THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, DC

GW Competition & Innovation Lab

No. 2025/18

Foundation Models and Generative AI Applications: What Competitive Concerns?

Antonio Manganelli

The GW Competition & Innovation Lab
Suite 621, 6th Floor, 805 21th Street NW
Washington, DC 20052
gwucil@email.gwu.edu

Foundation Models and Generative AI Applications: What Competitive Concerns?

Antonio Manganelli[♣]

Table of contents

Abstract This article analyses the competitive problems potentially arising in the foundation models and generative AI value chain, by exploring its structural complexities and distinguishing between upstream and downstream dynamics. While upstream markets, such as cloud services, proprietary data, and AI chips, have drawn much of the competition policy attention, tipping dynamics, consumer lock-in and foreclosures are most likely to arise in the downstream part of the value chain. This is particularly relevant for mobile ecosystems, where the integration of foundation models with operating systems, on one side, and the development of agentic systems, on the other side, significantly enhance those competitive risks. Consequently, as a complement to competition law scrutiny, ex-ante regulatory intervention seems necessary to ensure market fairness and contestability, according to principles already embedded in the Digital Markets Act.

Keywords: generative AI, foundation models, competition policy, mobile ecosystems, operating Systems, AI agents, gatekeepers

[♣] Professor of Competition Law and Policy - School of Economics and Management at the University of Siena, Italy.

1. Introduction: genAI systems and potential competitive issues

Artificial Intelligence (AI) is rapidly transforming industrial, economic and social systems, by fundamentally revising the structure of human-computer interactions and their impact on human behaviours. Traditional AI refers to systems designed for specific tasks that apply rule-based procedures and follow learning modes based on specified conditions for producing intended outcomes. In contrast, general-purpose AI represents a model with general capabilities used for different tasks across a variety of domains, more similarly to human cognition and "creativity".

The EU Artificial Intelligence Act (AI Act) defines a general-purpose AI model as "*trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications*."[1]

Generative AI (GenAI) is a subset of modern AI focused on creating new content - text, images, music, or code - based on statistical patterns learned from large datasets. Indeed, the same AI Act affirms that "*large generative AI models are a typical example for a general-purpose AI model, given that they allow for flexible generation of content, such as in the form of text, audio, images or video, that can readily accommodate a wide range of distinctive tasks.*"[2]

An increasing number of businesses and consumers are exploiting this technology. Indeed, genAI has gained immense popularity for its ability to produce "original" content and generate "human-like" responses, assisting in creative activities and enhancing productivity. GenAI applications cover a wide range of services and markets, e.g., customer support chatbots (e.g., ChatGPT by OpenAI, IBM watsonx Assistant); virtual assistants (e.g., OpenAI ChatGPT integrated into Microsoft Copilot, Amazon Alexa AI); search engines (e.g., Perplexity AI, Google Gemini); social networks (e.g., Meta AI Assistant, Snapchat My AI); productivity software (e.g., Microsoft 365 Copilot; Google Duet AI); image and video creation tools (e.g.,

[1] Art 3 (63) AI Act. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)

[2] Recital 99 AI Act.

Adobe Firefly, DALL·E by OpenAI); audio and music generation (e.g., ElevenLabs, Suno.ai); code generation and developer tools (e.g., GitHub Copilot, Amazon CodeWhisperer).

GenAI models are often built upon foundation models (FMs): large-scale, pre-trained models capable of adaptation to a variety of tasks. These models are trained on very large and diverse datasets to constitute the basis for a wide range of applications. FM can be trained with different type of data, thus defining its scope of action: large language models (LLMs) are trained on texts; image generation models are trained on images (accompanied with descriptive text); or multiple types of data can be used for multi-modal FM.

The genAI production cycle typically is composed of three main phases: (i) training, (ii) fine-tuning, and (iii) inference. Training on broad and diverse datasets, as mentioned, empower the model with its general capabilities, while fine-tuning (if present) further trains the model on industry-specific or task-specific data to enhance performance for particular use cases. Finally, during the inference phase, the model is made accessible via applications to end users: users provide inputs, and the model applies what it has learned to generate real-time outputs such as creative content, predictions, or decisions. The model may also be logged for future refinement through techniques such as "reinforcement learning from human feedback" (RLHF).

