BIS Papers
No 154

The AI supply chain

by Leonardo Gambacorta and Vatsala Shreeti

Monetary and Economic Department

March 2025

JEL classification: E31, J24, O33, O40.

Keywords: artificial intelligence, generative AI, market structure, big techs, competition, financial stability, cyber risk.

The views expressed are those of the authors and not necessarily the views of the BIS.

This publication is available on the BIS website (www.bis.org).

ISSN 1682-7651 (online)
ISBN 978-92-9259-843-3 (online)

The AI supply chain

Leonardo Gambacorta and Vatsala Shreeti*

Abstract

The rapid advancement of artificial intelligence (AI) relies on a complex supply chain comprising five key layers: hardware, cloud infrastructure, training data, foundation models and AI applications. This paper examines the market structure of each layer and highlights the economic forces shaping them: rapid technological change, high fixed costs, economies of scale, network effects and, in some cases, strategic behaviour by dominant firms. We also highlight the expanding influence of big tech companies across the AI supply chain. We discuss the challenges for consumer choice, innovation, operational resilience, cyber security and financial stability.

JEL classification: E31, J24, O33, O40.

Keywords: artificial intelligence, generative AI, market structure, big techs, competition, financial stability, cyber risk.

* Leonardo Gambacorta and Vatsala Shreeti work at the Bank for International Settlements (BIS). We would like to thank Iñaki Aldasoro, Giulio Cornelli, Juan Carlos Crisanto, Sebastian Doerr, Johannes Ehrentraud, Jon Frost, Gaston Gelos, Doh Shin Jeon, Priscilla Koo Wilkens, Anton Korinek Fernando Perez Cruz, Jermy Prenio, Hyun Song Shin and Jeffrey Yong for valuable comments and suggestions. The views expressed in this paper are those of the authors and not necessarily those of the BIS.

1. Introduction

When a user asks an artificial intelligence (AI) application on their phone a question, and receives a cogent, detailed response, the result often looks like magic. But to make this interaction possible, a long series of necessary steps had to take place involving different markets and players in countries around the world. The rapid advancement of AI thus depends on an increasingly complex supply chain, with multiple layers of technology that work together to power the AI applications that we interact with daily. Analysing the AI supply chain, and the economic forces that shape it, is critical to understanding how AI impacts social and economic welfare, innovation, operational resilience, cyber risk and financial stability.

This paper contributes to the literature by providing a comprehensive analysis of the market structure, economic forces and challenges along the five key input layers of the AI supply chain: hardware, cloud infrastructure, training data, foundation models and AI applications.[1] The first layer consists of specialised hardware – most notably specialised microprocessors or chips that perform the complex computations needed for AI model training and inference. The second layer is cloud computing, which provides the infrastructure required to build, store and use AI models. Training data are the next input layer: AI models feed on vast data sets, which include everything from text to images and videos, typically sourced from both public and proprietary repositories. Foundation models, the fourth layer, are large, pre-trained models that can be adapted to many different uses. These form the base for the last layer: downstream AI applications.

The market structure and the economic forces shaping the layers are different. The first two layers of the AI supply chain (hardware and cloud) are characterised by high fixed costs, economies of scale and scope, high switching costs and consumer inertia, and network effects. The substantial investment required for developing and maintaining AI infrastructure creates barriers to entry and favours larger firms with significant financial resources. In contrast, the market for training data, foundation models and downstream AI applications may currently be more contestable but could still be prone to "winner takes all" dynamics. In general, for foundation models and AI applications, there seems to be competition "for" the market, but "within" each market only a few firms tend to dominate.

At the same time, big technology companies (big techs) are expanding their footprint across the entire AI supply chain. They are active in every layer of the supply chain and are engaging in vertical integration, as well as the bundling and tying of different services and exclusive partnerships. In particular, big tech companies are well positioned to leverage their existing dominance in data and cloud services to gain advantages in both upstream and downstream layers. For instance, several big tech companies are producing their own AI hardware, acquiring several data-owning firms, building their own foundation models and even securing their own supply of nuclear power to fuel their data centres (CNBC (2024)). Driven by the potential consequences

[1] We do not analyse the human resources aspect of the AI supply chain. Specialised talent is in high demand throughout the supply chain and the ensuing labour market dynamics can affect the overall market structure and welfare outcomes. We leave this to future research.

of a concentrated AI supply chain, several jurisdictions are examining the economic structure of the AI market and the actions of its leading firms.

Market concentration in the provision of AI, and the expanding footprint of a few large technology companies can have broad consequences. Even beyond traditional competition policy concerns, concentrated AI provision can impact operational resilience, cyber security, financial stability and the direction of future innovation. Effective regulatory oversight of AI in this context is a challenging task given the diverse range of stakeholders and markets involved in the AI supply chain across national borders. As such, international and domestic cooperation between different competent authorities will be key.

The paper is organised as follows. In the next section, we describe each layer of the AI supply chain. In Section 3, we delve into the market structure of each layer, illustrating the economic forces shaping market structure as well as the strategic actions of dominant firms. In Section 4, we then trace the footprints of big tech companies in the AI supply chain. Section 5 highlights the challenges of a concentrated AI supply chain and policy considerations. Section 6 concludes by summarising the main findings.

