NETWORK LAW REVIEW

ARTICLE

Digital Empires Reinforced? Generative AI Value Chain

Ela Głowicka and Jan Málek

***

1. Introduction

Is it a problem to see artificial intelligence (AI) foundation model developers partnering with large incumbents in the big tech industry? Is potential future market power in the AI industry to be avoided under any circumstances? Against the saga of struggling antitrust enforcement in the digital industry, competition authorities all around the world worry that the tech giants will leverage their market power to the generative AI (GenAI) products. Or the other way around: the big tech companies could use the GenAI to fortify their market power in other digital markets. Hence, competition authorities worry that large tech firms integrate young AI developers into their ecosystems to prevent them from rising as independent competitors by discontinuing the target's or the acquirer's own substitute GenAI product. Another concern is about the AI market tipping to the monopoly. If the market tips early on, it is very difficult to reverse it. Even if from a static perspective a single large language model (LLM) was a natural monopoly due to economies of scale in data and computing power, it may not be optimal for consumers from the dynamic perspective, as it could result in reduced incentives to invest and innovate in the long term. Such monopolist could also weaken the pass on of any efficiencies realised in the upper layers of the AI value chain like the cloud or hardware level. Finally, the AI technology could make collusion of algorithms more likely and at the same time more difficult to detect.[2]

Several competition authorities have recently scrutinised partnerships between firms in the different layers of the vertical chain of GenAI and conduct market investigations. There is a substantial concern by authorities that identifying an abuse of market power is expected to be difficult, slow, and going to come too late, only after bad conduct has already taken place. Commissioner Vestager warned that "if we don't

act soon, we will find ourselves, once again, chasing solutions to problems we did not anticipate."[3] Hence, some go as far as claiming that the focus should be on limiting future market power for LLM. For example, Susan Athey said that: "we're very worried and I think we all should be incredibly worried about future market power. But I'm also cautiously optimistic that I see regulators around the world educating themselves quickly and trying to stay on top of it."[4] However, in some industries, it is the very prospect of large size and market power that is driving companies to invest, to enter, and to bring solutions that improve welfare, for example by lowering costs or increasing choice with innovative products and services. Pre-emptying market power in such industries would likely result in these markets never grow. A good example is pharmaceuticals where a large upfront R&D cost is required to come up with a marketable product: In those markets patent protection guarantees a temporary monopoly on the developed product to give the innovator an incentive to pay the fixed upfront cost. Is the GenAI an industry of that kind?

This note reviews the most prominent economic characteristics of the AI value chain, namely the high fixed cost, economies of scale and scope, network effects and (likely temporary) capacity constraints in hardware. We then discuss the currently observable tendency to vertical integration in the AI value chain and connect it to these factors. Finally, we discuss whether the AI value chain promotes market power or rather efficiency and close with the assessment of the risk of a non-competitive market outcome.

2. Economic factors driving vertical integration in the AI value chain

The vertical value chain of Artificial Intelligence (AI) can be structured around five layers:[5]

- *Hardware (HW) suppliers* manufacture processing units for training and running the AI foundation models. Several types of chips are in use, but GPUs (Graphics Processing Units) have become the core of AI development. Nvidia with 90% market share in AI chips currently produces the state-of-the-art GPU called H100.[6] Rival companies are stepping up efforts to challenge its position. As of the beginning of 2024, only AMD announced AI chips on par with Nvidia's GPU performance.

- *Cloud providers* own large stocks of processing units which they can offer together with additional services as computing power to train and run the AI models without the need for AI developers to invest heavily in the expensive infrastructure. They benefit from the large scale of operations and are attractive to users because users only pay for used services and do not need to incur large capital outlays to buy hardware. Large cloud services are often labelled hyperscalers[7] and include particularly Google, Microsoft,

and Amazon. Meta does not provide a public cloud service, but it operates large computing power centres for its own, internal use.

- *Data owners* own intellectual property rights to data that can be available for a price or for free. The above cloud providers belong to the group of prominent data owners. However, there are also others owning suitable data pools for development of AI models, for example large publishing houses or firms specialised in data collection and aggregation for AI training.

- *AI foundation model developers* use data and software engineering to develop LLM by utilising the computing power offered by cloud service providers. As of February 2024, there are 11 startups that raised at least 100 million US Dolars, including OpenAI, Anthropic, Inflection, and others. In addition, large digital companies, for example Google or Meta, are internally developing their own models. Whilst there is no consensus on whether a single winner has already emerged, benchmarking from February 2024 suggests that several models perform on a par.[8] Fierce competition between AI foundational developers is ongoing.

- Finally, *AI application developers* build on the general LLM for specific applications that are offered to users. Based on the LLM, which is an essential input, this layer is very competitive because of low entry barriers. For example, based on the OpenAI GPT4, there is a multitude of plug-ins developed for specific purposes by OpenAI itself and by third parties, and GPT4 even offers help in creating an application-specific GPT. [9]

Complex technical and competitive interactions take place horizontally within the layers and vertically across the layers. There are multiple economic factors driving these interactions: high fixed cost, economies of scale and scope, network effects and a likely temporary bottleneck in the computing power capacity. This section provides a deep dive into each of them.

2.1. High fixed cost

To train, run and deploy their models, GenAI providers need computing power, data and engineering talent, that constitute significant costs not only to enter, but also to run LLMs in day-to-day operations.

Fixed startup costs to develop an in-house AI model can be estimated at $350-400 million. For example, an estimated 6000 H100 GPUs were needed to train GPT3.[10] With an average cost of over $40,000 per GPU,[11] the cost of hardware is $240 million. Next generations of GenAI will need up to $320 million in GPU hardware alone. Additional hardware is required (e.g. Central Processing Units, mainboards, RAM, storage, networking equipment, cooling systems), resulting in

roughly $120 million more. Facilities to house the hardware set-up could range from $10-40 million to buy, depending on the size requirements and location.[12] Once the LLM is trained, the same hardware can be used in daily operations or to train new models, so the cost is not sunk, but it needs to be available to kick off.