Due to the huge data requirements, it is usually not effective to train and run these models on standard computer chips, e.g., Central Processing Units (CPUs). Indeed, specialised chips have been developed to accelerate computing and executing multiple operations in parallel. Namely, Graphical Processing Units (GPUs), which were originally designed for image processing in gaming, are the hardware accelerator that are mostly utilized. Moreover, a variable multitude of GPUs are used simultaneously to significantly enhance computational performance and efficiency.[3]

Therefore, training a FM requires both intense computational power and, usually, very large datasets. Due to the high costs of building the necessary hardware and software infrastructures, most FM developers are not making such large upfront investments and y rely on specialized cloud computing services. Cloud Service Providers (CSPs) must, in turn, acquire and deploy vast quantities of AI-optimized chips, (typically GPUs)

[3] For example, to train OpenAI's GPT-3 with 175 billion parameters 1024 NVIDIA A100 GPUs were used in parallel for several weeks. Parameters represent values "learned" and "adjusted" from data, that is what the model "knows". Therefore, more parameters imply potentially greater capacity to learn but also the need of more data, compute, and memory.

to meet the computational demands of model training and inference at scale.

Cloud computing services are not only a critical upstream input for training FMs, but they also play a central role in the downstream distribution of these models. Indeed, cloud infrastructures act also as interfaces between FM developers, deployers, and end users - facilitating both the release of FM-based services and their inference in real time.[4] The deployment of FMs in consumer-facing services also forms a key component of the AI stack's lower layer. GenAI applications arise either as enhancements to existing services by integrating FMs, or as GenAI-native applications, i.e., entirely new standalone products built around generative capabilities.[5]

As outlined above, this paper adopts a simplified distinction between upstream FM development and downstream FM deployment. Upstream FM development refers to the stage in the supply chain where FM developers build and train foundation models. Downstream FM deployment refers to integration of models into services and applications, as well as their distributed and usage, involving end-user interfaces or consumer-facing products. (figure 1)

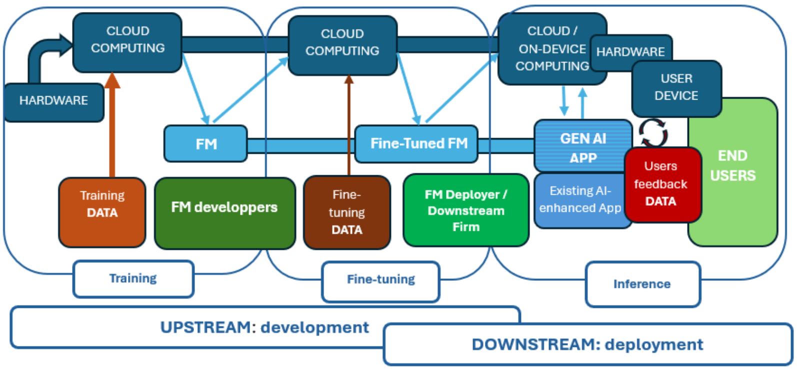

Figure 1 – A simplified version of the composite FM/genAI value chain

This two-tiered framework helps to distinguish roles within the AI value chain; these layers can overlap. The complexity of the AI stack, particularly in relation to the interdependencies between FM developers and downstream firms, implies that companies may operate across both layers. In some cases, FM developers also act as service providers, integrating their own

[4] At the lower layers of the AI stack, emerging alternatives to cloud computing is on-device computation, enabled by specialized AI chips embedded in consumer devices. These chips allow for local storage and inference of FMs, potentially reducing dependence on cloud-based services for certain applications- particularly where latency, privacy, or offline functionality are crucial.

[5] In practice, this distinction is often not clearcut, as many services may include elements of both.

models into end-user platforms and competing directly in downstream markets.

