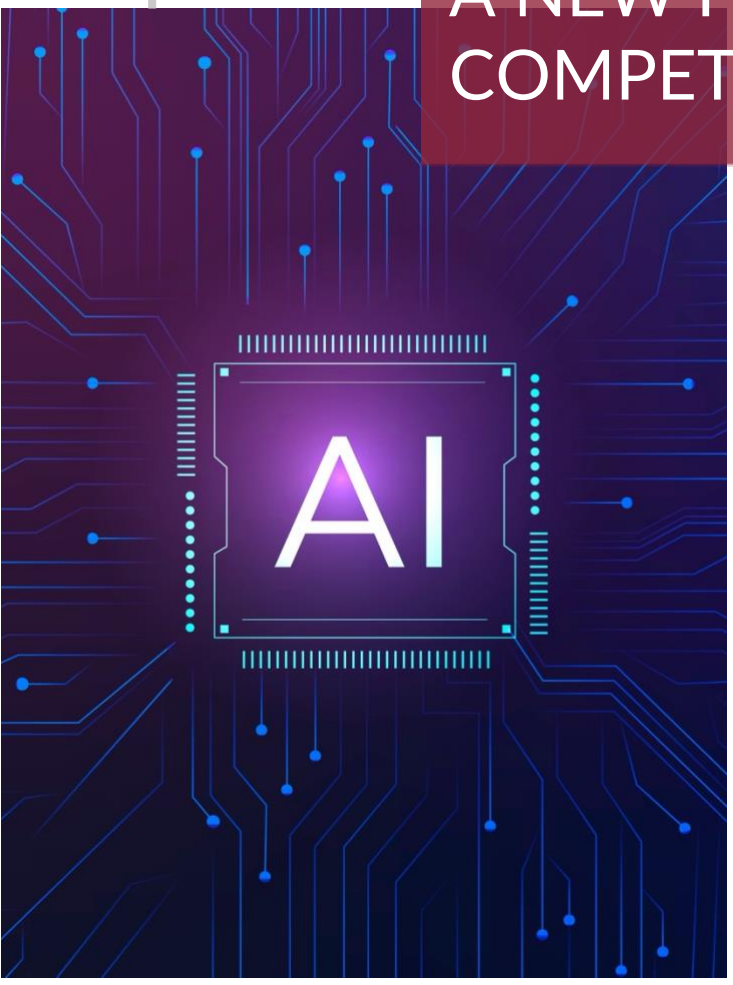
CIRANO
Knowledge into action

TERRITORIAL CONTROL OF DATA AND COMPUTE IN GENERATIVE AI: A NEW PARADIGM OF COMPETITIVE ADVANTAGE



FRÉDÉRIC MARTY, PHD
THIERRY WARIN, PHD

CS
2025s-27
WORKING PAPER

CIRANO
Knowledge into action

The purpose of the **Working Papers** is to disseminate the results of research conducted by CIRANO research members in order to solicit exchanges and comments. These reports are written in the style of scientific publications. The ideas and opinions expressed in these documents are solely those of the authors.

Les cahiers de la série scientifique visent à rendre accessibles les résultats des recherches effectuées par des chercheurs membres du CIRANO afin de susciter échanges et commentaires. Ces cahiers sont rédigés dans le style des publications scientifiques et n'engagent que leurs auteurs.

CIRANO is a private non-profit organization incorporated under the Quebec Companies Act. Its infrastructure and research activities are funded through fees paid by member organizations, an infrastructure grant from the government of Quebec, and grants and research mandates obtained by its research teams.

Le CIRANO est un organisme sans but lucratif constitué en vertu de la Loi des compagnies du Québec. Le financement de son infrastructure et de ses activités de recherche provient des cotisations de ses organisations-membres, d'une subvention d'infrastructure du gouvernement du Québec, de même que des subventions et mandats obtenus par ses équipes de recherche.

CIRANO Partners – *Les partenaires du CIRANO*

Corporate Partners – Partenaires Corporatifs	**Governmental partners - Partenaires gouvernementaux**	**University Partners – Partenaires universitaires**
Autorité des marchés financiers	Ministère des Finances du Québec	École de technologie supérieure
Banque de développement du Canada	Ministère de l'Économie, de l'Innovation et de l'Énergie	École nationale d'administration publique
Banque du Canada	Innovation, Sciences et Développement Économique Canada	de Montréal
Banque Nationale du Canada	Ville de Montréal	HEC Montreal
Bell Canada		Institut national de la recherche scientifique
BMO Groupe financier		Polytechnique Montréal
Caisse de dépôt et placement du Québec		Université Concordia
Énergir		Université de Montréal
Hydro-Québec		Université de Sherbrooke
Intact Corporation Financière		Université du Québec
Investissements PSP		Université du Québec à Montréal
Manuvie		Université Laval
Mouvement Desjardins		Université McGill
Power Corporation du Canada		
Pratt & Whitney Canada		
VIA Rail Canada		

CIRANO collaborates with many centers and university research chairs; list available on its website. *Le CIRANO collabore avec de nombreux centres et chaires de recherche universitaires dont on peut consulter la liste sur son site web*.

© Sept 2025. Frédéric Marty, Thierry Warin. All rights reserved. *Tous droits réservés.* Short sections may be quoted without explicit permission, if full credit, including © notice, is given to the source. *Reproduction partielle permise avec citation du document source, incluant la notice ©.*

The observations and viewpoints expressed in this publication are the sole responsibility of the authors; they do not represent the positions of CIRANO or its partners. *Les idées et les opinions émises dans cette publication sont sous l'unique responsabilité des auteurs et ne représentent pas les positions du CIRANO ou de ses partenaires.*

ISSN 2292-0838 (online version)

Territorial Control of Data and Compute in Generative AI: A New Paradigm of Competitive Advantage

Frédéric Marty[*], *Thierry Warin*[†]

Abstract/Résumé

The rapid advancement of generative artificial intelligence (AI) is increasingly shaped by control over two critical inputs: high-quality data and the compute infrastructure required to train and update large-scale model weights. This paper argues that these inputs – rather than algorithmic talent or novel architectures alone – have become the decisive strategic assets in generative AI, creating steep structural barriers to entry. We examine who controls these resources and how this control is territorially distributed across countries. Building on literature in industrial organization, competition policy, and international political economy, we highlight a gap in existing research: insufficient attention to the territorial concentration of "model-weight-setting" capacity, i.e. the ability to train cutting-edge foundation models. We find that the capacity to set foundation model weights is overwhelmingly concentrated in a few firms and regions, reinforcing market concentration and limiting the AI development sovereignty of most countries. While innovations in model architectures and efficiency (illustrated by the DeepSeek case) can reduce compute requirements at the margin, they do not eliminate the scale advantages conferred by privileged access to massive proprietary datasets and nation-scale computing clusters. The paper concludes with implications for competition and regulation, arguing that the territorial control of data and compute resources is a fundamental structural challenge for both market competition and global equity in AI.

Les progrès rapides de l'intelligence artificielle générative (IA) sont de plus en plus conditionnés par le contrôle de deux intrants essentiels : des données de haute qualité et l'infrastructure de calcul nécessaire pour entraîner et actualiser les poids de modèles à grande échelle. Cet article soutient que ces intrants – plutôt que le seul talent algorithmique ou la nouveauté des architectures – sont devenus les actifs stratégiques décisifs de l'IA générative, créant ainsi d'importantes barrières structurelles à l'entrée. Nous examinons qui contrôle ces ressources et comment ce contrôle se répartit territorialement entre les pays. En nous appuyant sur la littérature en organisation industrielle, en politique de concurrence et en économie politique internationale, nous mettons en évidence une lacune dans les recherches existantes : l'attention insuffisante portée à la concentration territoriale de la « capacité de réglage des poids des modèles », c'est-à-dire la faculté d'entraîner des modèles de fondation de pointe. Nos résultats montrent que cette capacité est largement concentrée dans quelques entreprises et régions, ce qui renforce la concentration des marchés et limite la souveraineté de la plupart des pays en matière de développement de l'IA. Bien que les innovations en matière d'architectures de

[*] Ancien membre du Collège de l'Autorité de la concurrence française au titre de personnalité qualifiée pour les professions réglementées. CNRS – GREDEG – Université Côte d'Azur. CIRANO

[†] HEC Montréal. CIRANO/OBVIA, CEIMIA/GPAI-OECD

modèles et d'efficacité (comme l'illustre le cas DeepSeek) puissent réduire les besoins en calcul à la marge, elles n'éliminent pas les avantages d'échelle conférés par l'accès privilégié à d'immenses ensembles de données propriétaires et à des grappes de calcul de dimension nationale. L'article conclut en soulignant les implications pour la concurrence et la régulation, en avançant que le contrôle territorial des données et des ressources de calcul constitue un défi structurel fondamental pour la concurrence sur les marchés et pour l'équité mondiale en matière d'IA.

Keywords/Mots-clés: Generative Artificial Intelligence, Data Sovereignty, Compute Infrastructure, Competition Policy, Territorial Concentration / Intelligence artificielle générative, Souveraineté des données, Infrastructure de calcul, Politique de concurrence, Concentration territoriale.

Pour citer ce document / To quote this document
Marty, F., & Warin, T. (2025). Territorial Control of Data and Compute in Generative AI: A New Paradigm of Competitive Advantage (2025s-27, Cahiers scientifiques, CIRANO.)
https://doi.org/10.54932/PIMA2204

1. Introduction

Since November 2022, generative AI—systems that learn from vast datasets to produce novel text, images, code, and other content—has emerged as a transformative general-purpose technology (OECD, 2025). A striking feature of this technological shift is its concentration in the hands of a few actors and regions. A small number of firms, largely headquartered in the United States and China, have achieved breakthroughs by deploying unprecedented amounts of computational power and training data. This concentration has raised concerns that generative AI markets could "freeze" into oligopolies dominated by these early movers, limiting competition and constraining other countries' ability to achieve technological sovereignty.

These concerns have gained renewed urgency with the market entry of DeepSeek in January 2025, which offers a more frugal alternative in terms of computational infrastructure compared to the major incumbents (the "Magnificent 7," according to Krause, 2025a). Does DeepSeek's arrival represent a Christensen-style disruption (Christensen, 2016) that mitigates lock-in risks identified by Marty and Warin (2025), or does it paradoxically risk reinforcing them, by strengthening arguments in favor of sovereignty-driven regulation and renewed forms of industrial policy?

Artificial intelligence now lies at the heart of debates on competitiveness and industrial policy, while rising international tensions are placing it firmly within a geopolitical frame (Vannuccini, 2025). The core questions driving this research are therefore: Who controls the data and compute resources that determine success in generative AI, and how is this control territorially distributed across countries?