These estimates do not account for the expected unprecedented increase in computing power demands needed to train large AI models. While in the 54 years leading up to 2012 computing power used doubled every two years in line with Moore’s Law, a 2018 OpenAI study showed that computing power needed to train large AI models doubled every 3.4 months.[13] To develop the best AI model, GenAI providers must employ the most computationally powerful hardware available and update the infrastructure frequently. This allows for more efficient training and refinement of the AI models, where efficiency gains are measured both in computing power and energy consumption.

In expectation, this trend will continue to push up the total cost of hardware infrastructure. Recent reports indicate that the cost of training AI models is expected to continue rising in the coming years. OpenAI has projected that the cost of training large AI models will increase from $100 million to $500 million by 2030.”***[14]*** The cost of infrastructure needed to serve user requests of an existing, trained LLM can be much higher. According to New Street Research, the cost of delivering ChatGPT to Microsoft Bing users is estimated to be $4 billion.[15] Apart from the upfront investments, the cost of running LLMs is substantial, and again mostly relates to the cost of computing power. Operating an AI tool such as ChatGPT is said to be very capital-intensive, costing $700,000 per day,[16] or $255.5 million annually. Other estimates talk about a single entry costing 36 cents to generate.[17] A substantial chunk of this is the cost of electricity. It can take up to 10 gigawatt-hour (GWh) of power just to train a single large language model like ChatGPT-3. To process user requests, a lot of electricity is needed on a daily basis, in case of ChatGPT nearly 1 GWh each day. [18]

The quality and cost of engineering talent come in addition. The talent needed to develop LLM is rare and needs access to all the needed infrastructure. For this reason, the AI innovation cannot happen in the proverbial garage in Silicon Valley. As recognised by the Commissioner Vestager, it is likely that “We’re not going to see disruption driven by a handful of college drop-outs who somehow manage to outperform Microsoft’s partner Open AI or Google’s DeepMind.”[19]

Another essential input for AI models is data, which is key to model’s quality. Differently from processing capacity, data are non-rival by their nature. The data needed to train LLM is abundant but sometimes can be excludable. Data owners like publishing houses engage in agreements in AI developers and provide them data to

train the models. One example is a global partnership between Axel Springer and OpenAI.[20]

Different types of data result in different levels of data costs: there exist both publicly available datasets and pay-wall data that may come with a price tag. It might get increasingly difficult to train LLM on data freely scrapped from the internet. For example, a new tool Nightshade plagues images when they are scraped in an unauthorised way. It confuses the models that use them for training. It this way it increases the cost of training on unlicensed data, such that licensing images from their creators becomes a viable alternative.[21],[22] Over time, data owners learn how to protect their data from unauthorised use for training LLM and the cost of data is likely to increase. The quality and quantity of data also play a role: More data requires more processing power, and low-quality data may not yield the best results, so one may need to use more complex models. [23] According to Dimensional Research, companies on average need close to 100,000 data samples for the effective functioning of their AI models.[24]

There are high fixed costs in upper layers of the AI value chain, too. Cloud providers purchase the stock of the hardware and offer it to the AI developers as a subscription-based infrastructure service, which includes computing power resources, servers, storage and networks. The fixed costs necessary to procure hardware are massive. For example, in January 2024, Meta announced to procure around 350,000 Nvidia H100s, estimated to cost $10.5 billion.[25] In this context, competition authorities worldwide have highlighted a trend in the cloud sector towards concentration in the hands of a few firms which have investment capabilities and can afford these high fixed costs.[26]

Significant upfront investment is required by hardware suppliers to manufacture AI chips, too, due to advanced technology and expensive rare metals used as inputs. The research and development, expensive equipment and facilities lead to investment in a fabrication plant (fab) exceeding $35 billion.[27] The materials used in semiconductor manufacturing, including silicon wafers and various rare metals, contribute to the cost.

Given the high startup costs and extensive operating costs, a long period of loss-making for new GenAI entrants must be expected. Without startup funding or major short-run profitability, a GenAI provider with in-house infrastructure would expect to suffer cumulative losses for several years. Even if the GenAI startup would receive $1 billion in funding (like OpenAI received in 2019),[28] this initial funding would likely run out within the first 2-4 years of development (assuming the GenAI does not amass sufficient profit in these years). For example, in 2022 OpenAI reported losses of around $540 million, roughly doubling its losses from the previous year. These

losses occurred in parallel to the development of ChatGPT and hiring key employees, reflecting massive costs associated in training ChatGPT ahead of its launch.[29]

Taken together, significant financial resources are needed at several layers of the value chain to develop and deploy foundation models. This results in a substantial barrier to enter the market for LLM products. This high fixed cost means that significant scale of operations is needed to recover them. If market players can expect large rents, they are more likely to incur these fixed costs, but if they expect to share the rent with too many competitors, their share of the rent may be too small to incentivise entry, particularly in the presence of economies of scale. [30]

2.2. Economies of scale and scope

The GenAI is characterised by economies of scale and scope in the sense of cost or quality advantages that enterprises obtain due to their scale of operation and the scope of their activities.

First, economies of scale and scope mean that the average total cost of using AI model by the end consumer decreases as the scale of the model is larger. This is typically achieved because hardware and the team of talent can be used for repeated training and computing power for queries can be optimised better. In this way resources achieve higher utilisation and knowledge spillovers across training different models within one company can be exploited.

Second, there are economies of scale and scope from using larger data repositories in terms of better LLM quality. Datasets with more observation points and with more types of observations (variables) are better for the prediction precision of the models trained on them. The empirical literature on economies of scale and scope in data aggregation find that prediction accuracy improves with more data, especially with user personalized data (Hocuk et al., 2023; Schaefer et al., 2018; Schaefer and Sapi, 2022; Lee and Wright, 2023; Muller et al., 2018; Neumann et al., 2023; Chiou and Tucker, 2017; Bajari et al., 2018). Put differently, the more complex and larger dataset, the better the prediction accuracy of LLM and quality of its output (Martens, 2020).