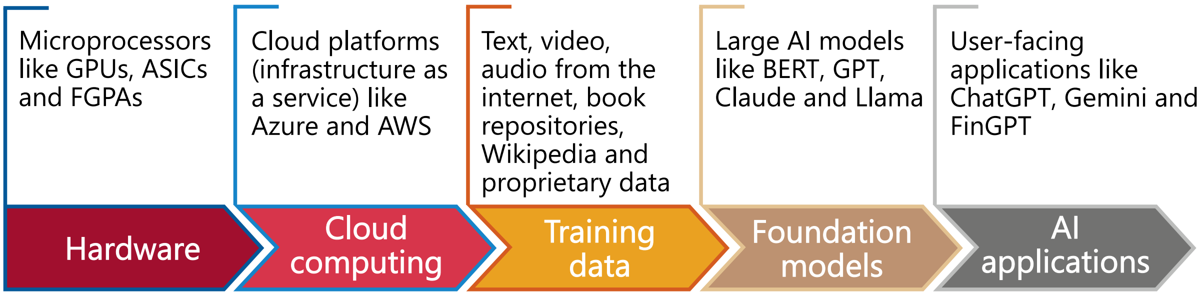
2. The AI ecosystem

The typical AI application is based on several layers of technology (Narechania and Sitaraman (2024); Hagiu and Wright (2025)). The key components of the supply chain are represented in Graph 1: (i) computing hardware; (ii) cloud computing infrastructure; (iii) training data; (iv) foundation models; and (v) user-facing AI applications. Consider a popular AI application like ChatGPT. It is based on a generative pre-trained transformer (GPT), a foundation model. GPT is trained with a vast corpus of text data stored on Microsoft Azure, a cloud computing service. For AI applications, Azure relies on several high-end microprocessors, typically graphics processing units (GPUs) that are produced by specialised providers like Nvidia.

In the first **hardware** layer, the most critical components are specialised AI chips such as field-programmable gate arrays (FPGAs), application-specific integrated circuits (ASICs) and GPUs. These chips excel at performing complex computations in parallel, which is particularly beneficial for AI models that rely on massive data sets. Parallel computing significantly enhances processing speed by handling thousands of computations simultaneously. Among these AI chips, GPUs are the most widely used, especially during the training phase of AI model development (Khan and Mann (2020)). Training large AI models, such as large language models (LLMs), typically requires several hundred, if not thousands of GPUs. Even for model inference, a substantial number of GPUs are necessary. For instance, some estimates suggest that it takes eight GPUs for Microsoft Bing to answer a single question in less than one second (CNBC (2023)).

The AI supply chain

Graph 1


Microprocessors like GPUs, ASICs and FGPAs
Cloud platforms (infrastructure as a service) like Azure and AWS
Text, video, audio from the internet, book repositories, Wikipedia and proprietary data
Large AI models like BERT, GPT, Claude and Llama
User-facing applications like ChatGPT, Gemini and FinGPT
Hardware
Cloud computing
Training data
Foundation models
AI applications

Source: authors' elaboration.

The next layer of the AI supply chain is **cloud computing**. In simple terms, cloud computing platforms provide a range of on-demand services, including data and model storage, processing, computation and analytics. Instead of being limited to local servers or personal computers, cloud computing infrastructure allows users to connect to these vast computational resources remotely over the internet. There are three main service models for cloud computing: (i) software as a service (SaaS); (ii) platform as a service (PaaS); and (iii) infrastructure as a service (IaaS) (Biglaiser et al (2024)). For SaaS, clients rent applications that run on a cloud provider's infrastructure, eg applications like Spotify or Netflix. In the case of PaaS, users run and manage applications on a cloud platform without maintaining the infrastructure itself. Finally, IaaS, the service model most relevant for AI applications, refers to the use of cloud providers' computing and storage resources. Given that AI models are large and computationally demanding, they are typically trained and stored using the IaaS model. Extensive cloud computing power is also needed to fine-tune AI models and use them for inference.

With hardware and computing in place, the next essential component of the AI supply chain is the **training data** required for building large AI models. These extensive data sets include text, audio, video and images from both public and proprietary sources. Training data can also be synthetic, meaning data that are outputs of other AI models. GPT-3, the initial model behind ChatGPT, used several years' worth of text from the internet, including Reddit posts, Wikipedia and book repositories (Narechania and Sitaraman (2024)). The next step after accumulating vast troves of data is to make it suitable for training AI models and typically involves automated or human-driven data labelling.

The training data sets are then fed into "**foundation models**", which are large AI models that can be adapted for various functions and applications. The performance of a foundation model depends not only on its technical architecture but also on the volume and quality of the training data sets.

Finally, we have the **application layer** of the AI supply chain, which allows end-users to interact with AI models. Prominent examples include ChatGPT, Claude, Gemini, FinGPT, DALL-E, AlphaFold, Perplexity or GitHub Copilot. These AI applications leverage the underlying foundation models and computing infrastructure to deliver user-facing functionalities that range from natural language processing and financial forecasting to image generation, protein folding prediction and coding assistance.

3. Market structure of the AI supply chain

In this section, we delve into the market structure of each layer of the AI supply chain, outlining the key participants in the market as well as the main economic forces at play.[2] We begin with the hardware layer, which is critical for AI applications.

Hardware. Consider the most important hardware for AI applications, namely microprocessors like GPUs. Nvidia – headquartered in Santa Clara, California, in the United States – serves most of the market for GPUs, with its market share reported to be larger than 90% (Graph 2.A; CNBC (2023); The Economist (2024)). It has gross margins of over 70% and has seen its revenues increase by 405% between 2023 and 2024 (Nvidia (2024)). Initially serving the video game market, Nvidia had a head start in leveraging the parallel computing capacity of its GPUs for AI models. Over time, it has built-up substantial intellectual property and a significant reputation, solidifying its position as the market leader for GPUs.

Apart from the GPUs themselves, Nvidia also produces complementary software. Nvidia's GPUs come in an exclusive bundle with CUDA, its parallel computing platform, which enables programmers and software developers to simplify the process of using GPUs and to enhance their performance. CUDA has become the industry standard for programmers and can only be used with Nvidia's GPUs (The Economist (2024)). Additionally, Nvidia has made strategic acquisitions to bolster its market position: in 2019, it acquired Mellanox, a technology company that provides the architecture to connect GPUs in a network. This acquisition allows Nvidia's GPUs to function more efficiently than its competitors (The Economist (2024)).