This question sits at the intersection of industrial organization (IO), competition policy, and international political economy (IPE). Traditional IO emphasizes how control over critical inputs can create structural barriers to entry and grant incumbents durable advantages. In AI markets, many policymakers and scholars highlight data and compute as such inputs. Competition authorities warn that firms controlling quasi "essential facilities[1]," such as proprietary datasets or cloud infrastructure, could use them to stifle

[1] We employ the term quasi-essential facility rather than essential facility insofar as the asset in question does not meet the requirements established in the Bronner judgment of the Court of Justice (C7-97, 26 November 1998), particularly the criterion of indispensability of access to the market. A rival can still operate on the market without such access, yet its performance would be significantly hindered and, consequently, it would

downstream competition (Autorité de la concurrence, 2023; 2024). In parallel, IPE scholars and security analysts increasingly frame AI as a sovereignty issue: control over AI's key inputs is seen as conferring not only economic but also strategic influence.While existing literature acknowledges the importance of data and compute, it rarely addresses their territorial concentration and its implications. Much of the competition debate has focused on platform data network effects and the dominance of a few cloud providers, without fully grappling with the uneven geography of the resources needed to train frontier models. National AI strategies often cite the "AI triad" (data, algorithms, compute), yet they underplay the fact that the ability to perform large-scale model training—requiring enormous datasets, specialized hardware, and expert talent—is concentrated in a handful of countries. Our contribution is to bring this territorial dimension to the fore: we systematically examine how the geography of data and compute shapes competition in generative AI.The entrenched position of large ecosystem operators in this domain can be understood through two key dynamics. The first, examined in Marty and Warin (2025), relates to dependencies arising from critical infrastructures. Addressing these requires both ex post interventions, through competition enforcement, and ex ante regulatory frameworks designed to mitigate risks of economic and technological dependence. The second dynamic concerns digital sovereignty (Falkner et al., 2025). Here, two trends converge. The first, emphasized in the Letta (2024) and Draghi (2024) reports, is a European awareness of the widening technological and economic gap with the United States and China, prompting calls for both regulatory reform and industrial policy initiatives (Vannuccini, 2025) that avoid capture by incumbents. The second trend, accelerated by COVID-19, is the return of neo-mercantilist strategies (Briganti Dini, 2025), fragmenting the global economy, especially in digital industries. In some ways, the paradigm has shifted from comparative advantage to zero-sum competition; geopolitically, rising tensions have placed national security imperatives at the center of trade and industrial strategies, spurring calls for strategic autonomy or, at minimum, friendshoring (Warin,

appear less attractive to consumers. For a broader perspective on the notion of a quasi-essential facility (where an access obligation may be justified by the risk that a rival would be unable to offer a service of equivalent quality) reference can be made to the Android Auto judgment of the Court of Justice (case C-233/23, 25 February 2025).

2025).These dynamics require a rethinking of competition and industrial policy in a context complicated further by disruptive entrants such as DeepSeek. Public action toward Big Tech in generative AI must now contend with this technological, competitive, and geopolitical landscape. In this respect, the critical question is whether Big Tech's control over essential assets (Marty & Warin, 2025) has become less of a lock-in factor than in the past, or whether it retains its full significance under new forms.In addressing this question, our article builds on and extends the argument that the decisive strategy in generative AI lies in controlling two inputs—data and compute—rather than transient advantages in talent or algorithms. We show how these inputs generate steep, self-reinforcing barriers to entry and argue that their territorial concentration is reshaping both competition and sovereignty. The paper proceeds as follows. We first situate our argument in the literature, drawing on IO, competition policy, and IPE. We then explain why data and compute constitute the *sine qua non* of generative AI progress, emphasizing the economic logics of scale and spillovers. Next, we document the territorial distribution of these resources, showing that the capacity to train frontier models is overwhelmingly concentrated in a small set of countries. Within this discussion, we highlight the case of DeepSeek, whose efficient model design narrowed the performance gap with incumbents, demonstrating that innovation can ease but not eliminate structural input constraints. Finally, we explore the implications of our findings: persistent concentration is likely unless proactive measures address disparities in access to data and compute, and sovereignty concerns will grow as countries depend on foreign AI infrastructures.We argue that meaningful policy responses—whether through antitrust action, data-sharing mandates, or public investment in AI infrastructure—must recognize that data and compute are the new strategic assets in the AI economy. It is against this backdrop that we assess how the evolving technological and competitive environment is shaping regulatory and industrial policy initiatives in the era of generative AI.

The paper is structured as follows. Section 2 presents our theoretical framework. Section 3 addresses the issue of data while Section 4 deals with the importance of computing capacities. Section 5 shows how the interactions between data and computing capacities reinforces competitive advantage and thus market concentration. Section 6 discusses the implications in terms of competition policy and regulation and Section 7 concludes.

2. Literature Review and Theoretical Framework

2.1 Industrial Organization: Data and Compute as Structural Barriers to Entry

The economics of information technology have long emphasized that certain inputs can function as barriers to entry by granting incumbents a cost or quality advantage not easily replicated by new entrants (Bain, 1956). In digital markets, network effects and economies of scale frequently produce "winner-take-most" outcomes (Belleflamme & Peitz, 2015). Classic examples include operating systems or social networks, where value grows with the number of users and accumulated data.

In the case of generative AI and foundation models, scholars are investigating whether similar dynamics apply to training data and computing capacity (Schrepel & Pentland, 2024; Carugati, 2024). Training state-of-the-art models requires tens of millions of dollars in compute resources and massive datasets, placing firms that already control such inputs in a structurally advantaged position. Recent analyses suggest that access to large-scale data and compute indeed creates formidable entry barriers. For example, researchers at the Brookings Institution highlight that "limited resources like talent, data, [and] computational power" are key hurdles for new entrants into foundation model markets (Brookings, 2023).

The economics of scale reinforce these challenges. In AI development, returns to scale are steep: using 10× more data or compute can yield disproportionate performance improvements, incentivizing ever-larger training runs by firms that can afford them. This creates a feedback loop reminiscent of classic industrial organization contexts where high fixed costs and scale economies lead to oligopoly or natural monopoly. In AI, the fixed costs are not only tied to physical infrastructure (e.g., data centers) but also to the one-time expense of producing high-quality model weights. Once trained, models can be deployed widely at relatively low marginal cost—a cost structure (high fixed, low marginal) long recognized as conducive to concentration (Varian, 2018).

Some scholars, however, stress that these barriers may be surmountable or transient. Abbott and Marar (2025), for example, challenge what they call "fears of data scarcity and monopolization." They argue that open data and data markets allow startups to access

sufficient training resources, while algorithmic advances and synthetic data reduce total requirements. Smaller models fine-tuned on the "right kind of data" may even outperform larger but less targeted ones. According to this perspective, superior talent and innovation can offset incumbents' resource advantages, echoing Schumpeter's idea that technological innovation can overturn monopolies. Empirical cases of independent AI firms producing competitive models with far fewer resources illustrate this possibility.

Yet even optimistic analyses concede that, absent intervention, current trends point toward growing concentration. Regulators such as the U.S. Federal Trade Commission (FTC, 2023), the UK's Competition and Markets Authority (CMA, 2024), and competition authorities in France (2024) and Portugal (2024) have warned that control of proprietary data or cloud infrastructure by a handful of firms could confer "unassailable" advantages and enable discriminatory practices. This raises questions reminiscent of the "essential facilities" doctrine in competition law: if critical inputs like data or compute are controlled by dominant players, should they be required to grant fair access to rivals?

The principal competitive risk, therefore, is that Big Tech firms—rather than being disrupted by entrants—may consolidate their positions through generative AI, leveraging both their resources and their ecosystems (Marty & Warin, 2025). Two competing narratives emerge. The first is one of creative destruction: established barriers no longer protect incumbents from challengers, whether from leading AI startups (OpenAI, Anthropic, Mistral) or new entrants such as DeepSeek. The second emphasizes consolidation: incumbents may integrate innovative firms as complementors or reinforce the essentiality of their existing resources.

Beyond this disruption-versus-consolidation debate lies the question of AI's systemic nature. Generative AI can be understood both as a General Purpose Technology (GPT) and as a Large Technical System (LTS). Its effective development depends on diverse resources—algorithms, data, storage infrastructures, compute—and is thus ecosystemic in character (Vannuccini & Prytkova, 2023). The dominant "bigger is better" paradigm (Varoquaux et al., 2024) may, however, be misleading, as illustrated by DeepSeek's innovations in more efficient training. Recognizing AI's ecosystemic dynamics suggests that traditional industrial policy tools—such as public pre-financing of infrastructures or

procurement strategies—are insufficient. Instead, a balanced mix of policies is required: fostering competition to prevent lock-in, while also encouraging cooperation to sustain ecosystem development (Vannuccini, 2025).

This tension between cooperation and competition is typical of digital ecosystems. Innovation and market access often depend on infrastructures developed by systemic incumbents, who can enable or hinder complementors depending on their strategic interests (Marty & Warin, 2023). Dominant firms can entrench their positions both through control of (quasi-)essential infrastructures and through acquisitions of potential disruptors (Marty & Warin, 2021). These dynamics highlight the importance of vertical coordination and complementarities across three key layers of generative AI: the physical (infrastructures), the code-related (standards, algorithms), and the content-related (data)[2]. Strategic assets across these layers share three characteristics (Fontana & Vannuccini, 2024): access to infrastructure is critical for complementors to develop and deploy applications; assets are shaped by path dependency (as with reliance on cloud resources); and they are marked by a scarcity of viable alternatives, creating economic and technological dependency. Such dependencies generate specific challenges not only for competition policy but also for industrial policy.

So, industrial organization perspectives highlight that data and compute function as entry barriers in generative AI. The high fixed costs and scale economies of training frontier models naturally concentrate activity in a small number of firms. Absent proactive intervention, economic theory predicts persistent concentration. Our analysis extends this understanding by stressing the territorial dimension: these inputs are not only concentrated in a few firms but also geographically concentrated in a handful of countries. The following section examines how competition policy is responding to this dual concentration, and what it implies for the international distribution of AI capacity.

[2] The question is particularly important in the field of generative AI, where instead of vertical growth operations through acquisitions, we increasingly observe partnerships between Big Tech firms and new entrants. Two interpretations of this phenomenon are possible. The first is to view such partnerships as a means of achieving, de facto, the same results as a formal acquisition. The advantage of this contractual integration is that it can more easily evade ex ante merger control. The second interpretation is that these are hybrid partnership arrangements between actors with complementary resources. The fact that such partnerships generally do not involve exclusivity clauses in favor of Big Tech would support this latter reading (Groza, 2025).

2.2 Competition Policy: Antitrust Concerns and Regulatory Debates

Competition authorities in the United States, the European Union, and other jurisdictions have begun to scrutinize generative AI through the lens of antitrust and competition law. A central concern is that control of foundational inputs—data, compute, and talent—by incumbent technology giants could enable them to distort competition in generative AI markets. In a June 2023 policy statement, the U.S. Federal Trade Commission (FTC) explicitly warned that if a small number of firms control these essential ingredients, they could dampen rivalry and ultimately wield "outsized influence over a significant swath of economic activity" as AI becomes ubiquitous. This reflects a broadening of competition policy beyond traditional product-market analysis toward the upstream level of inputs and infrastructure.