However, the economies of scope do not go on forever and are subject to diminishing returns. Adding more variables to a dataset initially increases the predictive power of an AI model significantly, but these benefits decrease as more variables are added beyond a certain threshold. In some cases, diminishing returns may set in at a very early stage (Pilaszy and Tikk, 2009); in other cases, they arrive when the number of observations increases many orders of magnitude (Varian, 2014).

Complex and large data is available at the cloud service providers. Public alternatives to the data in the cloud appear to attract less interest. For example, the European Commission has been supporting the development of common European data spaces to make more data available for access and reuse in a trustworthy and secure way.[31] As of January 2024, primary challenges are the deployment of and interconnection between data spaces and limited interest in participation and use of the data spaces.[32]

Third, at the level of cloud services, large scale of operations leads to more efficiency in terms of lower energy use, and cost savings. Cloud providers can spread fixed costs over a larger user base by managing large data centres and clustering of operations. This results in lower operational costs per unit. Large cloud operators invest in cutting-edge cooling technologies and renewable energy sources to power their data centres and in advanced cybersecurity tools to protect the data they store. These investments, combined with efficient server utilization, significantly reduce energy consumption, environmental impact and security compared to traditional on-premises data centres. Moreover, cloud providers can optimize the utilization of their physical hardware. This means they can serve more customers with less infrastructure, increasing efficiency. Large cloud operators can also offer services of higher quality resulting from data centres distributed worldwide, allowing them to deliver content and services with low latency. The large scale can also result in better negotiation power and lower purchase price of the hardware, which can then be further passed on to lower layers of the vertical chain. Overall, this allows for more competitive pricing and increases the quality of the provided services. The competition between several large hyperscalers has resulted in prices continuously dropping over the past years, even to such an extent that smaller competitors who are not able to afford the necessary scale are being driven out of the market.[33]

Summing up, there are economies of scale and scope in the AI industry, both at the level of AI foundation models and at the level of cloud services. These economies imply that the final service can be produced at a lower cost or in better quality when the scale is larger, i.e. a bigger provider can do it cheaper and/or better.

2.3. Network effects

Network effects benefit the quality of AI foundation models, as they can create a positive feedback loop of improvement and adoption. They manifest themselves by the fact that as more developers and users engage with foundation models, they contribute data, either directly through training and fine-tuning or indirectly through usage feedback. This expands the dataset and can lead to improvements in the model's performance. AI models benefit from scale and scope of the user base and improve with use. As more developers and organizations adopt these models for various applications, the models can learn from the vast array of interactions, leading

to performance improvements that benefit all users. The AI industry has yet to provide evidence for the magnitude of these effects.

Moreover, the development of tools, platforms, and applications designed to integrate or build upon AI foundation models enhances the value of the overall ecosystem. For example, tools that simplify the integration of AI models into existing software can increase the utility of foundation models, making them more attractive to a broader range of users. At the same time, they increase the utility from using existing software. A vibrant community of developers fine-tuning AI models can create a rich ecosystem of applications and solutions. Specialized models tailored for specific industries or tasks can trigger indirect network effects, as the availability of more targeted models can make the overall ecosystem more valuable to users with diverse needs.

The existence of an ecosystem with a seamlessly operating environment including an AI foundation model can lead to indirect network effects by making it easier for different products and services based on AI models to work together. This interoperability can increase the value of all participating products and services, as users benefit from a more integrated and cohesive ecosystem.

The risk that positive network effects brings with them is that tipping is more likely. Indeed, Prüfer and Schottmüller (2021) find in a theoretical model that duopoly markets tip at very mild conditions. After tipping, innovation incentives both for the dominant firm and the competitor are small. If consumers favour one LLM by and large, that incumbent will get a market position that is very hard to contest and hardly any competitive pressure exists. In the long term, this market position can lead to high prices, less innovation and lower quality.

2.4. Capacity constraints in hardware

The training of AI models and running them requires substantial computational power based on specialised hardware, allowing massive parallelism of computations.[34] Hardware for AI development includes different types of chips, but GPUs have generally become the golden standard.[35],[36]

Since 2022, the demand for specialized hardware has skyrocketed. The rapid take up of demand for AI chips has taken the manufactures by surprise.[37] Due to its long experience with the development of GPUs for the gaming industry, Nvidia was the best positioned to serve the suddenly exploding market.[38] The state-of-the-art chips are currently Nvidia’s H100,[39] manufactured by the TSMC.[40] As of beginning of 2024, Nvidia’s market share in AI chips was estimated at 90%.[41] Rival companies are stepping up efforts to challenge the position. AMD recently launched MI300X chip in December 2023.[42] This chip is perceived as presenting a challenge to NVIDIA but despite its lower price, it is unclear how much it can disrupt the market

given Nvidia's established software stack. Major orders have been however already placed by Microsoft and Meta.[43]

Currently, there are not enough GPUs to satisfy the demand of cloud providers and AI developers.[44] Majority of server GPUs are supplied to hyperscale cloud service providers, with other companies being rationed out from the market.[45] Major cloud-server providers had to limit their availability for customers.[46]In addition, there is evidence that developers must delay product releases, and face increased production costs.[47] OpenAI supposedly delayed training ChatGPT 5 because of these constraints.[48]

It is not clear for how long the AI chip shortage will be present. On the one hand, there are efforts to scale up production. Nvidia planned to triple the output of H100s up to 1.5 to 2 million in 2024.[49],[50] Furthermore, entry by AMD is expected in 2024, which can release the bottleneck, at least to some extent. Still, the AI chip shortage creates a bottleneck and the lack of scale-up possibilities, perceived as "the single greatest challenge in AI."[51]

3. Integration into vertical ecosystems

The current development in the AI industry gravitates towards vertical cooperation of different forms (investments, partnerships, alliances), with vertical ecosystems being created across the different layers of the vertical chain. To date in the western economy, five major clusters have formed around large tech companies: one around Google, one around Nvidia, one around Microsoft, one around Amazon, and one around Meta. While all are formed around large companies with deep pockets, there are important differences between these stacks. Each of the vertical stacks is strong at the different layer or offers different digital services based on heterogenous business models. While Nvidia supplies hardware for LLMs, Meta is pursuing an open-source approach for its LLM models. Google, Microsoft and Amazon are all strong in cloud computing, but their core/historical digital services range from the web search and advertising to PC operating system and software, to e-marketplace and retail. Having these companies compete against each other in the GenAI industry can generate strong competitive pressure. Importantly, apart from Google, every other stack misses on one layer to deliver the Gen AI and attempts to complete it by cooperating vertically with others. The status of vertical integration in the AI value chain is illustrated in Figure 1.