To be sure, several other firms, including startups and big techs, are also active in the market for AI hardware. Advanced Micro Devices (AMD), Intel[3] and big techs like Microsoft, Google and Amazon are all producing AI microprocessors to compete with Nvidia's GPUs, both for training AI models and for inference. Chinese companies like Alibaba, Baidu and Huawei are also starting to produce their own microprocessors, especially in light of geopolitical constraints. Some manufacturers, like Amazon, are providing new business models by allowing developers to rent microprocessors through their cloud service. There are concurrent industry efforts to counter the dominance of CUDA by building open-source software and tools to supplement AI microprocessors across different manufacturers. Nonetheless, there is significant uncertainty about how effective this competition will be in altering the dominant position of Nvidia. Nvidia enjoys both a first-mover advantage in this market, and benefits from bundling CUDA with its GPU offerings. Not many alternatives exist for CUDA and developers have been building downstream code on the platform for several years, which might make it challenging to migrate to other manufacturers' platforms.[4] Additionally, Nvidia is trying to integrate into other layers of the supply chain, notably by launching its own LLMs (Hagiu and Wright (2025)).

[2] Note that with rapidly evolving technologies like AI, the market structure can also change rapidly. We analyse the market structure based on the latest available data.

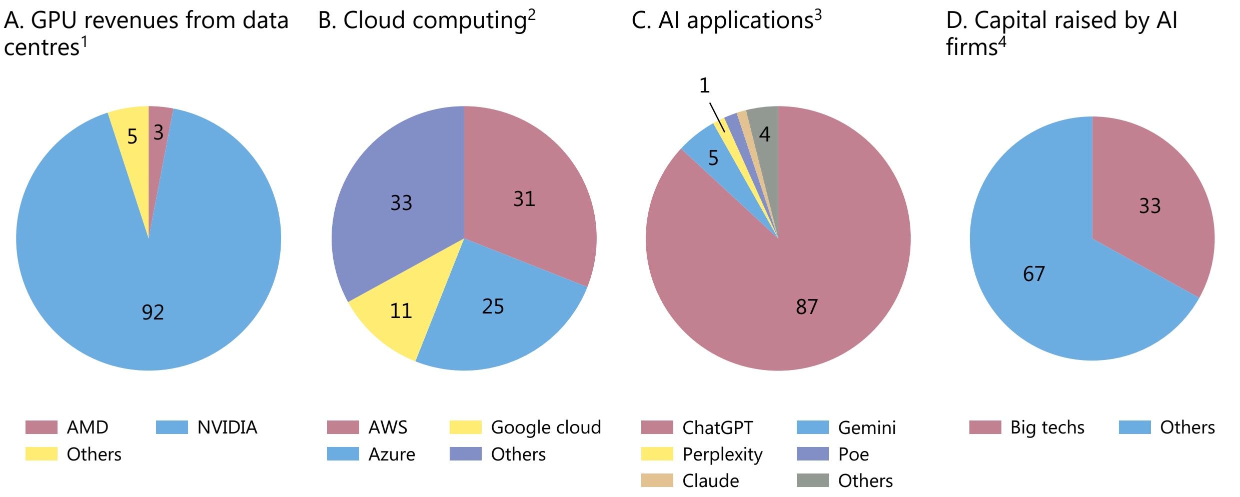
[3] AMD and Intel are also headquartered in Santa Clara.

[4] Notably, Nvidia's chips are produced by the Taiwan Semiconductor Manufacturing Company (TSMC), which itself produces over 60% of all semiconductors globally, and over 90% of the most advanced semiconductors (The Economist (2023a)). This further underscores the concentration in the hardware layer of the AI supply chain, as the reliance on a single manufacturer for the vast majority of advanced semiconductors introduces additional risks and dependencies.

Market structure of the AI supply chain

In per cent

Graph 2



[1] Based on global revenues of GPU producers for GPUs used in data centres in 2023. [2] Based on global cloud computing revenues for Q1 2024. [3] Based on monthly visits data. For further details see Liu and Wang (2024). [4] Based on total capital invested in 2023 in firms active in artificial intelligence and machine learning, as collected by PitchBook Data Inc. Big techs correspond to Alibaba Cloud Computing, Alibaba Group, Alphabet, Amazon Industrial Innovation Fund, Amazon Web Services, Amazon, Apple, Google Cloud Platform, Google for Startups, Microsoft, Tencent Cloud, Tencent Cloud Native Accelerator and Tencent Holdings.

Sources: Liu and Wang (2024); IoT Analytics Research (2023); PitchBook Data Inc; Statista; authors' calculations.

Cloud computing layer. Globally, the cloud computing market is dominated by three big tech companies: Amazon Web Services (AWS) with a market share of 31%, Microsoft Azure with 24% and Google Cloud Platform with 11% (Graph 2.B). In the European Union (EU), the estimated combined market share of AWS and Azure in 2020 was over 80% and their profit margins were also reported to be high, at 30% and 38%, respectively (Netherlands Authority for Consumers and Markets (2022)). In the case of the IaaS segment – the most relevant one for AI models – the market is even more concentrated. In 2023, AWS, Microsoft Azure and Google Cloud Platform together accounted for nearly 74% of the global market (Gartner (2024)). Many countries have similar market structures in the IaaS segment: these three cloud providers together accounted for 87% of the market in India in 2023, 77% in Australia in 2022 and 71% in Brazil in 2021 (Malik et al (2024); ACCC (2023); BNamericas (2021)).