Data has been the primary focus of these debates, partly because of analogies with earlier digital markets. Regulators recall how control over user data allowed Google and Facebook to cement dominance through targeted advertising[3] and network effects[4]. In generative AI, the question is whether access to more or better data—such as search query logs, private social media content, or proprietary text and image stocks—enables training fundamentally better models that rivals cannot replicate. The FTC notes two particular challenges for entrants: (1) incumbents accumulated massive datasets over many years via consumer-facing services unavailable to startups, and (2) incumbents have developed specialized tools and infrastructures to acquire data at scale. In specialized domains such as health or finance, access is even more restricted, giving incumbents with existing partnerships a structural advantage. While holding large datasets is not illegal per se, regulators worry that it can amount to a barrier to entry that prevents "fair competition from fully flourishing." Proposed remedies range from encouraging portability and open-data initiatives to more ambitious measures such as mandating dominant platforms to share data under the idea of "data commons" or as a condition in merger remedies.

[3] See the US Google Seach Case and the September 2nd, 2025 judgment.
[4] See the EU Commission decision Google Ad-Tech and data related Practices, case AT.40670, September 5th, 2025

Compute has more recently attracted regulatory attention as a second critical input. Training frontier models requires access to specialized processors (e.g., GPUs, TPUs) and vast computational capacity, resources controlled by only a few cloud providers. New entrants typically rent compute from Amazon AWS, Microsoft Azure, or Google Cloud in the West, or from firms like Alibaba Cloud in Asia. Regulators are concerned that these providers may privilege their own AI efforts or those of favored partners, raising risks of bundling, preferential pricing, or foreclosure. The aborted $40 billion merger of Nvidia and Arm in 2022—blocked partly due to competition concerns—remains a touchstone in these debates. Today, with demand for AI accelerators surging, authorities are wary of scenarios where dominant cloud providers might lock up supply or impose restrictive contracts, for example by charging high data egress fees that discourage switching.

Talent constitutes a third critical input. While not the focus of this paper, its scarcity is widely acknowledged: the limited pool of top AI researchers gives incumbents incentives to use non-compete clauses or no-poach agreements to restrict labor mobility (Chaiehloudj, 2025). Regulators stress that talent mobility is essential to allow new ventures to emerge, though talent alone is insufficient without access to data and compute—a point reinforcing the argument that these inputs form structural choke points.

These concerns have sparked debate about the appropriate regulatory approach: is the disruption possible or should we fear entrenched dominant positions (Hagiu and Wright, 2025)?. One school of thought, exemplified by Abbott and Marar (2025), cautions against "heavy-handed" interventions such as mandatory data sharing. They argue that such measures risk reducing incentives to collect and curate data, ultimately undermining innovation. Others, following Tirole (2023), highlight the dangers of "heavy-handed regulation," which can raise compliance costs, reduce profit expectations, and dampen innovative dynamism, while remaining prone to regulatory capture. Tirole instead advocates a "light-handed" approach aimed at lowering entry barriers through interoperability and portability, thereby fostering multi-homing and limiting lock-in—concerns also identified in the French Competition Authority's 2023 cloud inquiry.

Ex ante obligations of this kind already exist in the European Union. The Digital Markets Act (DMA), in force since 2022, and the Data Act, which entered into force in 2024, impose

requirements on data sharing and portability[5]. These obligations are designed to prevent irreparable competitive harm and to address conduct that, under competition law, would otherwise be sanctioned ex post as exclusionary abuses. They effectively shift some enforcement from an *ex post* to an *ex ante* framework (Bougette et al., 2025), while leaving room for traditional antitrust action[6].

Still, there is no consensus on remedies. Some scholars and regulators favor reliance on existing antitrust law to punish exclusionary conduct if and when it occurs, while others stress that by the time harm is observable, markets may already have tipped irreversibly due to high fixed costs and economies of scale. The debate highlights a key tension: fostering competition in generative AI likely requires ensuring more open access to data and compute, but how to do so without discouraging investment remains contentious.

Thus, competition policy perspectives underscore that control over data and compute is not merely a business advantage but a potential chokepoint for the entire AI ecosystem. Regulators are increasingly alert to scenarios in which incumbents' dominance over these inputs could solidify into lasting market power. Yet the appropriate policy response is still unsettled. What is clear is that interventions must take into account the global distribution of these resources, which leads directly to the industrial policy and international political economy dimensions that we examine next.

2.3 From the guarantee of a level playing field to geo-dirigiste industrial policies

The specific context of generative AI development has led to public policies that extend beyond the traditional goals of safeguarding undistorted competition and protecting citizens' personal data[7]. States are increasingly engaged in an "AI race," either to promote their national champions or to reduce the dependence of domestic firms on services

[5] For instance, Meta was sanctioned in April 2025 for extracting too much data from its users (considering its option 'consent or pay' was not satisfying). See EU Commission decision, April 23rd, 2025.
[6] See for instance the February 2025 Android Auto Judgment mentioned above and also the EU Commission decision related to the Google Ad Tech case in September 2025.
[7] Lacking major players in the Big Tech sector, the European Union has initially relied on competition and regulatory instruments to limit lock-in risks for its firms by targeting the control of critical resources of the digital economy—such as data and computational capacities, which are the focus here. These policies pursue a dual objective. Domestically, they aim to ensure a level playing field in competition and to safeguard values that are not purely economic, such as the protection of personal data. Externally, they seek to promote a regulatory model that, it is hoped, will be emulated by partner states (Bradford, 2012).

provided by Big Tech. In Europe, competitiveness is now closely tied to the management of international interdependencies (Farrand & Carrapico, 2022). The challenge lies not only in defining industrial policies that enable technological catch-up or preserve the autonomy of European firms (Seidl & Schmitz, 2023), but also in articulating a coherent strategy for digital sovereignty.

State intervention in this field—beyond competition policy—appears necessary given the risks of competitive lock-in and foreign control of ecosystems. Yet such intervention faces significant challenges of implementation. Two risks are particularly salient: the distortion of industrial policy for protectionist purposes, and regulatory capture by firms themselves (Vannuccini, 2025). Protectionist strategies, as in the case of state aid or tariff barriers, may generate collectively suboptimal outcomes and, in the longer term, harm the very state that pursues them through industrial and technological decline. The risks of regulatory capture are equally acute: firms have strong incentives to shape regulation to their advantage, especially in a context marked by imperfect information and heightened public concern about the ethical and sovereignty-related implications of AI. As Vannuccini (2025) observes, the central issue here is less about existential risks from AI than about managing competitive dynamics—both among firms and among states.

European debates (Letta, 2024; Draghi, 2024) should therefore be situated within a dual perspective: at the microeconomic level, enhancing competitiveness and reducing European firms' dependence on Big Tech; and at the macroeconomic level, strengthening strategic independence—or at least managing strategic interdependencies. Some proposed solutions, such as those in the Draghi Report (2024), focus on completing the internal market and reassessing the cost-benefit balance of existing regulations. Others, more demanding, concern the very definition of a European industrial policy in the context of potential technological catch-up.

One approach is to reduce dependency by developing autonomous infrastructures. This, however, raises questions about the likelihood of success in an investment race against states that already command vast infrastructures and enjoy easier access to critical assets and technologies, as well as the risk of industrial capture (Singh, 2025). A second option, also noted in the Draghi Report, is to adjust merger control so that efficiency gains are

weighed more heavily relative to competitive harm. In both cases, however, the risk of capture by powerful firms remains substantial. Alternatively, European industrial policy could seek to reduce dependencies not only vis-à-vis Big Tech, but also vis-à-vis trading partners that may adopt non-cooperative strategies. Initiatives to shape standards, promote open-source solutions, develop below-frontier AI applications (Martens, 2024) and decentralized learning capabilities all align with this approach.

2.4 International Political Economy: AI Development and Territorial Sovereignty

The rise of generative AI has not only economic and legal facets, but also a geopolitical one. Nations increasingly frame AI capabilities as matters of national competitiveness and security. This perspective belongs to the domain of international political economy (IPE) and technology governance, where questions of who leads and who lags in AI are deeply intertwined with concerns over sovereignty, dependence, and global power balances.

The issue of dependency has gradually evolved. Initially, it was framed at the level of firms—complementors vis-à-vis keystones in digital ecosystems—as illustrated, for example, by Regulation (EU) 2019/1150 on promoting fairness and transparency for business users of online intermediation services. At that stage, dependency was seen primarily through a B2B lens: a matter of contractual imbalance linked to asymmetries in bargaining power, only partially addressed under competition law. U.S. antitrust law does not sanction exploitative abuse, and only some EU member states recognize abuse of economic dependence. Nevertheless, such dependency raised concerns about dynamic efficiency, notably by constraining firms' innovative capacity and biasing innovation trajectories toward complementarity with keystone actors. In the European case, fairness rules were sometimes interpreted less as competition policy than as industrial or trade policy, aimed at protecting European firms in the absence of domestic platform champions (Radic et al., 2025). Related fiscal initiatives in member states were occasionally viewed as non-cooperative trade practices, amounting to de facto tariffs (Schramm, 2025).

With the advent of generative AI, these tensions have expanded from inter-firm relations to inter-state relations. The growing rivalry between the United States and China has led to export controls on advanced processors, underscoring the weaponization of technological

interdependence (Farrell & Newman, 2019). Global value chains once grounded in comparative advantage are giving way to strategies focused on strategic autonomy, resilience, and security. The discourse on digital sovereignty, while often presented in geopolitical terms, also encompasses legitimate industrial policy concerns. Yet it remains vulnerable to capture by domestic firms seeking regulatory protection.

A striking pattern in the geography of AI capability is its concentration in just a few states. The United States and China dominate in terms of talent, firms, large-scale datasets, and compute power. Metrics such as top AI publications, the largest model deployments, and the distribution of major AI data centers reflect this duopoly. Europe, India, Canada, and others have vibrant AI research communities, but lag behind in pretraining frontier models or scaling resources. European policymakers, for example, warn that without local capacity to train or adapt foundation models, the EU risks becoming a "rule-taker"—consuming technologies shaped by others' norms (Floridi, 2020; European Parliament Report, 2022). This has fueled calls for "AI sovereignty" or "digital sovereignty," understood pragmatically as the ability to develop and operate AI systems on domestic infrastructure, data, and workforce rather than relying entirely on foreign providers.

Empirical work highlights the stark unevenness of AI infrastructure across the globe. Hawkins, Lehdonvirta, and Wu (2025) mapped "compute sovereignty" by identifying countries with AI-capable data centers. Their study found that only 33 countries host facilities with accelerator hardware, and just 24 have the capacity to train full-scale foundation models. Over 160 countries lack any significant AI compute infrastructure. Of 132 major AI clusters identified worldwide, 26 were in the U.S. and 22 in China; together, those two countries account for roughly one-third of sites, and likely an even higher share of global compute capacity. EU member states collectively hosted 27 clusters, while other hubs included the UK, Canada, Japan, South Korea, Singapore, and the UAE. Entire continents such as Africa and South America are barely represented, with only South Africa and Brazil hosting notable facilities. The result is a deepening digital divide: if data centers are the engines of the AI economy, much of the Global South lacks the machinery.

This territorial concentration matters for several reasons. Economically, countries without AI infrastructure risk exclusion from innovation ecosystems and high-value industries,

while also facing brain drain as researchers migrate to the few global hubs. Strategically, reliance on foreign AI services exposes states to vulnerabilities: sanctions, export controls, or contractual restrictions could curtail access to models or cloud capacity critical for economic and defense activities. The U.S. restrictions on advanced GPUs to China, and China's subsequent push for indigenous AI chips, illustrate this techno-nationalist dynamic.