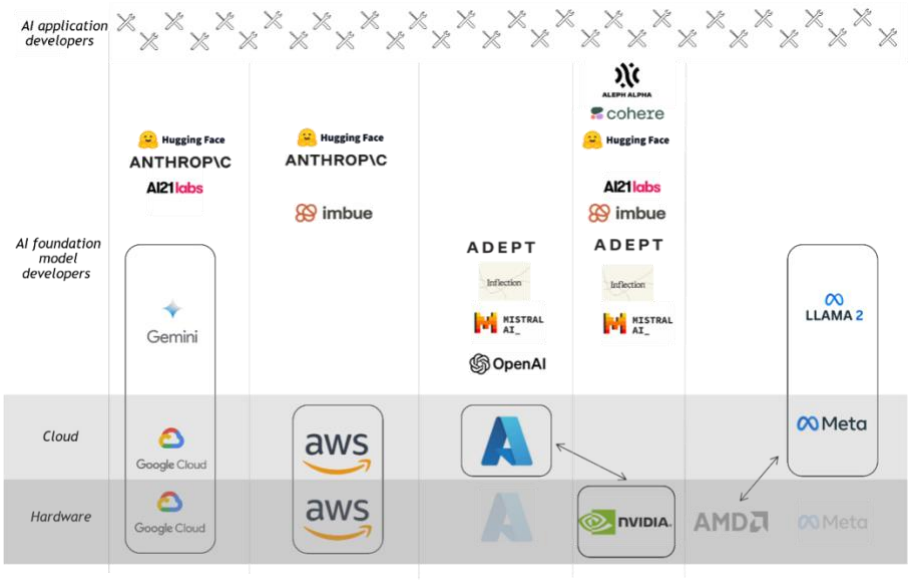


Figure 1: Vertical chain of AI and the relationships between the layers (March 2024)

Source: E.CA Economics. The figure visualises integration into vertical ecosystems around major cloud and hardware companies. Each row indicates a layer in the value chain of the AI, from upstream to downstream. Each column indicates one vertical stack. The GenAI layer includes both startups and models developed in-house. Assignments into stack depends on whether the cloud/chip company has invested into the GenAI startup (see Table 1 for more details). The black rectangles highlight the key player in the stack and layers where it is itself active. The shaded logos indicate areas in which entry is expected in 2024 as it was publicly communicated. The arrows between HW companies and cloud providers indicate announced alliances. Meta is listed in the cloud layer but uses cloud only for in-house services.

First, at the AI foundation model level, large companies are developing their own models (for example Google’s Gemini or Meta’s Llama), form alliances with GenAI developers, or do both at the same time. The following table provides an overview of the funding raised and the identity of their investors and lead investors for the 11 startups that received the most funding.

	Total Funding (Mio USD)	Share where involved	Google	Amazon	Microsoft	Nvidia	AMD	Meta
OpenAI	$11,300	97.3%			o			
Anthropic	$7,554	89.4%	o	o				
Inflection AI	$1,525	85.2%			o	o		
Aleph Alpha	$639	16.9%				x		
Mistral AI	$544	79.2%			x	x		
Cohere	$428	63.1%				x		
Adept AI	$415	84.3%			x	x		
Hugging Face	$395	59.5%	x	x		x	x	
AI21 Labs	$327	47.5%	x			x		
Imbue	$232	91.4%		o		x		
Character.ai	$150	0.0%						

Figure 2: Venture capital investment in GenAI businesses

Source: E.CA Economics based on data from Crunchbase and publicly available information in press releases as of February 28,2024. Note: x – listed as investor. O – listed as lead investor. The table includes GenAI developers outside of China that received at least one funding exceeding 150 million USD and which are listed by Crunchbase as developers of generative AI.

Large majority of the money raised by GenAI startups comes from cloud providers and hardware manufacturers. Most GenAI developers are backed by few large partners – either a cloud provider or AI chip manufacturer, or both. Typically, one of the large firms would be a lead investor in the financing rounds received by these startups. At the end of January 2024, only Hugging face received funding from more than three firms and had in place collaboration agreements with all major cloud providers and hardware providers.[52] Only Aleph Alpha has a low share of funding by major players in the AI value chain. Only Character.ai was not (yet) backed by any of the large cloud providers nor Nvidia. Note that the investments support both closed and open-source AI models, but some deals result in an end to the open-source strategy (e.g. the recent partnership between Microsoft and Mistral Large[53]).

All large public cloud providers invested in GenAI. Microsoft in Open AI[54] and Mistral,[55] Google and Amazon in Anthropic.[56],[57] Startups receiving money from these firms (Open AI and Anthropic) are also the most funded, by a large gap. We also see a pattern of large investor concentration – Microsoft was involved in virtually all investments in Open AI, whilst Anthropic raised almost 90% of its money in rounds where Google and AWS were involved. The companies also seem to

diversify their AI bets. While they may have the most favored AI developer, they do invest in competing developers too, although to a much smaller extent.