On this basis, the GenAI value chain can be broadly described as comprising multiple actors providing complementary inputs for the development and deployment of foundation models (FMs). These include: (i) FM developers; (ii) hardware component suppliers (e.g., AI chip manufacturers); (iii) cloud service providers (CSPs); (iv) data sources or curators; and (v) application providers that integrate FM capabilities into end-user products.

Within this complex ecosystem, a key concern from a competition policy perspective is whether any company might leverage its existing market power in one or more segments of the AI value chain to gain a competitive advantage in the FM segment.[6] This may occur at various levels of the value chain and may take the form of vertical integration, as well as may happen through exclusive partnerships or preferential access arrangements.

Regulatory debate increasingly focuses on Big Tech firms or other dominant digital tech incumbents since these vertically integrated players are active across multiple layers of the GenAI stack, i.e., hardware, cloud services, data, and, obviously, applications that integrate - as described - FMs into a variety of consumer and enterprise-facing services.

As for FM development, AI-native startups such as OpenAI and Anthropic achieved a first-mover advantage relative to Big Tech[7], but all major technology firms are now actively developing their own foundation models. In addition, strategic partnerships between major technology firms and FM developers have taken place, exerting significant influence on the structure and evolution of the GenAI market[8] - most notably,

[6] M. von Thun, D. Hanley (2024) Stopping Big Tech from Becoming Big AI: A Roadmap for Using Competition Policy to Keep Artificial Intelligence Open for All. Open Markets Institute 2024, Available at SSRN: https://ssrn.com/abstract=4990780 or http://dx.doi.org/10.2139/ssrn.4990780

[7] Also a dozens of innovative start-ups are developing proprietary and open-source models, such as Hugging Face, Mistral AI, Aleph Alpha, EleutherAI, Cohere, Adept, Midjourney, AI21 Labs, Technology Innovation Institute, Jasper, Inflection, Perplexity etc.

[8] As training costs continue to grow, collaboration - not competition – could become the dominant organizing principle in the AI sector, necessitating a reassessment of current competition policy paradigms. See B. Martens (2024) Why artificial intelligence is creating fundamental challenges for competition policy. Bruegel Policy Brief No. 16/2024.

Microsoft with OpenAI[9], Google with Anthropic[10], and Amazon with Anthropic[11]. Such partnerships can have positive effects, since FM developers benefit from access to the computational resources, cloud infrastructure, and financial backing provided by Big Tech firms, which - in return - gain early access, and/or potential exclusive integration rights for cutting-edge AI capabilities.[12] However, these partnerships also raise competition concerns, since dominant digital firms may further consolidate their market position, foreclose rivals, or inhibit the emergence of alternative models and innovation pathways.[13]

In this context, the UK Competition and Markets Authority has expressed concerns that incumbents might limit competition among FM, reasoning that "*the growing presence across the foundation models value chain of a small number of incumbent technology firms, which already hold positions of market power in many of today's most important digital markets, could profoundly shape these new markets to the detriment of fair, open and effective competition, ultimately harming businesses and consumers, for example by reducing choice and quality and*

[9] This partnership entails the following: (i) Microsoft has invested over $13 billion in OpenAI; (ii) OpenAI's GPT models (including GPT-4) are exclusively available via Microsoft Azure; (ii) Microsoft integrates OpenAI's models into its products (e.g., Copilot in Microsoft 365, GitHub, and Windows). In April 2025, the CMA concluded its inquiry, looking at the partnership between Microsoft and OpenAI, focusing on Microsoft's increasing investment and commercial agreements with OpenAI. The CMA concluded that there is no clear evidence that those dynamics resulted in Microsoft's effective control on OpenAI qualifiable as a merger under section 22(1) of the Enterprise Act 2002. Nevertheless, the CMA also underlined that this decision does not preclude future investigations. See CMA (2025) Microsoft Corporation's partnership with OpenAI, Inc. Decision on relevant merger situation.

[10] This partnership entails the following: (i) Google has invested over $2 billion in Anthropic (Claude AI); (ii) Anthropic uses Google Cloud's AI infrastructure to train and run its models; (iii) Google integrates Claude into its own services; (iv) Anthropic's models are available via Google Cloud's Vertex AI (Google AI marketplace service).

