grids, data centers, cooling systems, water, land, local permitting, and community tradeoffs.	Data-center electricity is not the same as AI electricity; local harms are case-specific.	Established finding / scenario-based projection
Counterforces mitigate the divide unevenly.	Stanford AI Index 2026, p. 72, p. 76-77; Pilz, Heim, and Brown, p. 1-7; NAIRR, p. 5, p. 8-9, p. 22; EuroHPC/EC AI Factories; OECD/Ofcom/CMA	Open-weight models, efficiency gains, public compute, structured access, and competition policy can reduce some barriers.	They do not automatically equalize frontier training, full reproduction, secure deployment, or regulatory testing capacity.	Counterargument / well-supported assumption

Appendix B: Figures And Data Tables

Figure B1: Core Feedback Model

The main feedback model is presented in Section 4.4 as Figure 1. It should be read as a conceptual model rather than a causal estimate. The key point is that infrastructure concentration, AI adoption, countercapacity, counterforces, and physical constraints interact rather than forming a one-way deterministic chain.

Figure B2: Four Capacity Layers

Layer	Capacity question	Short description
1	Use access	Can actors use AI systems through interfaces, APIs, open-weight models, or applications?
2	Production capacity	Can actors build, train, adapt, or substantially shape advanced AI systems?
3	Deployment capacity	Can actors operate AI systems reliably and securely at scale?
4	Evaluation and reproduction capacity	Can actors test claims, reproduce results, inspect evidence, and evaluate risks independently?
5	Governance and contestation capacity	Can actors regulate, negotiate, challenge, or democratically contest high-impact AI systems?

The compute divide is not primarily about whether actors can use AI tools. It concerns whether they can move upward from use access to production, deployment, evaluation, reproduction, governance, and contestation capacity.

Figure B3: Ireland Data-Center Electricity Growth

Year	Data-center metered electricity consumption (GWh)
2015	1,238
2016	1,480
2017	1,760
2018	2,180
2019	2,488
2020	3,028
2021	4,010
2022	5,270
2023	6,335
2024	6,969

Ireland illustrates why data centers should be treated as local infrastructure, not only as abstract cloud capacity. Official CSO data report that data centers' share of metered electricity rose from 5% in 2015 to 22% in 2024.

Appendix C: Counterforce Matrix

Counterforce	What it improves	What it does not solve
Open-weight models	Use, adaptation, local research, competition	Frontier training, full reproduction, deployment infrastructure
Small and domain models	Cost, latency, local deployment	General frontier capability
Algorithmic efficiency	Compute requirements for some capabilities	Evaluation, governance, security, total energy growth
Public compute	Research and public-sector access	Private frontier scale unless sustained and large
Structured access	Third-party evaluation without full release	Independence if access remains discretionary
Competition policy	Lock-in, switching barriers, market power	Technical audit access by itself
AI safety institutes	Evaluation expertise, testing norms	Broad democratic contestation capacity
Transparency obligations	Documentation and accountability	Verification without access and staff