Meta is the only large digital company involved in the AI chain that does not offer cloud services publicly to others but uses its computing capacity for a development of its in-house model, Llama 2. This likely drives the very different strategy Meta has adopted. Meta has not invested into any GenAI startup. Instead, it has formed an alliance with over 50 founding members and collaborators globally, including model developers, research institutes, and hardware providers such as AMD, Intel, and IBM but not Nvidia.[58] In addition, this seems to be the only digital ecosystem that pursues open-source LLM.[59]

Lastly, as the supplier of hardware, Nvidia follows yet a different investment pattern. While Nvidia has been the lead investor in the third most backed startup, Inflection AI, it is involved in far more startups than any other firm from the upper layers of the vertical chain – it invested in 8 out of 11. It is also typically co-investing with others, as the share of money where Nvidia was involved as an investor is much lower compared to the leading startups backed by large hyperscalers. This is interesting particularly with respect to reports that as the owner of the GPU bottleneck,[60] Nvidia prefers customers with nice brand names, startups with strong pedigrees,[61] or startups they have a relationship with.[62] This can also indicate that Nvidia tries to insure its position: it diversifies from being a pure supplier of hardware to being a player in the downstream layer of the chain, in the market for GenAI services. This strategy seems to be a hedge given the competitors (AMD) entering the hardware layer and GenAI developers expanding or considering expansion of their own activities into the hardware layer of the vertical chain. For example, OpenAI have been looking into the options how to engage into the production of their own specialised chips, addressing the shortage of compute power but also creating a vertically integrated AI developer with direct access to computing power. [63] All other large tech firms (Meta, Microsoft, Amazon, Google) try to counter the market power of hardware producers, too. Amazon is developing its own GPUs – Trainium and Inferentia. Google is partially already relying on own tensor processing units. Finally, Microsoft and Meta announced plans for the development of their own AI chips. [64]

4. Competition between vertical ecosystems

The consolidation into several vertically integrated ecosystems is a clear trend in the GenAI industry. Deep-pocketed cloud and hardware firms invest into GenAI developers, while cloud and GenAI developers enter AI chip manufacturing. Several clusters are being formed, with ecosystem leaders being strong in different layers of the chain and entering the other layers of the chain. Only Google has its own, full

vertical stack. The key question is whether this development can result in a competitive AI market where several ecosystems exert pressure on each other.[65]

The incentives to compete between vertically integrated ecosystems can be stronger than in the „mix-and-match" environment where firms at each vertical layer compete against each other horizontally but also vertically. If a vertically integrated ecosystem loses customers downstream, it also loses sales and scale upstream. This marginally reduces revenues and increases costs in the upper layers of the value chain. The loss in a vertical chain is larger than in a „mix-and-match" system where a loss in one layer is independent of a loss in another layer. This incentive, along with the discussed economic forces likely brings vertical integration to generating competitive pressure and some procompetitive effects.

First, partial or full integration to ecosystems **facilitates a scalable entry of AI developers and new entry in chip manufacturing.** Vertical partnerships can give more startups a chance to grow, commercialize and bring their products to end-customers within the ecosystems.[66] This is the more important, the longer it takes to define the business case for AI models. Currently AI developers are still heavily loss-making, and investors start discounting stocks of the major developers due to mounting costs.[67]

Vertical partnerships also strengthen the incentives for the cloud providers to compete for computing power capacity. Computing power is excludable, and preventing a competitor from using it means that it cannot continue further developing their AI models, providing seamless services, or acquire new customers. For an integrated ecosystem, the lack of computing capacity not only translates into direct loss of market share of their AI model, it also translates in a loss of scale at the cloud level and hence hurts twice. Beyond competition for capacity, this additionally motivates entry into the chip manufacturing layer to release the bottleneck itself. While in-house chip sourcing has a potential to reduce the costs of running AI queries by one third,[68] it also comes with an additional competitive benefit of technological differentiation as new chips can specialise depending on the need of the models developed by the ecosystems.

An additional benefit might come from the emergence of ecosystems that can compete with Google's. Google already has a fully integrated ecosystem. Creating competing ecosystems at this early stage of the GenAI industry could ensure that there is no tipping of the GenAI services to Google. Competition of a few strong ecosystems will give consumers more choice and innovation, higher quality and lower prices compared to a quasi-monopoly after the market tipped. On the contrary, if vertical integration in the ecosystems competing with Google is slowed down, tipping to the existing and fully integrated ecosystem becomes more likely. Note also

that both closed and open-sourced ecosystems are being created. This again helps to stimulate competition and to provide a differentiated offer to the consumer.

Second, **several types of cost savings** can be achieved. Vertical partnerships in the AI value chain motivate cloud providers to expand the scale and scope of the cloud services and achieve higher capacity utilization, better performance and lower operational costs due to economies of scale. For AI developers, this can be beneficial as they can rely seamlessly on one technological solution that allows to scale up or down operations flexibly. Cost savings can be also expected from the abundance of data available in the vertical ecosystems. This data advantage can be a significant competitive lever. By controlling more stages of the data collection and processing pipeline, companies can ensure that the data they collect is of high quality and relevant to their specific needs, leading to the development of more accurate and efficient AI models. This can raise the bar for what is considered state-of-the-art in AI technologies and push competitors to innovate. Admittedly, this also raises barriers to entry for new AI developers, who would need to find access to equivalent datasets to be able to compete.

Partial vertical integration can lead to cost savings by eliminating the need for third-party suppliers or intermediaries (e.g. reducing the cost of purchasing data, negotiating contracts, and managing external relationships expenses). It allows to implement uniform data privacy and security measures across all stages of data collection, storage, and processing, which can help in better complying with data protection regulations and in building trust with users and customers by ensuring their data is handled securely.

Third, tough competition in the AI market between vertically integrated firms also ensures that important benefits realised in the upper chain by achieving scale or countervailing Nvidia are going to be **passed on to consumers**. The marginal cost of generating output of AI tools is non-zero along the different layers, and it is in the interest of the vertical ecosystem to reduce the cost or increase the quality in the upper part of the vertical chain and pass such efficiencies on to expand market share at the downstream level. [69] Positive network effects within the ecosystem strengthen this incentive. This is how ecosystems can generate larger market share in the AI market by transferring the advantages from the upper layers to the lower layers of the value chain.