There are several economic reasons to expect this market structure in the cloud computing market. First, the market is characterised by high switching costs for end users (Biglaiser et al (2024)). Different cloud providers often have different interfaces and technical features that make switching difficult without extensive retraining of engineers (exogenous switching costs). Switching costs might also depend on the choices that users make – for example, if they use proprietary software on a cloud platform, switching to another cloud service might be more costly than if they use more flexible, open-source platform (endogenous switching costs). Second, leading cloud providers in the market often charge an "egress fee" which imposes a cost on users for transferring data out of the cloud to a rival platform (Ofcom (2023)). Egress fees charged by the big cloud service providers exceed not only the incremental cost of transferring data but are also significantly higher than the fees charged by smaller competitors (Biglaiser et al (2024)).[5] Providers like Microsoft Azure also charge large licensing penalties and fees for using Windows applications on other cloud platforms (Reuters (2024b)). Third, cloud service providers also often offer an "ecosystem" of vertically integrated services, often at a discount, which can entrench their dominance across markets (Netherlands Authority for Consumers and Markets (2022)). Finally, the market is also characterised by high fixed costs, lack of interoperability and significant direct and indirect network effects in usage (Biglaiser et al (2024)).

Training data. So far, frontier AI models have been trained using vast troves of publicly available data. However, as the stock of public data rapidly declines, firms are turning to other data sources. Going forward, larger firms, particularly technology companies, can have a favourable position for three reasons. First, large firms like big techs often hold extensive proprietary user data from their primary business activities (such as social media). Second, certain firms may be able to benefit from the "data-network-activities feedback loop" and enjoy increasing returns to scale (Agrawal et al (2018); BIS (2019); Gans (2024)). For some AI applications, models that attract more users can use the data generated by these users to train future, improved iterations of the model, attract even more users, generate further training data and so on. Third, larger firms with greater financial resources may be able to strategically acquire or partner with smaller data owners to gather new training data. We discuss this last aspect in more detail in the next section.

At the same time, there are countervailing forces that may weaken the data feedback loop (Hagiu and Wright (2025)). Not all data generate meaningful feedback loops. If the use of AI applications generates data that does not necessarily improve the performance of the underlying model, then the competitive advantage of having access to such user data is limited. Moreover, there may be diminishing returns from additional training data, though empirical evidence on this is mixed (Gans (2024); Bajari et al (2019); Klein et al (2023); Schaefer and Sapi (2023)). The ultimate strength of the data feedback loop will vary with the nature of the AI applications; some applications will generate usable data that reinforces the loop, while others will not.[6]

Foundation models. At first glance, the market for foundation models is dynamic and rife with competitors. There are over 300 foundation models in the market, provided by 14 different firms (CMA (2024); Korinek and Vipra (2024)). There are also competing business models – while some firms choose to offer proprietary foundation models (like OpenAI and Google DeepMind), others have adopted a relatively more open approach (notably Meta with its open-source Llama models and, more recently, DeepSeek). Proprietary models offer limited flexibility to users in terms of model deployment and configuration and can be prohibitively costly for many users. On the other hand, open-source models can enhance competition and innovation by providing greater flexibility and customisation to users.[7] Nevertheless, the market for foundation models is currently dominated by only a handful of firms like OpenAI, Google DeepMind, Anthropic and Meta (CMA (2024)). In 2023, despite numerous competing foundation models, OpenAI's GPT-4 accounted for 69% of the market for generative AI in terms of global revenue (Korinek and Vipra (2024)). Given

[5] As Biglaiser et al (2024) note, egress fees may or may not turn out to be anticompetitive, but they can be one way for cloud service providers to entrench their dominant positions in the market.

[6] For a detailed analysis of the economic forces shaping data feedback loops, see Hagiu and Wright (2023 and 2025).

[7] Note that producers of open-source models may sell complementary services in exclusive bundles that may ultimately dampen competition (Lerner and Tirole (2000)). Open-source models may also be more vulnerable to misuse and cyber attacks.

the dynamic nature of the market and the potential to realise efficiencies, the hierarchy may shift rapidly.

What are the economic and strategic forces that shape the market for foundation models? The first is cost structure. Building and training foundation models involves high fixed costs arising from acquiring training data and computational resources. For example, the cost of training GPT-4 was estimated to be over $100 million (The Economist (2023b)). While these costs can decline over time as model training and inference become more efficient, they remain sizeable in absolute terms.[8] At the same time, the variable costs of operating the models are relatively low (Korinek and Vipra (2024)). High fixed costs combined with low variable costs typically give rise to extensive economies of scale in the production of foundation models. There might also be economies of scope as the same foundation model can be deployed across many different downstream markets.

The second aspect is competition "for" the market and the potential for market tipping (Korinek and Vipra (2024)). Even though the investment required to build and operate foundation models is substantial, several firms are willing to enter the market. This is likely because there is extensive competition "for" the market. Driven by economies of scale, economies of scope and inertia in user behaviour, the first firms to enter the market may enjoy significant advantages and make the market less contestable, as is the case with some digital platforms (Korinek and Vipra (2024)). This means that even when fixed costs of building foundation models decline, if there is extensive first mover advantage and a tendency among most users to single-home, competition "in" the market may be limited.[9]

There is also a tendency for producers of foundation models to vertically integrate with both upstream markets like hardware and cloud computing, and downstream markets like AI applications (CMA (2024)). While vertical integration can enhance efficiency in many cases, it can also create distortions, reduce competition and undermine innovation if vertically integrated firms restrict their rivals from accessing essential inputs or downstream markets.

On the other hand, as foundation models will have broad use cases, there is room for different providers in the market. Hagiu and Wright (2025) note that foundation models can be employed for varied use cases, such as language processing, content creation, customer service, image generation, protein structure and prediction, among others. These use cases can either be served by general foundation models with wide applicability or more specialised models. In either case, the nature of competition will vary with the use cases, depending on how the relevant market is defined for each of them. Within each market (of a use case), it is possible that economies of scale and scope prevail, and only a few foundation models operate.