The analogy with oil is particularly resonant in IPE debates. Just as control over oil resources granted geopolitical leverage in the 20th century, control over AI compute may confer strategic power in the 21st. This includes not only ownership of data centers but also control of the semiconductor supply chain. Here, interdependence is acute: cutting-edge AI chips are manufactured primarily in Taiwan (TSMC) and South Korea (Samsung), even if designed by U.S. firms such as Nvidia and AMD. The global AI infrastructure thus depends on a fragile geographic nexus, with Taiwan's political status casting long shadows over AI's future. Efforts such as the August 2022 U.S. CHIPS Act, China's vast semiconductor investments, and the September 2023 EU Chips Act are all attempts to secure sovereignty over this compute backbone.

Finally, territorial concentration also shapes norms and governance. Advanced AI models embody not only technical architectures but also the values of their originators. U.S. models may privilege certain liberties, while Chinese models may embed stricter state-aligned constraints. Countries that adopt foreign models import not only technologies but also implicit normative frameworks, raising concerns of ideational sovereignty and what some scholars describe as "digital colonialism."

Taken together, an IPE lens reveals that the uneven distribution of AI's inputs—data, compute, and chip manufacturing—is producing a global hierarchy of AI capability. A handful of states lead; most depend. Calls for "AI sovereignty" seek to mitigate this imbalance, yet achieving technological self-determination is profoundly difficult in a landscape where critical resources remain so unequally distributed. The following sections examine the two key inputs—data and compute—in detail, before analyzing how their control translates into competitive advantage and what the case of DeepSeek suggests about the prospects for disruption in this landscape.

3. Data as a Strategic Resource in Generative AI

Data is often dubbed "the new oil" of the digital economy, and in the realm of AI, it is indeed the fuel that powers model training (Warin et al, 2014). However, not all data is equal. The success of generative AI models depends on access to high-quality, domain-relevant, and diverse datasets. In this section, we examine the role of data in generative AI, who controls the most valuable data, and why possession of large datasets constitutes a barrier that new entrants struggle to overcome. We also discuss whether this barrier is absolute or if there are emerging ways to circumvent data limitations.

3.1 The Importance of Scale and Quality in Training Data

Generative models, especially large language models (LLMs) and other foundation models, require exceptionally large training corpora. For instance, GPT-3 (2020) was trained on approximately 300 billion tokens of text, sourced from Common Crawl, digital libraries, and other large-scale internet datasets—effectively a significant slice of the web's textual content (Brown et al., 2020). A general principle has emerged that more data tends to yield better performance, up to the point of diminishing returns. This is supported by research on scaling laws, which demonstrate that model performance improves in predictable power-law relationships as a function of training data volume, model size, and compute budget (Kaplan et al., 2020). Thus, a firm able to assemble a trillion-token dataset holds a structural advantage in training more fluent and knowledgeable models compared to a firm with access to only a billion tokens. This scale effect constitutes a quantitative barrier to entry.

Equally important, however, is the quality and diversity of data. A dataset spanning multiple domains and styles enables a model to generalize broadly—hence the term "foundation model." Incumbent developers often leverage web-scale data collected through large-scale crawling (e.g., Common Crawl, Wikipedia, Reddit, news articles). Beyond these public resources, leading technology firms augment training with proprietary datasets: Google has access to years of search query logs and clicked results, Meta (Facebook) to billions of social media posts and images with metadata, and Microsoft to LinkedIn and GitHub datasets on professional networks and software code. Such

proprietary caches are not readily accessible to rivals. New entrants must therefore rely on open data or purchase licenses from third parties. While commercial data markets are expanding, the most valuable categories of high-quality, domain-specific data (such as medical records, financial transaction data, or user interaction data reflecting human preferences) are tightly controlled or restricted by privacy and regulatory frameworks.

A further dimension is the feedback loop in data advantage. The more users a company has for its AI products, the more interaction data it collects to improve its models, notably through techniques like reinforcement learning from human feedback (RLHF). This creates data network effects: products become better as more people use them, because their usage generates data that enhances performance. OpenAI's ChatGPT illustrates this dynamic: by rapidly reaching 100 million users, it collected vast volumes of feedback that could be used to fine-tune and align its models. Competing chatbots with smaller user bases lack access to such volumes of feedback data. The U.S. Federal Trade Commission (FTC, 2023) has noted that such "positive feedback loops" could enable incumbents to secure durable performance leads simply by virtue of scale in usage and data. This process is analogous to Google's search engine, which improved over decades by learning from billions of search queries—a level of cumulative data advantage difficult for new entrants to replicate[8].

3.2 Data Control and Ownership

From a competition standpoint, the question of who owns or controls valuable data is crucial. Public data, such as most webpages, remains freely scrapable, and smaller AI companies have made use of large portions of internet text. Yet as generative AI has expanded, content owners have begun to push back. Major platforms such as Reddit, Stack Exchange, and Twitter have restricted free API access or imposed paid licenses for the use of their content in AI training. This illustrates that the supply of freely accessible, high-quality data is not limitless. As valuable sources become monetized, incumbents with deep financial resources gain an advantage. Firms that already hold large datasets through their

[8] Some of the behavioral remedies imposed by the September 2nd 2025 judgement in the Google Search case tend to address this issue.

core business activities are spared such costs, while independent startups may now face high licensing fees for textual or image data that was once openly available. This shift risks reinforcing the incumbency advantage.

Personal and user-generated data form another contested category. Legislation such as the European Union's General Data Protection Regulation (GDPR) imposes strict limits on how personal data can be processed. These rules can be double-edged: they constrain the ability of large firms to use personal data, but they also make it more difficult for new entrants to gather comparable data legally. In addition, there is a linguistic and cultural dimension. The majority of accessible training data exists in English and a few other dominant languages, which structurally benefits firms operating in those linguistic domains—often U.S.-based companies for English. By contrast, countries with weaker representation online risk a "data deficit." European policymakers, for instance, have expressed concerns that foundation models trained primarily on English and Chinese data underperform for European languages and domain-specific applications unless fine-tuned. This has become part of broader debates over AI sovereignty (European Commission, 2023).

Finally, data ownership also extends to government-held datasets. Several countries are exploring ways to leverage public resources such as administrative records, national archives, or health system data to support domestic AI development. National healthcare systems, for instance, contain rich medical datasets that could, if mobilized responsibly, provide a sovereign data advantage in training medical AI. Yet exploiting these resources raises ethical and governance challenges and requires significant policy coordination. To date, leading AI laboratories have been primarily private-sector ventures, with limited reliance on government datasets. Nonetheless, it is foreseeable that states will increasingly treat certain categories of data as strategic national assets—resources to be guarded carefully or pooled selectively for domestic innovation.

3.3 Can Data Barriers Be Overcome?

The critical question is whether the absence of proprietary data constitutes a definitive barrier for aspiring competitors, or whether ingenuity and alternative strategies can

compensate. While data barriers are significant, there are reasons to believe they are not insurmountable in every case.

The first mitigating factor lies in the availability of open-source and shared datasets. The AI community has a strong tradition of open access. Projects such as LAION have compiled billions of images with captions for public use, while The Pile and other large text corpora aggregate diverse materials into massive datasets accessible to all. These initiatives enable motivated groups to assemble reasonably large and capable datasets without relying on proprietary caches. Indeed, many open-source models, such as EleutherAI's GPT-Neo series and Stability AI's Stable Diffusion, were trained primarily on open data. However, these datasets often lack the curation and domain-specific content that proprietary corpora offer. Open data may include duplication, noise, or gaps in specialized areas—for instance, much high-quality code, scientific literature, and domain-specific text remains behind paywalls.

A second dimension is the use of synthetic data. An intriguing workaround involves employing AI itself to generate additional data for training. A baseline model can produce synthetic text or images that, if sufficiently high in quality, expand the training set. This approach has proven useful in reinforcement learning, where self-play generates vast quantities of training data, and researchers have suggested it could help in language modeling as well. However, synthetic data raises the risk of a feedback loop: models may end up training on their own outputs, potentially leading to degraded performance. Recent scholarship has described this risk as "model collapse," whereby iterative retraining on AI-generated data reduces diversity and quality (Shumailov et al., 2023).

Transfer learning and fine-tuning provide a third mitigating pathway. Competitors may bypass the need for trillion-token datasets by fine-tuning smaller pre-trained models on highly targeted proprietary datasets. A startup unable to train a 100-billion parameter model from scratch could instead adapt an open-source base model with domain-specific corpora, such as legal documents or medical records, to achieve superior performance in that niche. This strategy is increasingly common given the lower compute requirements for fine-tuning. Yet the approach depends on the availability of robust base models. If base models are controlled by incumbents, dependency persists. The release of Meta's LLaMA-2 in

2023 as an open, high-quality model mitigates this issue by providing a strong starting point for fine-tuning. Policymakers frequently stress the importance of such open models for democratizing AI development (Vannuccini & Prytkova, 2023).

A fourth mitigation factor is the potential role of regulation and mandated data sharing. As above mentioned, regulators could, in principle, require dominant firms to make certain datasets available to competitors or researchers under privacy-safe conditions. For example, if a handful of companies possess uniquely rich datasets from consumer devices, policy could mandate portability or interoperability through APIs. Precedents exist in other sectors, such as telecom interconnection rules, though applying similar mechanisms to AI data raises complex questions of privacy, intellectual property, and incentives. While no regulation currently requires AI training data to be shared, such policies are actively debated in competition and technology governance circles (Tirole, 2023).

Despite these mitigating factors, it remains largely true that firms such as Google, Meta, OpenAI (with Microsoft), and Baidu, Tencent, or Alibaba in China possess durable data advantages. Abbott and Marar (2025) downplay long-term concerns about data scarcity, arguing that once a model has "enough" of the right data, marginal returns diminish and other factors—such as algorithms or user experience—become more decisive. They compare this to face recognition models, where performance gains from adding more faces beyond a threshold are minimal. While this observation holds in narrow domains, the challenge in generative AI is that the threshold of "enough data" for broad, open-ended tasks may be extraordinarily high—potentially encompassing much of the internet's content. Even if returns diminish at the frontier, new entrants must first reach the incumbents' performance level. If leading firms have already trained on most of the world's high-quality text, followers face the choice of relying on lower-quality or non-text modalities, or innovating radically more data-efficient methods.

In sum, data remains a foundational asset in AI, one that provides durable—though not entirely insurmountable—advantages to those who control it. Control over data is distributed unevenly, with global platform companies and a handful of nations, especially the United States and China, holding the bulk of valuable resources. This asymmetry reinforces the structural lead of incumbents and leading states. That said, data advantages

are ultimately inseparable from compute capacity: making use of large datasets requires the infrastructure to process them at scale. We therefore turn next to the second critical input: compute infrastructure.

4. Compute Infrastructure: The Backbone of Foundation Model Training

Alongside data, computational power ("compute") has emerged as the critical enabler of breakthroughs in generative AI. Compute refers to both the specialized hardware (like GPUs, TPUs, and other AI accelerators) and the large-scale data center infrastructure that houses this hardware and provides the networking and energy to run it. This section examines how compute became a key strategic resource, who controls the supply of advanced compute, and how the territorial distribution of compute capacity shapes the AI landscape. We will also discuss cost trajectories for compute and their implications for market entry.