On the negative side, an oligopoly of several vertically integrated ecosystems brings some risks for competition. Collusion of algorithms and the abuse of market power over locked-in customers can turn out to be an issue. For this reason, active and competent competition authorities are needed to watch the business practices of vertical GenAI ecosystems with market power.

5. Outlook

This note explains how vertical integration along the GenAI value chain into several ecosystems shapes competition. Given the high fixed and operating costs and barriers to entry in multiple layers of the value chain, vertical partnerships and/or integration can facilitate competition in the downstream AI market through several channels: increasing scale and scope of cloud services and realising cost efficiencies, facilitating use of high-quality data for better performance of the AI models, and alleviating financial constraints for GenAI developers.[70] The efficiencies from scale incentivise vertical ecosystems to compete for computing capacity or even tackle the supply bottleneck by entering on their own manufacturing. Strong competition between several ecosystems along the vertical chain also ensures that it is in the ecosystem's interest to pass efficiencies down to the consumers[71] and makes market tipping less likely.

Despite these procompetitive effects, risks for competition are also present: large tech companies might leverage market power from other markets into the AI industry or use AI to fortify the market power in other markets. Competition authorities would like to see "mix-and-match" competition at and between all layers of the vertical chain rather than vertical integration. Indeed, some of their recent actions indicate heightened levels of scrutiny on the developments in this market.[72]

In its Digital Agenda, the European Commission has identified the lack of interoperability and compatibility needed for such mix-and-match approach as one of the significant obstacles to a thriving digital economy. This model of competition sets incentives to compete on price, on innovation, on quality, and other competition parameters. However, this model has practical problems for implementation and the trade-offs between benefits and costs of a higher degree of interoperability are complex. Obtaining and maintaining compatibility often involves a sacrifice in terms of product variety or restraints on innovation.[73] Vertical interoperability may reduce economies of scale and scope.[74]

Vertical ecosystems offer a level of efficiency that mix-and-match solutions are unlikely to mirror. In addition, mix-and-match may not drive the same level of competition between foundational models: competition would focus on individual product features and pricing rather than on creating a comprehensive ecosystem.[75] This potentially leads to less intense overall competition as less is at stake at each layer separately than for an ecosystem as a whole. Finally, the competitive dynamics can lead to duplication and inefficient use of infrastructure and reduced incentives to innovate and enter. All these effects need to be considered when choosing preferred market structure in the GenAI industry.

An ultimate risk is tipping at the end of the chain, with only one GenAI model dominating the space. Tipping could happen also because of hoarding computing capacity by a specific stack or when one of the GenAI developers gets ahead of the others because it comes up with technological feature that makes the model superior in quality. The AI market has characteristics which make it prone to tipping: these include positive network effects, switching cost (e.g. egress fees and potential interoperability issues), free essential facilities (e.g. ChatGPT 3.5), data-enabled learning, trust and complementary offerings of ecosystems (e.g. complementary applications). [76] However, some market characteristics work against tipping: constant innovation in technology and business model, consumer heterogeneity and product differentiation tend to sustain multiple networks. The existence of several integrated vertical ecosystems with similar service quality can limit the likelihood of market tipping. With each vertical stack owning unique datasets that competitors cannot easily obtain, exclusivity can be crucial in development of heterogeneous models, where the quality and uniqueness of data can significantly impact the performance of AI models. The market equilibrium with multiple incompatible products then reflects the social value of variety.[77] Some AI users will also likely multi-home, making it difficult for a single 'gatekeeper' to unduly influence consumers' choices. In the presence of these counter-balancing features, it is very unclear whether the AI market will tip in favour of a single ecosystem. The future is still to come but right now, four digital empires and a chip manufacturer compete fiercely for the AI spotlight -this is a development to be praised.

Ela Głowicka
Director at E.CA Economics.

Jan Málek
Manger at E.CA Economics.

***

Citation: Ela Głowicka and Jan Málek, *Digital Empires Reinforced? Generative AI Value Chain*, Dynamics of Generative AI (ed. Thibault Schrepel & Volker Stocker), Network Law Review, Winter 2024.

References

- [1] We would like to thank Jakub Kałużny and Alena Kozáková for productive discussions and comments. We also thank Zachary Herriges and Giulia Santosuosso for research assistance. The note presents authors' own professional views. We do not work for companies active in the GenAI value chain. Our employer E.CA Economics works for Apple currently and worked for Meta in the past. The note reflects the market situation and information available as of March 2024.

- [2] Justin Johnson, Daniel Sokol, AI Collusion (Algorithm), Global Dictionary of Competition Law, Concurrences, Art. N° 115883.
- [3] https://ec.europa.eu/commission/presscorner/detail/en/speech_24_931
- [4] CEPR Competition Policy webinar on GEN AI & Market Power: What Role for Antitrust Regulators? https://cepr.org/system/files/2023-07/Competition%20Policy%20RPN%20Webinar_Gen%20AI%20%26%20Market%20Power_12%20July%202023_Transcript.pdf
- [5] Portuguese competition authority provides a deep-dive at the stratification of the AI vertical chain https://www.concorrencia.pt/sites/default/files/documentos/Issues%20Paper%20-%20Competition%20and%20Generative%20Artificial%20Intelligence.pdf. See also Thomas Höppner & Luke Streatfeild. (2023) ChatGPT, Bard & Co.: an introduction to AI for competition and regulatory lawyers, 9 Hausfeld Competition Bulletin, 1/2023.
- [6] https://www.smbom.com/news/13181
- [7] https://www.bruegel.org/sites/default/files/2023-12/WP%202023%2019%20Cloud%20111223.pdf
- [8] https://mistral.ai/news/mistral-large/
- [9] This categorisation broadly aligns with the "tech stack of AI" presented at the conference organised by Bundeskartellamt in February 2024. While this first layers broadly overlap (Chips, Datacenters, Data, Foundation models), we group the other four layers (Tooling, Applications, Distribution, Developers and users) into the developer layer. International Cartel Conference 2024 , Berlin
- [10] https://community.juniper.net/blogs/sharada-yeluri/2023/10/03/large-language-models-the-hardware-connection?CommunityKey=44efd17a-81a6-4306-b5f3-e5f82402d8d3
- [11] https://www.businessinsider.com/openai-boss-sam-altman-wants-his-own-chip-supply-2024-1
- [12] https://www.loopnet.com/search/warehouses/san-francisco-ca/for-sale/?sk=4a1233d9bf4c75fed5156e7bd84fa3e0, accessed on 14 February 2024.