AI applications and user-facing layer. The last stage of the AI supply chain, the user-facing layer, follows the playbook of digital platforms and mobile applications. Since the "ChatGPT moment" of AI, applications built on top of foundation models have been proliferating in various sectors of the economy including health, education, backend processing and compliance, software development and others. Nonetheless, and as with digital platforms, there can be a risk of "winner takes all" dynamics

[8] An illustration of declining costs is the recent launch of DeepSeek, a Chinese foundation model, reported to be trained at a much lower cost than GPT-4 (Financial Times, (2025a,b)).

[9] Consumers are said to single-home when they patronise only one service, platform or provider. Consumers multi-home when they use more than one service or platform, for example several competing foundation models at the same time.

emerging in the markets for AI applications. While it is a Herculean task to trace the market for AI applications in every sector, the market for chatbots can be instructive. As we show in Graph 2.C, despite a flurry of similar interfaces, ChatGPT still accounted for 60% of the chatbot market (measured by the total number of monthly visits) in 2024, highlighting the importance of being first to market.[10]

4. Role of big techs: towards big AI?

Perhaps the most notable development in the AI market is the increasing influence of (especially US and Chinese) big tech companies across the AI supply chain. Big techs already hold market power in many digital markets and are extending it to emerging AI markets. Big techs are investing heavily in AI: in 2023, they accounted for 33% of the total capital raised by AI firms, and nearly 67% of the capital raised by generative AI firms (Graph 2.D; Financial Times (2023)). Big techs are also actively partnering with and investing in several AI startups. The most notable deals include Microsoft's investment worth $10 billion in OpenAI, Builder.ai and Inflection AI, and Google and Amazon's investments in Anthropic and Hugging Face.

Big techs are vertically integrating across all layers of the AI supply chain, leveraging their unprecedented access to data, deep financial resources and dominance in the cloud computing market. The linchpin of AI is cloud computing, and as previously noted, only a handful of big tech companies provide most of the global cloud computing infrastructure. Furthermore, big tech cloud providers also impose exclusivity conditions when they partner with or invest in AI startups. For instance, the partnership between Microsoft and OpenAI binds the latter into using Microsoft Azure cloud for providing its services (OpenAI (2023)). Similarly, Amazon's investment in Anthropic is conditional on the startup using its cloud infrastructure (Lynn et al (2023)).

Big techs' significant advantage in training data can give them a further edge. The success of foundation models like GPT was facilitated by vast amounts of free and publicly available web-scraped data from repositories like Common Crawl. However, researchers estimate that the stock of high-quality language data will likely be exhausted by 2026 (Villalobos et al (2022)). OpenAI has already used much of the available textual data and is now relying on new sources such as audio from YouTube videos and podcasts for speech recognition (The New York Times (2024)).

Without enhancing data efficiency or finding new sources of data, the progress made by AI models may slow in the future. This is where big techs have a substantial advantage: they can train AI models on both publicly available data and their own proprietary data. Most big techs have their own potential pool of training data. As Hagiu and Wright (2025) note, Meta has Instagram, Facebook and WhatsApp; Google has Gmail, Maps, Play Store and Google Search; and Microsoft has Bing, LinkedIn and Microsoft 365. While data controlled by big techs are subject to data protection laws and privacy regulations in some jurisdictions, they have nevertheless been modifying their privacy policies and terms of service. For example, Google has broadened its terms of service to allow the use of data from Google Docs, Google Sheets and

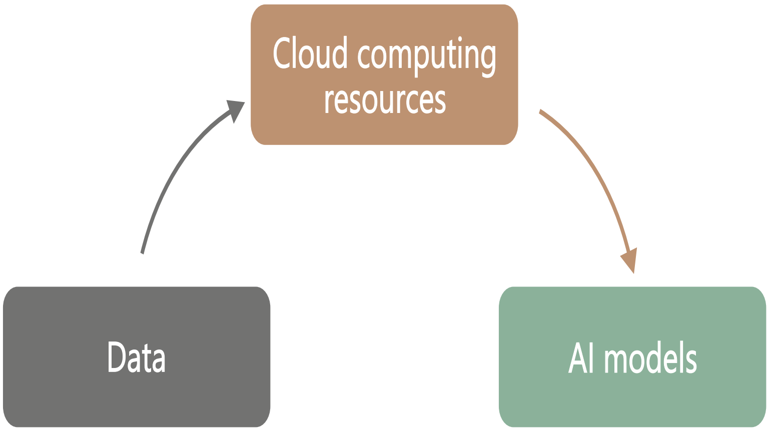
[10] A distinct issue related to AI and competition is the use of pricing algorithms in retail markets. Theoretical and empirical evidence show that algorithms may lead to supracompetitive pricing, even without explicit communication between firms (Calvano et al (2020); Assad et al (2024)). This issue is different from market structure concerns in the provision of AI, which is the focus of this paper.

Google Maps to feed into its AI products (The New York Times (2024)). Other big techs like Meta and Microsoft have adopted similar strategies (Microsoft (2023); Business Insider (2023)). Big techs are also acquiring or partnering with data owners. In 2019, Google acquired Fitbit, a fitness company with access to the health data of millions of users. In 2022, Meta, Google, Amazon and Apple partnered with Shutterstock, a company that provides images, videos and music files (Reuters (2024a)). These deals, along with several others, presumably aim to secure new data sources. Of course, the ultimate value of new training data is constrained by the strength of the data feedback loop.

Big techs are also active in the foundation model layer, either through partnerships or as producers of foundation models themselves. Due to their control over cloud infrastructure and proprietary data, AI startups have strong incentives to partner with big techs. Another dimension of these partnerships is emerging as big techs like Amazon, Google, Meta and Microsoft integrate into another layer of the supply chain: semiconductor design and AI chip manufacturing. For example, as part of its deal with Amazon, Anthropic must use specialised AI chips produced by Amazon for developing and training its models in the future (Amazon (2024)). Ultimately, this means that big techs can potentially develop foundation models more cheaply than other firms by using their own computational resources, AI chips and data. Big techs also provide AI applications: Google embeds Gemini into its services and search results, Microsoft provides Copilot across its applications, and Meta has integrated an AI-based assistant on Facebook, Instagram and WhatsApp. Meanwhile, Amazon also offers an AI companion on its marketplace.