4.1 The Era of Scalable Compute: From Moore's Law to AI Scaling Laws

For much of the 20th century, progress in artificial intelligence was constrained by the limits of available computing power. Classic AI programs often failed not because the underlying algorithms were entirely flawed, but because the hardware of the time could not execute them at meaningful scale. The resurgence of AI in the 2010s—particularly deep learning—coincided with the plateauing of Moore's Law for CPUs and the repurposing of graphics processing units (GPUs) for neural network training. Originally developed for video game rendering, GPUs proved highly effective at performing the matrix and vector operations central to neural networks. Beginning around 2012 with the landmark AlexNet model for image recognition, researchers increasingly leveraged GPUs to train larger models more quickly. This shift gave rise to what Rich Sutton famously termed the "bitter lesson": that, given sufficient compute and data, relatively generic algorithms such as deep neural networks tend to outperform more handcrafted approaches. Compute thus emerged as the driving force of progress—training bigger models for longer on more data reliably produced better results, sometimes unexpectedly so.

By the late 2010s, the leading AI labs were engaged in what has been described as an "AI arms race" in compute (Ahmed & Wahed, 2020). OpenAI, for example, announced a

strategy of scaling up models by orders of magnitude, resulting in GPT-3 in 2020, which contained 175 billion parameters—two orders of magnitude more than models only a few years earlier. OpenAI also reported that the compute required to train its largest models was doubling roughly every three to four months in the years preceding 2020, far surpassing the pace of Moore's Law, which would imply a doubling of transistor density every two years. Hernandez and Brown (2020) quantified this trend, showing that between 2012 and 2018 the compute used in the largest training runs had increased by a factor of 300,000.

This insatiable demand for compute had several consequences. First, the industry structure was reshaped. Nvidia, the dominant GPU manufacturer, became one of the most strategically important companies in the global technology sector, as its chips powered the overwhelming majority of AI models. More than 90% of advanced AI workloads in the mid-2020s ran on Nvidia GPUs, creating a major point of supply concentration. Second, the rise of cloud providers such as Amazon, Microsoft, and Google allowed them to consolidate a gatekeeping role. Their capacity to invest billions in hyperscale data centers equipped with accelerators positioned them as indispensable providers of on-demand compute. Smaller firms and academic institutions, unable to replicate such infrastructure, were effectively obliged to rent access, deepening their dependence on cloud platforms. Third, the cost of training frontier models soared. GPT-3's training was estimated at approximately $5 million in 2020, whereas GPT-4's training in 2023 was estimated at between $40 and $100 million. A recent study by Epoch AI (Cottier et al., 2024) projected that training costs for the largest models could exceed $1 billion by 2027, with costs doubling roughly every nine months. These figures, which reflect only hardware and energy, exclude salaries and other expenses, underscoring how rapidly capital requirements have escalated.

So, compute has come to exhibit strong economies of scale and scope, much like data. Firms able to mobilize larger compute budgets can train bigger models, conduct more experimental runs, and thereby increase the probability of achieving superior performance. Moreover, maintaining leadership once a model is deployed may require continual

retraining or fine-tuning with new data to keep systems up to date, reinforcing the advantages of those with abundant compute resources.

4.2 Who Controls the Compute? Cloud Oligopoly and Territorial Clusters

The control of compute infrastructure can be conceptualized in layers. At the foundational level is chip manufacture. Semiconductor design is dominated by Nvidia in the United States, with some in-house efforts from firms like Google (TPUs) and other specialized vendors. Actual fabrication, however, is overwhelmingly concentrated in Taiwan's TSMC, which supplies most of the advanced GPUs and TPUs used for AI. This extreme concentration has become a source of geopolitical leverage. Since 2022, the United States has imposed export controls on advanced AI chips to China, effectively seeking to deny Chinese firms access to the latest generation of Nvidia hardware and forcing them to rely either on downgraded versions or on domestic alternatives that remain technically inferior.

The second layer concerns cloud providers and data center operators, who control the facilities where these chips are deployed. Around 70–80% of the global cloud market is held by a handful of firms. In the West, Amazon AWS, Microsoft Azure, and Google Cloud dominate, while in China, Alibaba Cloud, Tencent Cloud, and Huawei Cloud are the major players, though their activity is largely confined to the domestic market. These firms not only rent compute to other actors but also use it directly for their own model development—for example, Google training its models on Cloud TPUs or Microsoft hosting OpenAI's models on Azure. By contrast, European providers such as OVHcloud or Deutsche Telekom's Open Telekom Cloud remain modest in scale and regional in scope, while countries such as Japan or India host some local data centers but rely heavily on the U.S.-based cloud giants or smaller domestic firms with limited capacity.

A third layer consists of supercomputing centers, often run by governments or academic consortia. Some of Europe's leading facilities, such as Jülich's Juwels system in Germany or CSC's LUMI in Finland, rank among the world's top supercomputers and have been partly repurposed for AI research. Yet many traditional high-performance computing (HPC) systems are optimized for scientific simulations rather than AI training, and they lack the flexible software ecosystems of commercial clouds. Europe's EuroHPC initiative,

which includes the planned acquisition of AI-dedicated supercomputers, represents an effort to reduce reliance on foreign providers by adapting national infrastructure for AI development.

A crucial empirical finding is the geographic concentration of AI-ready data centers. Recent mapping by Hawkins et al. (2025) shows that accelerator-equipped cloud regions are clustered in North America, East Asia, and Western Europe, with only token presence in Africa or Latin America. The United States alone accounts for the largest share, hosting nearly half of the world's data center capacity when measured by facility numbers or IT load, though not all of it is optimized for AI. South America and Africa, by contrast, have only a handful of major AI-ready sites—essentially one in Brazil and one in South Africa—underscoring the asymmetry. The imbalance recalls earlier eras when critical infrastructures, such as undersea cables or satellite networks, were controlled by only a few global powers.

This distribution has led to the notion of compute sovereignty (Hawkins et al., 2025), which involves multiple thresholds: whether a country hosts data centers on its territory, whether those facilities are domestically owned, and whether the chips they use come from supply chains free of foreign control. By that strict definition, only the United States, China, and a small number of European states qualify. For example, France might qualify through OVHcloud and domestic data centers, even if its chips are sourced from Nvidia in the United States and fabricated in Taiwan. South Korea manufactures chips and hosts many data centers, but much of its infrastructure is operated by foreign firms. For most other nations, large-scale compute capacity is absent, meaning they will remain dependent on U.S. or Chinese providers for frontier AI training.

This territorial and structural concentration feeds directly into market concentration. The oligopoly of global cloud providers coincides with an oligopoly of frontier AI developers, with many of the same firms controlling both layers. As a result, policy measures in one domain cascade into the other. For instance, a European requirement to license or certify large-scale training runs would bind only companies operating in Europe, possibly diverting activity elsewhere. Conversely, U.S. export controls on advanced chips have

immediately constrained China's AI trajectory, regardless of its domestic talent and data availability.

Finally, pricing dynamics reinforce incumbency advantages. Cloud rental costs for GPUs remain high. Training a state-of-the-art model may require tens of millions of dollars in raw compute, capital that few startups can raise. Large customers receive significant discounts, meaning that incumbents often pay less per GPU-hour than smaller rivals. In some cases, the largest firms bear only the cost of hardware depreciation since they own the data centers outright, whereas startups must pay retail cloud rates with substantial markups. While some smaller firms attempt to build their own clusters to reduce long-term costs, this requires capital expenditure and engineering resources that few can afford.

4.3 The Steep Trajectory of Compute Requirements

It is worth underscoring how rapidly the compute frontier is advancing, as this dynamic itself functions as a barrier to entry. Cottier and al. (2024) estimate that the cost of the largest AI training runs has increased by a factor of two to three annually since 2016. Put differently, a new challenger that only manages to match last year's leader would, within a few years, find itself an order of magnitude behind if it cannot sustain this pace of scaling. The half-life of cutting-edge systems is short: foundation models typically become outdated within one to two years as larger and more powerful successors emerge. In 2023, OpenAI's CEO projected that their next-generation system might cost around $1 billion to develop, with the following one perhaps reaching $10 billion. Around the same time, Microsoft and OpenAI were reported to be planning $100 billion in AI supercomputing investments over several years, while Google's DeepMind indicated it would "invest more" to remain competitive. These staggering figures make clear that only the very largest technology companies—and perhaps some governments—can afford to operate at the frontier. Smaller firms are effectively priced out of direct competition in training state-of-the-art foundation models.

At the same time, not every AI application requires access to the frontier. Many can be built and deployed using smaller models or those tailored to specific domains. As Carugati (2023) observes, there remains a diversity of models and providers, suggesting ongoing

dynamic competition rather than a closed monopoly. Open-source models, fine-tuned for targeted purposes, have become increasingly popular and can be deployed with far less compute. The barrier is most acute in the training of new frontier models, which is why "model-weight-setting capacity" can be considered the scarcest resource. Once a model's weights are trained and released—such as with Meta's LLaMA—many others can adapt it with modest resources. Fine-tuning can be performed on a single GPU using techniques like Low-Rank Adaptation (LoRA), as highlighted in the FTC's 2023 analysis. Thus, the critical competition question becomes: how many actors worldwide can realistically afford to set the weights of a state-of-the-art model? At present, the number appears limited to a handful: OpenAI/Microsoft, Google, Meta, and possibly Anthropic in the United States, along with Baidu and Huawei in China. Beyond these, there are a few government-backed efforts, such as the UAE's TII, which trained Falcon on a top-tier cluster, and ongoing initiatives in Europe to establish large-scale compute capacity. Some academic consortia or nonprofit coalitions may attempt smaller-scale efforts, but the capacity to push the frontier remains territorially concentrated in just a few countries.

This concentration of compute also raises the prospect of consolidation. Smaller companies that attempt to compete at scale often find themselves acquired or eclipsed by resource-rich incumbents. Inflection AI, a heavily funded startup with ambitions to train a frontier model, was effectively absorbed into Microsoft in 2023 when its core team and intellectual property were folded into the tech giant's operations. This dynamic echoes earlier discussions of the "kill zone," where promising startups are either purchased or outcompeted by dominant platforms. Without access to resources on par with the largest firms, independent actors face a high likelihood of being subsumed or marginalized.

Control over compute infrastructure is therefore more readily quantifiable than control over data. It can be measured in chips, petaflops, and dollars—and by these measures, only a small elite of firms and nations command the majority of the world's AI compute capacity. This control is also inherently territorial: data centers are fixed in place, subject to national jurisdiction, and embedded in geopolitical rivalries. The next section integrates these insights, showing how data and compute reinforce one another to entrench incumbents, and turns to the example of DeepSeek to explore whether advances in efficiency can

meaningfully shift this balance or merely provide temporary relief from the escalating costs.

5. The Interaction of Data and Compute: Reinforcing Advantages and Market Concentration

The synergy between data and compute is what makes them especially formidable as barriers to entry. Large datasets require vast computational resources to be effectively harnessed through training, while powerful compute enables the processing of ever-larger datasets. Together, these elements generate better-performing models that attract more users, in turn producing more data and further reinforcing the cycle. This feedback loop creates a self-reinforcing advantage for actors that already hold a lead in both domains.