- [13] https://www.technologyreview.com/2019/11/11/132004/the-computing-power-needed-to-train-ai-is-now-rising-seven-times-faster-than-ever-before/
- [14] https://insight.openexo.com/how-much-is-invested-in-artificial-intelligence/ and, specifically, https://mpost.io/ai-model-training-costs-are-expected-to-rise-from-100-million-to-500-million-by-2030/?ref=insight.openexo.com.
- [15] https://www.cnbc.com/2023/03/13/chatgpt-and-generative-ai-are-booming-but-at-a-very-expensive-price.html
- [16] https://timesofinsider.com/openai-annual-revenue-2023-surpasses-1-3-billion-mile/.
- [17] https://www.spiegel.de/wirtschaft/unternehmen/nvidia-und-die-ki-wie-jensen-huang-microsoft-amazon-und-google-attackiert-a-094ebb7f-17ad-43cc-9a87-0a7644bd8437
- [19] https://ec.europa.eu/commission/presscorner/detail/en/speech_24_931
- [20] https://www.axelspringer.com/en/ax-press-release/axel-springer-and-openai-partner-to-deepen-beneficial-use-of-ai-in-journalism
- [21] https://venturebeat.com/ai/ai-poisoning-tool-nightshade-received-250000-downloads-in-5-days-beyond-anything-we-imagined/?utm_source=substack&utm_medium=email
- [22] https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html
- [23] https://www.akkio.com/post/cost-of-ai.
- [24] https://www.kdnuggets.com/2021/05/shaip-budgeting-ai-training-data.html
- [25] https://uk.pcmag.com/ai/150532/zuckerbergs-meta-is-spending-billions-to-buy-350000-nvidia-h100-gpus
- [26] Carugati, C. (2023) ‘The competitive relationship between cloud computing and generative AI’, Working Paper 19/2023, Bruegel
- [27] https://www.bcg.com/publications/2023/navigating-the-semiconductor-manufacturing-costs

- [28] https://www.crunchbase.com/organization/openai/company_financials, accessed on 14 February 2024.
- [29] https://www.theinformation.com/articles/openais-losses-doubled-to-540-million-as-it-developed-chatgpt
- [30] William J. Baumol, Robert D. Willig, Fixed Costs, Sunk Costs, Entry Barriers, and Sustainability of Monopoly, *The Quarterly Journal of Economics*, Volume 96, Issue 3, August 1981, Pages 405–431, https://doi.org/10.2307/1882680
- [31] https://digital-strategy.ec.europa.eu/en/policies/data-spaces
- [32] https://digital-strategy.ec.europa.eu/en/library/second-staff-working-document-data-spaces
- [33] https://www.inderscienceonline.com/doi/abs/10.1504/IJCC.2023.129772
- [34] https://www.synopsys.com/blogs/chip-design/ai-chip-architecture.html
- [35] Graphics Processing Units (GPUs) were originally designed for rendering graphics and become a popular choice for parallel processing tasks. They consist of thousands of small cores optimized for handling vector and matrix operations. GPUs are created by connecting different chips — such as processor and memory chips — into a single package.
- [36] Other used chips include CPUs (Central Processing Units) used for tasks that require less parallel processing. TPUs (Tensor Processing Units) offer high efficiency for AI computations specifically for neural network machine learning. There are also some other technologies, for example FPGAs (Field-Programmable Gate Arrays). The GPUs are dominated by Nvidia. The CPU segment is dominated by Intel. The TPU are only developed and accessible though Google.
- [37] Quotes of many executives highlight how serious the excess demand is: "GPUs are at this point considerably harder to get than drugs." https://www.youtube.com/watch?v=nxbZVH9kLao "In fact, we're so short on GPUs, the less people use our products, the better". https://www.cnn.com/2023/05/17/tech/sam-altman-congress/index.html There are also reports that black markets for AI GPUs have emerged in recent months. https://www.computerworld.com/article/3707412/chip-industry-strains-to-meet-ai-fueled-demands-will-smaller-llms-help.html

- [38] https://www.spiegel.de/wirtschaft/unternehmen/nvidia-und-die-ki-wie-jensen-huang-microsoft-amazon-und-google-attackiert-a-094ebb7f-17ad-43cc-9a87-0a7644bd8437?giftToken=c271777d-78d7-4fc6-8829-12a42cd86d8b
- [39] https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand/
- [40] https://www.businessinsider.com/openai-boss-sam-altman-wants-his-own-chip-supply-2024-1
- [41] https://www.smbom.com/news/13181
- [42] https://www.tweaktown.com/news/95001/amd-mi300x-vs-nvidia-h100-battle-heats-up-says-it-does-have-the-performance-advantage/index.html
- [43] https://www.barrons.com/livecoverage/microsoft-alphabet-google-amd-earnings-stock-price-today/card/amd-is-taking-on-nvidia-in-ai-chips-here-s-the-market-share-it-s-expected-to-take-1jFpSMKQy9lVrF89wvGH
- [44] https://albertoromgar.medium.com/200-million-people-use-chatgpt-daily-openai-cant-afford-it-much-longer-55bf2373d01c
- [45] https://www.tomshardware.com/tech-industry/nvidia-ai-and-hpc-gpu-sales-reportedly-approached-half-a-million-units-in-q3-thanks-to-meta-facebook
- [46] https://www.theinformation.com/articles/ai-developers-stymied-by-server-shortage-at-aws-microsoft-google
- [47] https://medium.com/@Aethir_/shortage-of-ai-chips-and-its-impact-on-the-ai-boom-7c8a9818b78b
- [48] https://albertoromgar.medium.com/200-million-people-use-chatgpt-daily-openai-cant-afford-it-much-longer-55bf2373d01c
- [49] https://www.tomshardware.com/news/nvidia-to-reportedly-triple-output-of-compute-gpus-in-2024-up-to-2-million-h100s
- [50] https://www.tomshardware.com/news/nvidia-to-reportedly-triple-output-of-compute-gpus-in-2024-up-to-2-million-h100s
- [51] https://www.barrons.com/articles/nvidia-ai-chips-taiwan-jensen-huang-9f549651
- [52] At the end of January 2024, Hugging face annnounced that they are the only AI startup to have commercial collaborations with all major cloud