Big techs' expanding footprints over the AI supply chain can lead to the emergence of a new cloud model data loop (Graph 3). With their control over computational resources and comparative advantage in producing, storing and analysing data, big techs may be able to provide better AI models. The use of these models generates more data, which can be optimally utilised by big techs' computational resources to improve their AI models and applications, making it more efficient to process and analyse data (so-called data gravity). This loop will be reinforced if substantial network effects arise from model use. Ultimately, it is likely that big tech companies will dominate significant parts of the AI supply chain.

Big techs in the AI supply chain — Graph 3



Source: authors' elaboration.

5. Challenges of a concentrated AI supply chain and policy implications

The market structure of the AI supply chain is consequential for several distinct policy goals. Consider welfare and innovation: a concentrated AI supply chain may limit consumer choice, lead to consumer lock-in effects, restrict smaller firms' access to critical inputs and lead to rent extraction. An example of the type of distortions emerging from concentration comes from the semiconductor industry. In 2021, amid a global semiconductor shortage, TSMC, the largest semiconductor manufacturer with a market share of over 60%, prioritised selling its available chips to Apple and assigned a lower priority to smaller firms. In the market for AI chips, large technology firms find it easier to secure Nvidia's GPUs than startups and researchers do (The New York Times (2023)). Moreover, a major risk to competitive outcomes is the tendency of large cloud providers to extend their dominant positions in the cloud market to other layers of the AI supply chain (Hagiu and Wright (2025)). If only a few firms dominate the AI market and entrench dominant positions going forward, they will also control the direction of innovation. While the private sector has had an important role in shaping innovative activity in the economy, control by a few firms of the AI supply chain increases the risk of misalignment between socially desirable innovation and privately profitable innovation (Acemoglu (2021)).

Concentration in the AI supply chain also has an impact on the operational resilience of critical infrastructure and cyber security (BIS (2024)). Relying on a few major suppliers for critical AI components or AI models themselves can create single points of failure. If any of these key players face disruptions – due to supply chain issues, operational errors, regulatory changes, climate change or geopolitical conflicts – the entire industry can be affected. This risk is of particular importance in the banking and finance sectors (BCBS (2024)). A recent example of the consequences of concentration for operational resilience is the CrowdStrike incident of 2024. A faulty software update from CrowdStrike, a large cyber security software provider, caused nearly 8.5 million computers using Microsoft Windows to crash. This was the largest information technology (IT) operational failure in history and disrupted operations in several industries including airlines, banking and manufacturing. Concentration also increases cyber security risks by centralising critical infrastructure and data in the hands of a few players. While these players tend to have very sophisticated cyber defences, a successful attack on one major provider could have a cascading effect across numerous organisations that rely on its services.

A related aspect is the relationship between the concentrated AI supply chain and financial stability. When AI models are used in finance, a concentrated supply chain can create systemic risk (Aldasoro et al (2024)). If financial institutions use the same AI models, or the AI models themselves rely on similar data sets, there is a heightened risk of procyclicality during times of financial stress and herding (Aldasoro et al (2024); Leitner et al (2024)). Additionally, the use of similar algorithms can lead to flash crashes, market volatility and illiquidity during times of stress (OECD (2021)).[11]

Given the diverse range of stakeholders and markets involved in the AI supply chain, policy remedies to influence the market structure and degree of concentration are not straightforward. The path towards effective regulation has several challenges.

[11] Another dimension relevant for the financial sector is the difficulty in assessing third party dependency risks and setting regulatory perimeters, particularly for big tech companies. See Crisanto et al (2024).

First, achieving consensus on policy measures is difficult as the AI supply chain contains many different markets that fall under the ambit of different regulatory authorities that often have competing goals.[12] Second, even if domestic policy consensus is achieved, international cooperation can be more elusive. The cross-border nature of AI demands joint policy efforts across jurisdictions. However, jurisdictions differ in their legal frameworks, geopolitical goals and regulatory approaches, making international coordination difficult to attain. Third, the pace of technological progress in the field of AI is usually much faster than regulatory capacity, making it challenging to design and enforce effective policy remedies. This points to the need to have a constantly evolving skillset among policymakers and regulators. Fourth, most often, antitrust measures are applied ex post. In digital markets, which are prone to winner takes all dynamics, ex post remedies may not restore competition sufficiently.[13] Finally, policy remedies need to balance static effects on competition with dynamic effects on future innovation.

Nevertheless, several policy measures are being considered to address the consequences of market concentration in the AI supply chain.[14] These include measures to enable data-sharing between firms, creating public data sets for training AI models, non-discrimination requirements for access to foundation models, multi-cloud strategies and common application programming interface (API) standards to reduce switching costs for firms (Korinek and Vipra (2024)). The first step will be to gather evidence on market structure and market conduct, since concentration by itself is not evidence for anticompetitive outcomes, though it may have consequences for outcomes beyond the ambit of competition policy, as outlined above. To this end, regulatory authorities in several jurisdictions are launching inquiries to evaluate market conduct and safeguard competition. Recently, the US Federal Trade Commission (FTC), the US Department of Justice (DOJ), the UK Competition and Markets Authority (CMA) and the European Commission (EC) issued a joint statement outlining the competition risks of AI (European Commission et al (2024)). In the United States, the FTC has initiated an investigation into the partnerships between Microsoft and OpenAI, Amazon and Anthropic, and Google and Anthropic. Concurrently, the US DOJ is investigating the acquisition of Run:ai, a firm that optimises the use of GPUs, by Nvidia. The CMA and EC are also looking into the Microsoft-OpenAI partnership.