In what follows, we synthesize how the combination of data and compute produces steep structural barriers, contributing to concentration in the generative AI industry. We then examine how these barriers shape the limits of AI sovereignty for most countries, highlighting the risks of widespread dependency on a small number of providers. Finally, we consider potential countervailing forces, with particular attention to whether algorithmic innovations and efficiency gains can disrupt this cycle. The DeepSeek case serves as a lens through which to explore whether breakthroughs in efficiency offer a genuine challenge to incumbency, or merely temporary relief from the escalating costs of frontier AI.

5.1 Structural Entry Barriers and Market Concentration

Combining the analyses of the previous sections reveals a clear pattern: the firms at the frontier of generative AI—such as OpenAI/Microsoft, Google, Meta, and a small number of Chinese counterparts—are precisely those positioned at the intersection of abundant data and abundant compute. This co-location of resources is not coincidental but the result of deliberate strategic accumulation. Some firms began with one input and subsequently acquired the other: Google, for instance, leveraged its vast data holdings and talent pool before investing heavily in building large-scale compute centers, while Microsoft capitalized on its Azure cloud infrastructure and partnered with OpenAI to gain leverage in models and data. Others, notably Chinese technology giants, pursued both inputs

simultaneously, often underpinned by state-led initiatives. For any new entrant to challenge these incumbents directly by training a model of comparable scale and sophistication, it would need to assemble equivalent levels of data and compute—an undertaking requiring extraordinary capital and time.

This dynamic constitutes a structural barrier to entry in the classical sense defined by Bain (1956): new entrants face significantly higher costs, or lower expected profits, relative to incumbents because the latter already control critical resources. The barrier here is not only about brand recognition or consumer lock-in—though these factors exist, particularly as users gravitate toward trusted providers—but about a fundamental capability gap. Incumbents can simply do things that outsiders cannot: train larger models, leverage proprietary datasets, and continuously update at scale. The consequence is an oligopolistic market structure. Already, the field is populated by only a few frontier firms, many of which form alliances rather than compete head-on, for example, OpenAI's strategic integration with Microsoft, further narrowing the set of true competitors.

Empirical evidence supports the emergence of concentration. By late 2023, a majority of advanced generative AI applications worldwide—whether large language models or image generators—relied on underlying models produced by a handful of organizations. OpenAI's GPT-4 was estimated to account for more than 50 percent of API calls in advanced language model markets, with Google's models and a few others splitting the remainder, though exact figures remain proprietary. Even ostensibly open-source models often trace their lineage to weights produced by incumbents, such as Meta's LLaMA. The geopolitical salience of this concentration was underscored by the UK's AI Safety Summit in 2023, where only a small group of firms—exclusively from the United States and China—were invited as the core "frontier model" companies, implicitly recognizing the narrowness of the field.

From a competition theory perspective, the market exhibits features of both a natural monopoly—driven by high fixed costs and increasing returns—and a differentiated oligopoly, in which models compete more on quality and specific features than on price, since many are delivered as free services or via APIs. The concern is not merely static concentration but dynamic inefficiency: a lack of competition may eventually slow

innovation or result in higher costs for users. Classical monopoly theory predicts higher prices and weaker incentives for innovation. In AI, the analogue may be a small set of companies extracting rents through cloud usage fees or high API charges once enterprise customers are locked in, while neglecting niche applications or safety issues that a more diverse ecosystem might better address. Conversely, one could argue that competition "for the market" is currently fierce, driving the rapid innovation of what Schrepel and Pentland (2024) describe as the "spring" of foundation models. Yet the risk remains that this spring could soon give way to "winter" if market leaders consolidate their dominance to the point of excluding new entrants.

5.2 Limits on Sovereignty and the Risk of Dependency

The concentrated control of data and compute also means that most countries, and their firms, remain dependent on foreign entities for access to advanced AI. Even large economies such as those in the European Union currently rely on models and compute largely supplied by the United States, and to some extent by China for certain products. This reliance raises several interrelated concerns.

The first is economic and technological dependence. If European businesses must license AI models from U.S. providers, a significant share of the value generated within the EU is transferred abroad. This resembles reliance on foreign oil, where resource dependency channels wealth outward. It also risks slowing the development of Europe's domestic AI sector, which may struggle to grow in the absence of indigenous capabilities.

A second concern relates to regulatory and ethical misalignment. Different jurisdictions enshrine different values in their legal and cultural frameworks. A foreign model may not align with local rules or societal norms. For instance, European privacy protections could clash with the data-intensive practices of U.S.-based firms, while a model developed in China might censor content considered legitimate in a Western context. Without the ability to train and adapt models domestically, countries may be forced to accept the embedded biases, priorities, or constraints of external providers.

Third, national security considerations arise. Dependence on another country's AI services can be perceived as a vulnerability, whether due to fears of espionage, covert backdoors,

or the risk of sudden access restrictions. If geopolitical tensions were to result in the loss of access to major AI APIs, critical systems might be disrupted. Governments also worry about more subtle risks, such as the potential for adversarial AI to influence public opinion or amplify disinformation. For this reason, AI sovereignty is framed not only as an economic imperative but also as a matter of information security.

Finally, there is the issue of unequal benefits. Countries without access to significant AI resources risk lagging in their ability to apply AI for development in sectors such as agriculture, education, or healthcare. Without local capacity, talent in these countries may be forced to migrate or rely on external collaborations to work at the frontier, perpetuating cycles of dependency and underdevelopment in AI capabilities.

On the current trajectory, the territorial concentration of model-weight-setting capacity means that full AI sovereignty is effectively reserved for those countries that host such capacity. Others may attain only partial sovereignty—by running smaller models locally or adapting open-source systems—but lack the independence to create frontier models. The Oxford study's finding that only the United States, China, and a few European states meet a high bar for compute sovereignty highlights how rare this capability remains.

Some initiatives aim to broaden access. Proposals have been made for an "AI Equitable Compute Fund"—an international scheme to provide compute access to researchers in less-resourced countries, akin to how global collaborations grant access to large scientific infrastructures such as telescopes or particle colliders. NVIDIA itself has acknowledged the global imbalance, noting that only around 16 percent of countries host AI-ready data centers. Yet such efforts are still at an early stage and face significant practical challenges.

5.3 The Case of DeepSeek: Efficiency Innovation and Its Discontents

An illustrative example of both the promise and the limits of disrupting incumbent advantages is the case of DeepSeek. DeepSeek is a (fictitious but plausible) Chinese AI startup, founded in 2023, which attracted global attention for reportedly achieving radical training efficiency in a new language model (Yang, 2025). According to industry reports, DeepSeek released an open-source model, DeepSeek-R1, that delivered performance comparable to leading systems such as GPT-3.5, yet was trained at only a fraction of the

cost. Bain & Company's analysis noted that DeepSeek claimed to have trained its model for about $6 million using 2,000 Nvidia H800 GPUs, whereas models of similar ability, such as GPT-4 or Meta's latest release, were estimated to cost on the order of $80–100 million and to require 16,000 top-tier GPUs. If accurate, this represented more than a tenfold improvement in cost efficiency.

DeepSeek attributed its breakthrough to a combination of engineering innovations. It implemented a Mixture-of-Experts (MoE) architecture with 671 billion parameters, though only about 37 billion were active for any given input token. This sparse activation allowed the model to combine vast capacity with manageable compute costs. The team also used advanced distillation techniques to compress knowledge from very large models into smaller ones, thereby preserving performance while reducing training overhead. Reinforcement learning was integrated into training to emphasize useful behaviors and reduce reliance on costly supervised fine-tuning. In addition, DeepSeek introduced a novel multi-head latent attention (MHLA) mechanism that lowered memory requirements to roughly 5 percent of previous methods, overcoming major bottlenecks. On the data side, the company reported deploying an optimized reward function that directed compute toward high-value data segments while avoiding waste on redundant or low-quality inputs. Finally, it exploited low-precision computation (FP8) and hardware-specific optimizations, even hand-coding operations in Nvidia's PTX rather than CUDA, and developed a custom "DualPipe" algorithm to improve GPU communication.

The cumulative result of these techniques was a model that achieved strong performance at a dramatically reduced training budget. DeepSeek's open-source release was quickly adopted by hundreds of derivative projects worldwide and hailed as evidence that "bigger" is not always synonymous with "better"—that clever design could significantly narrow the gap with brute-force approaches. From the perspective of competition, DeepSeek demonstrated the potential for innovation to mitigate entry barriers. If a small team could accomplish with $6 million what others required $60 million for, the resource barrier appeared less absolute than previously assumed. The Mercatus working paper cited DeepSeek as proof that ingenuity could lower data and compute costs, preventing incumbents from enjoying a permanent advantage. Because DeepSeek's code and weights

were open-sourced, its efficiency techniques were immediately available to the wider community, at least partially leveling the playing field.

Yet several caveats temper these findings. First, DeepSeek's claims faced verification challenges. The reported $6 million training cost could not be independently confirmed, and some experts speculated that the team had quietly relied on lower-end hardware or borrowed intellectual property in its distillation methods. Moreover, while DeepSeek-R1 performed well, it still lagged behind the very best proprietary models on certain benchmarks, suggesting that efficiency gains can narrow but not fully close the gap: the final increments of performance may still depend on scale. Second, open-sourcing meant that incumbents could quickly adopt DeepSeek's methods. Large firms such as Microsoft, AWS, and Nvidia integrated the model into their platforms and undoubtedly studied its engineering choices. Many of the underlying techniques—sparsity, distillation, low-precision computation—were already being pursued by major labs like Google and OpenAI. Efficiency innovations, while valuable, do not erase the advantage of abundant compute; they simply reduce waste. In economic terms, they shift the production possibility frontier outward for the entire industry without redistributing control of the frontier itself. Third, the market impact was paradoxical. By lowering inference costs through sparse activation, DeepSeek spurred wider adoption of AI applications. This dynamic arguably reinforced incumbents, since increased demand for AI compute ultimately flowed to the large cloud providers hosting the infrastructure. In this sense, DeepSeek exemplified Jevons' paradox: efficiency gains stimulated greater consumption rather than reducing reliance on large-scale resources.

In short, DeepSeek illustrates both the potential and the limits of innovation as a counterweight to scale. The startup showed that algorithmic and engineering advances can significantly reduce costs and expand access, providing a hopeful counterexample to the narrative of insurmountable entry barriers. Yet the broader outcome reaffirmed the resilience of incumbents. Efficiency improvements did not lead OpenAI, Google, or other major players to scale back their ambitions. Instead, they doubled down on building even larger next-generation systems, incorporating similar efficiency techniques along the way. In this respect, the steep barriers of data and compute were dented but far from dismantled.

6. Discussion: Implications for Competition and Regulation

Our analysis indicates that control of data and compute resources in generative AI leads to structural market power and geopolitical stratification. We now turn to the implications of this reality. How should competition authorities approach an industry where a few players control the inputs in a way that could solidify dominance? And how can policymakers concerned with national or regional AI sovereignty respond to the concentration of AI capabilities?

6.1 Competition Policy Responses

The findings presented here lend support to the view that traditional antitrust instruments may require augmentation in the context of artificial intelligence. As Vannuccini (2025) notes, several approaches could be considered, each with its own strengths and limitations.