providers and hardware providers. Source: https://twitter.com/ClementDelangue/status/1751982306896519616?utm_source=substack&utm_medium=email

- [53] https://azure.microsoft.com/en-us/blog/microsoft-and-mistral-ai-announce-new-partnership-to-accelerate-ai-innovation-and-introduce-mistral-large-first-on-azure/
- [54] https://fourweekmba.com/openai-microsoft/
- [55] https://www.handelsblatt.com/technik/ki/kuenstliche-intelligenz-mistrals-naechster-coup-schlaegt-frankreichs-ki-star-jetzt-openai-01/100016986.html
- [56] https://www.cnbc.com/2023/10/27/google-commits-to-invest-2-billion-in-openai-competitor-anthropic.html
- [57] https://www.geekwire.com/2023/amazon-pours-up-to-4b-into-ai-startup-anthropic-boosting-rivalry-with-microsoft-and-google/
- [58] https://www.infoworld.com/article/3711468/ai-alliance-led-by-ibm-and-meta-to-promote-open-standards-and-take-on-aws-microsoft-and-nvidia.html
- [59] https://llama.meta.com/
- [60] Nvidia's position in markets relating to AI has led to increased interest in our business from regulators worldwide, including the European Union, the United States, and China. For example, the French Competition Authority collected information from us regarding our business and competition in the graphics card and cloud service provider market as part of an ongoing inquiry into competition in those markets. We have also received requests for information from regulators in the European Union and China regarding our sales of GPUs and our efforts to allocate supply, and we expect to receive additional requests for information in the future". https://d18rn0p25nwr6d.cloudfront.net/CIK-0001045810/7df4dbdc-eb62-4d53-bc27-d334bfcb2335.pdf
- [61] https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand/
- [62] https://gpus.llm-utils.org/nvidia-h100-gpus-supply-and-demand/
- [63] https://www.businessinsider.com/openai-boss-sam-altman-wants-his-own-chip-supply-2024-1
- [65] An interesting question is how many ecosystems are needed for competition. Economic literature provides some answers in the context of

more traditional industries. For example, Bresnahan and Reiss (1991) studied five retail and professional industries and showed that in concentrated markets, only the first five market players matter for competition, the sixth entrant has little effect on competitive conduct. Bresnahan, T. F., & Reiss, P. C. (1991). Entry and competition in concentrated markets. Journal of political economy, 99(5), 977-1009.

- [66] Microsoft said that it will increase its investments in the deployment of specialized supercomputing systems to accelerate OpenAI’s AI research and integrate OpenAI’s AI systems with its products while “introducing new categories of digital experiences.” https://techcrunch.com/2023/01/23/microsoft-invests-billions-more-dollars-in-openai-extends-partnership/

- [67] https://www.reuters.com/technology/investors-punish-microsoft-alphabet-ai-returns-fall-short-lofty-expectations-2024-01-31/

- [68] https://www.spiegel.de/wirtschaft/unternehmen/nvidia-und-die-ki-wie-jensen-huang-microsoft-amazon-und-google-attackiert-a-094ebb7f-17ad-43cc-9a87-0a7644bd8437

- [69] Genakos, C., & Pagliero, M. (2022). Competition and pass-through: evidence from isolated markets. American Economic Journal: Applied Economics, 14(4), 35-57.

- [70] Fumagalli et al. (2020) show that a takeover by the incumbent may be pro-competitive if the incumbent has an incentive to develop a project that an independent startup would have not been able to pay for. Fumagalli, C., Motta, M., & Tarantino, E. (2020). Shelving or developing? The acquisition of potential competitors under financial constraints.

- [71] Loertscher, S., & Reisinger, M. (2014).

- [72] In November 2023, Bundeskartellamt assessed the partnership between Microsoft and OpenAI under the German merger control law. In December 2023, the UK’s Competition and Markets Authority said that it was scrutinising the partnership between Microsoft and OpenAI under UK merger control rules. In January 2024, FTC started to scrutinise Big Tech investments in OpenAI and Anthropic. Google, Amazon and Microsoft must produce documents describing the rationale behind their investments. In January 2024, the European Commission announced that it is looking into some of the agreements that have been concluded between large digital market players and generative AI developers and checking whether Microsoft’s investment in OpenAI might be reviewable under the

EU Merger Regulation. Most recently, media reported that the European Commission plans to look into Microsoft's new partnership with French artificial intelligence company Mistral.

- [73] Katz, M. L., & Shapiro, C. (1994). Systems competition and network effects. Journal of economic perspectives, 8(2), 93-115.
- [74] Bourreau, M., & Krämer, J. (2022). Interoperability in digital markets. Available at SSRN 4181838.
- [75] Ibidem.
- [76] Bedre-Defolie Ö. & R. Nitsche (2020). When Do Markets Tip? An Overview and Some Insights for Policy. Journal of European Competition Lae & Practice. Vol 11. No. 10, pp. 610- 622.
- [77] Katz, M. L., & Shapiro, C. (1994).