6. Conclusions

The AI supply chain consists of five key layers: hardware, cloud infrastructure, training data, foundation models and AI applications. At the moment, the market structure of the first two of these layers exhibit significant concentration. The market for end-user facing AI applications is flourishing with the availability of many new applications across sectors (competition for the market) but winner takes all dynamics can easily emerge, as in the case of other digital platforms. Market structure in the AI supply chain is driven by high fixed costs, economies of scale and scope, network effects and

[12] Take, for example, the competing goals of maintaining user privacy and having data portability and interoperability to ensure a level playing field for smaller firms.

[13] While ex post antitrust measures can often be "too little, too late", ex ante measures will necessarily be based on more limited or less accurate information. This gives rise to an important trade-off for competition policy, particularly in the context of digital markets. See Crisanto et al (2021) for an overview of these trade-offs in the context of big tech companies.

[14] For a cross-country review of AI regulations focused on the financial sector, see Crisanto et al (2024).

strategic behaviour by dominant firms. At the same time, big tech companies are also expanding their footprint over the AI supply chain. These firms are uniquely positioned to leverage their existing dominance in data and cloud services to expand their control over the entire AI ecosystem. This may result in anticompetitive outcomes and barriers to entry for smaller firms.

Concentration in the AI supply chain poses several risks, including reduced consumer choice, control of the direction of innovation by a handful of firms, operational vulnerabilities, increased cyber security threats and potential financial instability. At the same time, given the global nature of AI, the diverse range of stakeholders and the nature of the economic forces shaping these markets, designing effective policy remedies may not be straightforward. As market concentration does not imply anticompetitive outcomes by itself, the first step is to collect evidence and monitor these markets. In this spirit, regulatory authorities in several countries are investigating market conduct across different layers of the AI supply chain with the view that a competitive and inclusive AI market will enhance innovation, resilience and welfare in the long run.

Global cooperation in regulating AI is crucial to effectively address these challenges (Aldasoro et al (2024)). AI technologies and their impacts are not confined by national borders, making international collaboration essential.[15] Harmonising regulatory frameworks, sharing best practices and coordinating enforcement actions can help mitigate risks associated with market concentration and ensure a level playing field for all market participants. By working together, different countries and jurisdictions can foster an environment that promotes innovation while safeguarding the public interest, ultimately leading to a more resilient and equitable AI ecosystem.

[15] Global collaboration on AI focuses on ensuring safety and transferring knowledge and best practices to ensure that all regions of the world can benefit from AI advancements responsibly. Initiatives like the G7 Hiroshima Process (signed in December 2023) and the transatlantic Trade and Technology Council (last meeting in April 2024) underscore the importance of international collaboration in establishing standards for the safe and ethical use of AI. More recently, at the AI Action Summit in Paris (February 2025), around 60 countries signed a declaration promoting inclusive and sustainable AI, emphasising global cooperation, safety and ethical development. Co-hosted by France and India, the agreement aimed to ensure that AI benefits all nations, particularly the Global South. The summit also saw the launch of "Current AI", a $400 million initiative supporting public interest AI projects. The declaration reflects growing international efforts to balance innovation with responsible AI governance.

References

Acemoglu, D (2021): "Harms of AI", *NBER Working Papers*, no 29247.

Agrawal, A, J Gans and A Goldfarb (2018): *Prediction machines: the simple economics of artificial intelligence*, Harvard Business Press.

Aldasoro, I, L Gambacorta, A Korinek, V Shreeti and M Stein (2024): "Intelligent financial system: how AI is transforming finance", *BIS Working Papers*, no 1193.

Amazon (2024): "Amazon and Anthropic deepen their shared commitment to advancing generative AI".

Assad, S, R Clark, D Ershov and L Xu (2024): "Algorithmic pricing and competition: empirical evidence from the German retail gasoline market", *Journal of Political Economy*, vol 132, no 3, pp 723–71.

Australian Competition and Consumer Commission (ACCC) (2023): *Report on the expanding ecosystems of digital platform service providers*, Digital Platform Services Inquiry.

Bajari, P, V Chernozhukov, A Hortaçsu and J Suzuki (2019): "The impact of big data on firm performance: an empirical investigation", *AEA Papers and Proceedings*, vol 109, pp 33–37.

Bank for International Settlements (BIS) (2019): "Big tech in finance: opportunities and risks", *Annual Economic Report 2019*, June, Chapter III.

——— (2024): "Artificial intelligence and the economy: implications for central banks", *Annual Economic Report 2024*, June, Chapter III.

Basel Committee on Banking Supervision (BCBS) (2024): *Digitalisation of finance*, May.

Biglaiser, G, J Crémer and A Mantovani (2024): "The economics of the cloud", *TSE Working Papers*, no 1520.

BNamericas (2021): "Who leads Brazil's cloud market and in which verticals?", November.

Business Insider (2023): "A long list of tech companies are rushing to give themselves the right to use people's data to train AI", 13 September.

Calvano, E, G Calzolari, V Denicolò and S Pastorello (2020): "Artificial intelligence, algorithmic pricing and collusion", *American Economic Review*, vol 110, no 10, pp 3267–97.

CNBC (2023): "Meet the $10,000 Nvidia chip powering the race for AI", 23 February.

——— (2024): "Why big tech is turning to nuclear to power its energy-intensive AI ambitions", 16 October.

Competition and Markets Authority (CMA) (2024): *AI foundation models: update paper*, April.

Crisanto, J, J Ehrentraud, A Lawson and F Restoy (2021): "Big tech regulation: what is going on?", *FSI Insights*, no 36.