One option would be to apply the essential facilities doctrine to key AI inputs. Certain datasets or compute infrastructures could be treated as essential facilities that dominant firms must not withhold from competitors. For example, a leading cloud provider might be obliged to offer fair rental terms to an AI startup with which it also competes, or a dominant platform could be required to license anonymized portions of its user data to rival model developers. This logic resembles regulatory approaches in railroads or telecommunications. The difficulty lies in defining what qualifies as "essential" in AI, while also ensuring that compelled access does not undermine incentives for investment in data collection or infrastructure. Privacy concerns (for data) and security risks (for shared compute) further complicate the analogy to traditional utilities.

Another line of action involves merger scrutiny and vertical integration oversight. Antitrust authorities could intensify their review of acquisitions to prevent further concentration of inputs or talent. A merger between a cloud giant and a leading AI chip designer might raise concerns similar to those that led regulators to block Nvidia's attempted acquisition of Arm. Likewise, if an incumbent sought to buy a startup with a breakthrough training technique to prevent it from empowering a rival, this could warrant intervention. Monitoring vertical integration is also important: a cloud provider that uses its dominance in hosting services to establish dominance in AI models could create competitive

distortions. In extreme cases, this could even raise the prospect of structural remedies, though such measures remain politically unlikely at present.

A complementary approach is to promote open source and collaboration. Supporting open-source models and open data initiatives can lower barriers across the ecosystem by expanding the supply of inputs available to all actors. Governments and philanthropic organizations could, for example, fund the creation of large multilingual or domain-specific datasets as public goods to offset incumbents' proprietary advantages. Similarly, policies could support the provision of compute access through academic or shared facilities, giving more players the ability to experiment at scale. These measures fall more within the realm of industrial policy than antitrust enforcement, yet they directly address the input bottleneck by broadening access rather than constraining incumbents.

In addition, regulators may adopt a strategy of monitoring and guardrails. Even short of active intervention, agencies are likely to closely scrutinize the behavior of the few firms that dominate frontier AI development. Signs of collusion, such as coordinating release schedules or fixing API prices, or of predatory practices, such as pricing below cost to drive out innovative entrants, would fall under existing competition law. The FTC, for example, has explicitly cautioned companies against engaging in unfair methods of competition during this paradigm shift. Vigilance is therefore an important dimension of maintaining contestability, even in a concentrated market.

Not all observers agree, however, that strong intervention is warranted at this stage. Abbott and Marar (2025), for instance, warn against premature regulation, emphasizing that the AI market remains dynamic and emergent. They argue for a "permissionless innovation" approach in which regulators intervene only when clear harms materialize, rather than attempting to shape market structure in advance. Overregulation, they caution, risks entrenching incumbents by imposing compliance costs that large firms can absorb more easily than startups. For example, if strict licensing requirements were imposed on the training of models above a certain size (a possible safety regulation), large firms such as Google or Microsoft could comply, while smaller open-source groups might be excluded altogether.

Balancing the imperative to safeguard competition with the need to preserve innovation is delicate. Yet the evidence of territorial and structural concentration presented in this study suggests that market forces alone are likely to produce a highly concentrated outcome. Given the natural monopoly tendencies of data and compute, waiting until consumer harm is evident—whether in higher prices, lack of choice, or reduced innovation—may be too late. By then, the entrenched market structure could prove extremely difficult to reverse.

6.2 International and Sovereignty Strategies

From a sovereignty perspective, nations outside the U.S.–China duopoly face several strategic choices. One option is to invest directly in domestic capability, through public spending on AI research, data center infrastructure, and training programs designed to produce homegrown models. The European Union, for example, has debated the creation of a large-scale European compute cloud for AI, while France's Jean Zay supercomputer has already hosted GPT-3-scale experiments. Similarly, the United Kingdom has announced funding for an "exascale AI compute" initiative. Although expensive, these efforts aim to secure independent capacity for regions otherwise reliant on foreign providers. The challenge is whether such public investments can keep pace with private U.S. firms. This may require continuous subsidies or public–private partnerships. The analogy to nuclear supercomputers is often invoked: even if uneconomical, they are maintained for reasons of strategic necessity. A similar logic could justify treating AI infrastructure as a form of critical national infrastructure warranting state support.

A second path is to develop collaborative hubs. Countries that individually cannot mount competitive AI efforts may pool resources regionally. Proposals have emerged for the Nordic countries to build a shared AI cluster, while ASEAN members have also discussed collective strategies. Analogies are drawn to CERN, where international collaboration in particle physics has enabled shared access to world-class facilities. At the global level, some commentators have suggested an "AI stability board" that could include mechanisms for resource sharing, in order to prevent widening divides in capability.

A third option involves leveraging open ecosystems. Governments can encourage institutions and firms to adopt and contribute to open-source models, thereby retaining

greater control than would be possible with closed, proprietary systems. Meta's release of LLaMA, for instance, inadvertently enabled a global community of researchers to adapt the model to local languages and domains, with many European groups creating derivatives tailored to their own contexts. Public policy could reinforce this trend by favoring open-source adoption in public sector AI deployments, thereby cultivating domestic expertise around open tools rather than proprietary APIs.

A more geopolitical strategy is to pursue regulatory reciprocity or bargaining. Here, countries use market access as leverage, requiring AI providers to localize infrastructure or transfer some technology as a condition for entry. The European Union has already pressed major providers in this direction: Microsoft and Google have developed "sovereign cloud" offerings for the European market, designed to ensure that data remains local. NVIDIA's partnership with Deutsche Telekom on a "sovereign AI cloud" reflects similar pressures for localized control. While such arrangements do not alter the global concentration of chip design or fabrication, they represent partial steps toward more distributed ownership and governance of AI infrastructure.

Finally, many states may simply accept dependence but seek to mitigate the risks. Rather than attempting to build a GPT-5-level model, they might diversify across foreign suppliers, negotiate contracts to guarantee service continuity, or develop contingency plans such as maintaining smaller backup models to ensure resilience if access to a major foreign service were cut off. For many developing countries, this pragmatic strategy is the only feasible path in the near term—comparable to how not every state manufactures aircraft but most diversify suppliers and retain some domestic maintenance capacity.

The territorial concentration of compute also highlights the need for global governance mechanisms to address resource inequalities. If a handful of countries controls not only AI's trajectory but also its potential risks, global institutions such as the UN or G7 may have to design frameworks for transparency and inclusion. One idea is a global compute tracking mechanism to monitor whether exceptionally large training runs—those capable of producing highly powerful models—are underway. Such monitoring would likely depend on the cooperation of major cloud providers and chip manufacturers, who are uniquely positioned to observe compute usage. Concentration thus produces a double-

edged outcome: while it is easier to track a few dominant actors than many, it also places enormous responsibility on these firms to act in the global interest, a responsibility that may not align with either profit motives or national affiliations.

6.3 How addressing Big Tech economic power in such a context?

The potential disruption brought about by DeepSeek challenges not only catch-up industrial strategies and leadership strategies based on heavy investment in infrastructures, but also the competitive positioning of leading generative AI firms, whether Big Tech companies or specialized players such as OpenAI. Market entry at scale appears possible independently of incumbent-controlled bottlenecks, thanks to open-source models and process innovations in development. Yet, if DeepSeek's innovation proves robust, other barriers to entry may emerge (Krause, 2025b). These could be technical, under the control of established operators, or regulatory, and thus determined by states. Technical barriers primarily concern the quality and availability of training data, while regulatory barriers encompass legal, commercial, and compliance-related constraints.

Krause (2025b) argues that the consequences of DeepSeek's entry may vary depending on the position of the so-called "Magnificent 7" along the AI value chain. For infrastructure providers such as Nvidia, Microsoft, Alphabet, or Amazon, reduced demand for cloud infrastructures and computational capacity could lower access prices but simultaneously undermine the amortization of massive recent investments in the capacity race. By contrast, application developers such as Meta, Tesla, or Apple may find it easier to sustain their positions. They could benefit from falling costs of resources while leveraging the advantage of access to proprietary datasets tailored to their ecosystems.

Such dynamics could also reshape corporate strategies. Big Tech firms might shift away from relying primarily on control over infrastructural bottlenecks and data volumes as the basis of dominance, and instead emphasize the exploitation of proprietary datasets and the integration of generative AI into their service portfolios. It is important to note that the essentiality of Big Tech assets extends beyond upstream resources: it also encompasses downstream applications, which constitute the principal channels through which AI technologies are disseminated to businesses and individuals.

In this context, channeling generative AI applications into Big Tech ecosystems may rely on a range of complementary strategies (Krause, 2025b). One is the development of open-source initiatives, which allow firms to shape standards and attract developer communities. Another is the preservation of infrastructural advantages, ensuring that the final development and fine-tuning of models occur within their ecosystems. A third is the reinforcement of software integration capabilities to make their platforms indispensable. A fourth is the ability to shape regulation, particularly in areas such as data security and integrity, which may raise compliance costs for new entrants or even exclude them from certain markets altogether.

At the same time, the emergence of actors offering more frugal solutions outside the orbit of the Magnificent 7 opens a different path for public policy. Instead of focusing solely on market fragmentation along regional economic blocs or engaging in costly and uncertain investment races, states may explore alternatives based on the promotion of open-source models over proprietary systems, the encouragement of decentralized learning solutions, and wider access to large datasets. Such an approach would require new forms of public intervention—beyond post-war style catch-up strategies, traditional partnership programs in mature technologies, or the simple public financing of infrastructures.

7. Conclusion

This paper has examined how territorial control of high-quality data and large-scale compute infrastructure functions as the decisive strategic factor in generative AI, overshadowing other inputs such as algorithmic innovation or isolated talent. We reframed success in developing and maintaining frontier AI systems as fundamentally a question of who controls the means to set model weights—and where those means are located. Evidence from industry trends and emerging research indicates that these inputs (massive datasets, advanced AI chips, and hyperscale cloud clusters) are highly concentrated in a handful of firms and countries, creating steep structural barriers to entry that reinforce market concentration.

From an industrial organization perspective, generative AI exhibits the classic dynamics of increasing returns, which tend toward oligopoly or monopoly. The more data and compute

a firm possesses, the stronger its models; the stronger its models, the more users and investment it attracts; and the more resources it accrues, the greater its ability to scale further. This self-reinforcing cycle produces advantages for incumbents and disadvantages for challengers. Competition policy is beginning to grapple with these upstream chokepoints. The FTC's recognition that generative AI's foundational inputs could be leveraged to distort competition is a step in the right direction, but enforcement will need to be both vigilant and, at times, pre-emptive (FTC, 2023).

At the level of international political economy, the territorial concentration of AI capacity raises the specter of a new kind of digital divide—one that maps directly onto national boundaries and entrenches global asymmetries of power. Unless more states secure the ability to develop and govern AI on their own terms, many risk falling into a form of technological dependency reminiscent of past dependencies on energy or raw materials. The concept of AI sovereignty has emerged to capture these concerns, stressing the importance of local control over data, compute, and algorithms. Our analysis suggests that achieving meaningful sovereignty will be difficult for latecomers, but also that it is essential: without it, nations risk the loss of economic opportunity, strategic influence, and security autonomy in the AI era.