Crisanto, J, C Leuterio, J Prenio and J Yong (2024): "Regulating AI in the financial sector: recent developments and main challenges", *FSI Insights*, no 63.

The Economist (2023a): "Taiwan's dominance of the chip industry makes it more important", 6 March.

——— (2023b): "Large, creative AI models will transform lives and labour markets", 22 April.

——— (2024): "Why do Nvidia's chips dominate the AI market?", 27 February.

European Commission, UK Competition and Markets Authority, US Department of Justice and US Federal Trade Commission (2024): "Joint statement on competition in generative AI foundation models and AI products", 23 July.

Financial Times (2023): "Big tech outspends venture capital firms in AI investment frenzy", 27 December.

——— (2025a): "OpenAI says it has evidence China's DeepSeek used its model to train competitor", 29 January.

——— (2025b): "DeepSeek's 'aha moment' creates new way to build powerful AI with less money", 29 January.

Gans, J S (2024): "Market power in artificial intelligence".

Gartner (2024): "Gartner says worldwide IaaS public cloud services revenue grew 16.2% in 2023", 22 July.

Hagiu, A and J Wright (2023): "Data-enabled learning, network effects and competitive advantage", *The RAND Journal of Economics*, vol 54, no 4, pp 638–67.

——— (2025): "Artificial intelligence and competition policy", *International Journal of Industrial Organization*, 103134.

IoT Analytics Research (2023): "Generative AI market report 2023-2030."

Khan, S and A Mann (2020): *AI chips: what they are and why they matter*, Center for Security and Emerging Technology, Georgetown University, April.

Klein, T, M Kurmangaliyeva, J Prüfer and P Prüfer (2023): "How important are user-generated data for search result quality?", *CEPR Discussion Papers*, no 17934.

Korinek, A and J Vipra (2024): "Concentrating intelligence: scaling and market structure in artificial intelligence", *NBER Working Papers*, no 33139.

Leitner, G, J Singh, A van der Kraaij and B Zsámboki (2024): "The rise of artificial intelligence: benefits and risks for financial stability", *Financial Stability Review*, May.

Lerner J and J Tirole (2000): "Some simple economics of open source", *The Journal of Industrial Economics*, vol 50, no 2, pp 197–234.

Liu, Y and H Wang (2024): "Who on earth is using generative AI?", *World Bank Policy Research Working Paper*, no 10870.

Lynn B, M von Thun, and K Montoya (2023): *AI in the public interest: confronting the monopoly threat*, Open Markets Institute, November.

Malik, P, B Das and H Jagadeesh (2024): *A competition analysis of the Indian cloud computing market*, ICRIER Prosus Centre for Internet and Digital Economy, August.

Microsoft (2023): *Microsoft services agreement*.

Narechania, T and G Sitaraman (2024): *An antimonopoly approach to governing artificial intelligence*, Vanderbilt Policy Accelerator for Political Economy and Regulation.

Netherlands Authority for Consumers and Markets (2022): *Market study into cloud services*.

The New York Times (2023): "The desperate hunt for the AI boom's most indispensable prize", 16 August.

——— (2024): "How tech giants cut corners to harvest data for AI", 6 April.

Nvidia (2024): "Nvidia announces financial results for fourth quarter and fiscal 2024", press release, 21 February.

Organization for Economic Cooperation and Development (OECD) (2021): *Artificial intelligence, machine learning and big data in finance: opportunities, challenges and implications for policymakers*.

Ofcom (2023): *Cloud services market study*, final report, April.

OpenAI (2023): "OpenAI and Microsoft extend partnership", 23 January.

Reuters (2024a): "Inside big tech's underground race to buy AI training data", 5 April.

——— (2024b): "Google complains to EU over Microsoft cloud practices", 25 September.

Schaefer M and G Sapi (2023): "Complementarities in learning from data: insights from general search", *Information Economics and Policy*, vol 65, 101063.

Villalobos P, J Sevilla, L Heim, T Besiroglu, M Hobbhahn, and A Ho (2022): "Will we run out of data? An analysis of the limits of scaling datasets in machine learning", arXiv:2211.04325v1.

Previous volumes in this series

No	Title	Issue date
BIS Papers No 153	Fast payments and financial inclusion in Latin America and the Caribbean	March 2025
BIS Papers No 152	Faster digital payments: global and regional perspectives	December 2024
BIS Papers No 151	Central bank digital currencies and fast payment systems: rivals or partners?	December 2024
BIS Papers No 150	Navigating uncharted waters: opportunities and risks for central banks	October 2024
BIS Papers No 149	Quantum computing and the financial system: opportunities and risks	October 2024
BIS Papers No 148	Keeping the momentum: how finance can continue to support growth in EMEs	September 2024
BIS Papers No 147	Embracing diversity, advancing together - results of the 2023 BIS survey on central bank digital currencies and crypto	June 2024
BIS Papers No 146	Central bank capital and trust in money: lessons from history for the digital age	June 2024
BIS Papers No 145	Generative artificial intelligence and cyber security in central banking	May 2024
BIS Papers No 144	The economic implications of services in the metaverse	February 2024
BIS Papers No 143	Central banking in the Americas: Lessons from two decades	November 2023
BIS Papers No 142	Inflation and labour markets	November 2023
BIS Papers No 141	Will the real stablecoin please stand up?	October 2023
BIS Papers No 140	Central banks, macro-financial stability and the future of the financial system	October 2023
BIS Papers No 139	Digital safety nets: a roadmap	September 2023
BIS Papers No 138	Financial stability risks from cryptoassets in emerging market economies	August 2023
BIS Papers No 137	Building an integrated surveillance framework for highly leveraged NBFIs – lessons from the HKMA	July 2023
BIS Papers No 136	Making headway – Results of the 2022 BIS survey on central bank digital currencies and crypto	July 2023

All volumes are available on the BIS website (www.bis.org).