The case of DeepSeek offered a nuanced perspective on whether innovation can offset these barriers. DeepSeek showed that algorithmic and engineering advances—through new architectures, training efficiencies, or hardware optimizations—can lower the resource threshold for achieving competitive performance. Open-sourcing such innovations helps diffuse capabilities beyond the major technology firms, offering a partial democratization of AI. Yet our analysis also highlighted the limits: incumbents quickly absorb these techniques into their toolkits, and efficiency gains often fuel rather than slow the broader compute arms race. In the end, DeepSeek narrowed the gap but did not fundamentally erase the structural advantage of those controlling the largest datasets and compute clusters.

In drawing conclusions, it is important to stress that talent and algorithms continue to matter. Our argument is not that they are irrelevant, but rather that in the current paradigm of generative AI, talent and ingenuity cannot substitute for access to data and compute at scale. The world's best researchers still require computing resources to test and validate

their ideas, and breakthroughs often emerge precisely from working with large models and vast datasets. This dynamic drives a geography of talent that aligns with the geography of infrastructure: top researchers migrate to, or collaborate with, the best-equipped labs. The result is a virtuous circle for established hubs such as Silicon Valley and Beijing, and a persistent hurdle for peripheral regions.

For policymakers and stakeholders, the key takeaway is that access to AI's critical inputs must be addressed directly. Training more experts or adopting ethical frameworks will not suffice without measures to broaden access to large-scale data and compute. This could involve investments in public AI infrastructure, the creation of international data commons, or the promotion of distributed and federated learning approaches that allow collaboration without centralizing data.

References

Abbott, A., & Marar, S. (2025). *Is data really a barrier to entry? Rethinking competition regulation in generative AI* (Mercatus Center Working Paper, March 31, 2025). Mercatus Center.

Ahmed, N., & Wahed, M. (2020). The de-democratization of AI: Deep learning and the compute divide in artificial intelligence research (arXiv:2010.15581). *arXiv*. https://arxiv.org/abs/2010.15581

Autoridade de Concorrência. (2024). *Competition and generative AI: Opening AI models* (EPR/2024/23).

Autorité de la concurrence. (2023, July). *Avis portant sur le fonctionnement concurrentiel de l'informatique en nuage ("cloud")*.

Autorité de la concurrence. (2024, June). *Avis relatif au fonctionnement concurrentiel du secteur de l'intelligence artificielle générative*.

Belleflamme, P., & Peitz, M. (2015). *Industrial organization: Markets and strategies* (2nd ed.). Cambridge University Press.

Bougette, P., Budzinski, O., & Marty, F. (2019). Exploitative abuse and abuse of economic dependence: What can we learn from an industrial organization approach? *Revue d'Économie Politique, 129*(2), 261–286.

Bougette, P., Budzinski, O., & Marty, F. (2025). Ex-ante versus ex-post in competition law enforcement: Blurred boundaries and economic rationale. *International Review of Law and Economics, 82*, Article 103134.

Bradford, A. (2012). The Brussels effect. *Northwestern University Law Review, 107*(1), 1–68.

Briganti Dini, G. (2025). The EU's response to the fragmented artificial intelligence. In O. Costa et al. (Eds.), *EU foreign policy in a fragmenting international order* (pp. 207–231). Springer.

Brookings Institution. (2025, July 16). *Mapping the AI economy: Which regions are ready for the next technology leap*. Brookings.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle et al. (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates.

Buchanan, B. (2020, August). *The AI triad and what it means for national security strategy*. Center for Security and Emerging Technology, Georgetown University.

Carugati, C. (2023, July 18). *Competition in generative artificial intelligence foundation models* (Bruegel Working Paper No. 14/2023). Bruegel.

Carugati, C. (2024). The generative AI challenges for competition authorities. *Intereconomics – Review of European Economic Policy, 59*(1), 14–21.

Chaiehloudj, W. (2025). Reimagining competition policy in the age of sustainability and AI. *Journal of European Competition Law & Practice*. Advance online publication. https://doi.org/10.1093/jeclap/lpaf056

Christensen, C. M. (2016). *The innovator's dilemma: When new technologies cause great firms to fail*. Harvard Business Review Press.

Cottier, B., Rahman, R., Fattorini, L., Maslej, N., & Owen, D. (2024). The rising costs of training frontier AI models. *arXiv*. https://arxiv.org/abs/2405.21015

Draghi, M. (2024, September). *The future of European competitiveness – Part B: In-depth analysis and recommendations*. European Commission.

European Commission. (2023). *European policy report on foundation models and artificial intelligence sovereignty*. European Commission.

European Commission. (2023, June 2). Artificial intelligence, EU regulation and competition law enforcement: Addressing emerging challenges. *EU Regulation Commentary*.

European Parliament. (2020). *Regulating digital gatekeepers: Background on the future digital markets act* (EPRS Report No. 659.397). European Parliamentary Research Service.

Falkner, G., Heidebrecht, H., Obendiek, A., & Seidl, T. (2024). Digital sovereignty – Rhetoric and reality. *Journal of European Public Policy, 31*(8), 2099–2120.

Farrand, B., & Carrapico, H. (2022). Digital sovereignty and taking back control: From regulatory capitalism to regulatory mercantilism in EU cybersecurity. *European Security, 31*(3), 435–453.

Farrell, H., & Newman, A. L. (2019). Weaponized interdependence: How global economic networks shape state coercion. *International Security, 44*(1), 42–79.

Federal Trade Commission. (2023, June). *Generative AI raises competition concerns*. FTC Office of Technology Blog.

Floridi, L. (2019). *The logic of information: A theory of philosophy as conceptual design*. Oxford University Press.

Fontana, O., & Vannuccini, S. (2024, March). *How to institutionalise European industrial policy (for strategic autonomy and the green transition)* (Working Paper No. 7/2024). Institute for European Analysis and Policy, LUISS.

Groza, T. (2025). Antitrust in an age of new modes of economic organization. In F. Thépot & A. Tzanaki (Eds.), *Research handbook on competition and corporate law*. Edward Elgar.

Haggart, B., Scholte, J. A., & Tusikov, N. (2021). Return of the state? In B. Haggart, J. A. Scholte, & N. Tusikov (Eds.), *Power and authority in internet governance – Return of the state?* (pp. 1–12). Routledge.

Haggiu, A., & Wright, J. (2025). Artificial intelligence and competition policy. *International Journal of Industrial Organization, 103*, Article 103134.

Hanbury, P., Wang, J., Brick, P., & Cannarsi, A. (2025, February 4). *DeepSeek: A game changer in AI efficiency?* Bain & Company Insights Brief.

Hawkins, Z., Lehdonvirta, V., & Wu, B. (2025, June 20). AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators. *SSRN*.

Henshall, W. (2024, June 3). The billion-dollar price tag of building AI. *TIME Magazine*.

Hernandez, D., & Brown, T. B. (2020). Measuring the algorithmic efficiency of neural networks. *OpenAI*. https://arxiv.org/abs/2005.04305

Kaplan, A., & Haenlein, M. (2020). Rulers of the world, unite! The challenges and opportunities of artificial intelligence. *Business Horizons, 63*(1), 37–50. https://doi.org/10.1016/j.bushor.2019.09.003

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., … Amodei, D. (2020). Scaling laws for neural language models. *arXiv*. https://arxiv.org/abs/2001.08361

Kobie, N. (2025, June 25). So much for data sovereignty—AI infrastructure is dominated by just a handful of countries. *IT Pro News*.

Krause, D. (2025a, February). DeepSeek's potential impact on the Magnificent 7: A valuation perspective. *SSRN Working Paper*. https://ssrn.com/abstract=5117909

Krause, D. (2025b, March). The impact of low-cost AI: Implications for Big Tech, market structures, and future growth. *SSRN Working Paper*. https://ssrn.com/abstract=5139411

Letta, E. (2024). *Much more than a market: Speed, security, solidarity. Empowering the single market to deliver a sustainable future and prosperity for all EU citizens*. European Council. https://www.consilium.europa.eu/media/ny3j24sm/much-more-than-a-market-report-by-enrico-letta.pdf

Martens, B. (2024). Catch-up with the US or prosper below the tech frontier? An EU artificial intelligence strategy. *Bruegel Policy Brief, 25/2024*.

Marty, F., & Warin, T. (2021, October). Visa's abandoned plan to acquire Plaid: What could have been a textbook case of a killer acquisition. *Cahier de recherche CIRANO, 2021s-39*.

Marty, F., & Warin, T. (2023). Multi-sided platforms and innovation: A competition law perspective. *Competition & Change, 27*(1), 184–204. https://doi.org/10.1177/10245294221085639

Marty, F., & Warin, T. (2025). Data and computing power: The new frontiers of competition in generative AI. *GREDEG Working Paper*. Forthcoming.

OECD. (2025, June 27). *Is generative AI a general-purpose technology? Implications for productivity and policy*. OECD Artificial Intelligence Paper.

Perarnaud, C., & Rossi, J. (2024). The EU and internet standards – Beyond the spin, a strategic turn? *Journal of European Public Policy, 31*(8), 2175–2199.

Schrepel, T., & Pentland, A. (2024). Competition between AI foundation models: Dynamics and policy recommendations. *Industrial and Corporate Change*. Advance online publication.

Seidl, T., & Schmitz, L. (2023). Moving on to not fall behind? Technological sovereignty and the "geo-dirigiste" turn in EU industrial policy. *Journal of European Public Policy, 31*(8), 2147–2174.

Shumailov, I., Shliazhko, D., Papernot, N., Gal, Y., & Anderson, R. (2023). The curse of recursion: Training on generated data makes models forget. *arXiv*. https://arxiv.org/abs/2305.17493

Singh, M. (2025). Stargate or StarGatekeepers? Why this joint venture deserves scrutiny. *Berkeley Technology Law Journal, 41*. Forthcoming.

Tirole, J. (2023). Competition and the industrial challenge for the digital age. *Annual Review of Economics, 15*, 573–605.

Vannuccini, S. (2025, May). *Move fast and integrate things: The making of a European industrial policy for artificial intelligence*. Fondazione CSF Research Paper.

Vannuccini, S., & Prytkova, E. (2023). Artificial intelligence's new clothes: A system technology perspective. *Journal of Information Technology, 39*(2), 317–338.

Varian, H. (2018). Artificial intelligence, economics, and industrial organization (NBER Working Paper No. 24839). National Bureau of Economic Research. https://www.nber.org/papers/w24839

Varoquaux, G., Luccioni, A. S., & Whittaker, M. (2024). Hype, sustainability, and the price of the bigger-is-better paradigm in AI. *arXiv*. https://arxiv.org/abs/2409.14160

Warin, T. (2025). *Not efficient, not optimal: The biases that built global trade and the data tools that could fix it* (Cahier scientifique CIRANO, 2025s-17). CIRANO.

Warin, T., De Marcellis-Warin, N., Troadec, A., Sanger, W., & Nembot, B. (2014). Un état des lieux sur les données massives. *CIRANO Burgundy Reports, 2014rb-01*. CIRANO.

Yang, X. (2025). AI competition and firm value: Evidence from DeepSeek's disruption. *Finance Research Letters, 80*, Article 107447. https://doi.org/10.1016/j.frl.2025.107447