[11] This partnership entails the following: (i) Amazon committed $4 billion to Anthropic in 2023; (ii) Anthropic's Claude models are hosted and optimized for AWS Bedrock (Amazon's AI marketplace service); (iii) Amazon's AI chips (Trainium and Inferentia) are used for training Anthropic's models; (iv) AWS customers get early access to Claude models.

[12] See for example D. Spulber (2024) Antitrust and Innovation Competition: Investments and Partnerships in Artificial Intelligence, in (edited by T. Schrepel, A. Abbott) Artificial Intelligence and Competition Policy, where the author suggests these collaborations may also serve as innovation-enabling alternatives to full vertical integration.

[13] These dynamics are described also by a recent Federal Trade Commission - Office of Technology staff - report analysing the mentioned partnerships. See FTC (2025) Partnerships Between Cloud Service Providers and AI Developers - Staff Report on AI Partnerships & Investments 6(b) Study. The report found that CSPs often impose cloud spending requirements that lock developers into specific ecosystems, elevating switching costs and raising barriers to entry, thus reinforcing their existing positions of power in both upstream and downstream markets. Moreover, as training costs continue to grow collaboration—not competition—may become the dominant organizing principle in the AI sector, necessitating a reassessment of current competition policy paradigms

increasing price."[14] The French Autorité de la concurrence echoed those concerns because "*major digital companies enjoy preferential access to the inputs needed to train and develop foundation models. ... The vertical integration of certain digital operators and their service ecosystems may give rise to a number of abusive practices*."[15]

In this multifaceted context, this paper seeks to contribute to the ongoing effort to clarify and disentangle the complex and often ill-defined competition concerns arising within the generative AI value chain. Section 2 distinguishes between upstream and downstream segments of the AI stack, identifying potential competition issues at each layer and arguing that the most significant risks are likely to emerge downstream, where FMs are deployed and integrated into end-user applications and platforms. On this basis, the paper explores the interdependent and diverse downstream relationships in the AI stack to identify conditions under which lock-in effects and market tipping dynamics are more likely to emerge. Section 3 contextualizes these risks by focusing on two key areas: (i) the integration of FMs with mobile operating systems (OS), and (ii) the emergence of GenAI agents, as both developments raise critical concerns around market fairness and contestability.

2. Where competitive problems are most likely to arise

2.1. Upstream vs downstream

Across the GenAI value chain, anticompetitive strategies may emerge in both upstream and downstream segments. To date, however, competition policy discussions have primarily focused on upstream concerns, where entrenched market power is held by firms controlling essential inputs for FM development. In these segments, potential competitive leverage is often more direct and observable, as companies may restrict access to key inputs - such as hardware, computational resources, cloud infrastructures, and data - to favour their own FM operations or to compel third-party developers into exclusive partnerships.

Such strategies can create bottlenecks in the supply chain, granting privileged access to critical inputs to proprietary integrated FM or affiliated FM developers, while raising barriers to entry for rivals. In turn, this may shield dominant firms from

[14] CMA (2024) AI Foundation Models: Technical update report.

[15] Autorité de la concurrence (2024) Opinion 24-A-05 on the competitive functioning of the generative artificial intelligence sector

competition and entrench their position across the broader AI stack.[16]

As previously noted, the most powerful and entrenched technology firms are (i) active across nearly all segments of the AI ecosystem and (ii) enjoy preeminent market positions in specific segments controlling critical upstream inputs, such as proprietary AI chips, cloud services, and unique datasets essential for FM training.

As for hardware, all Big techs are developing or investing in AI chips for FM training[17], although another tech giant, NVIDIA, has emerged as the dominant player, by securing a first-mover advantage in the development of high-performance graphics processing units (GPUs) for AI training.[18] Its flagship products, such as the A100 and H100 chips, currently account for over 90% of the market for AI model training accelerators. Moreover, NVIDIA built a strong market position its proprietary software platform which enables developers to optimize code for its GPUs, i.e., CUDA (Compute Unified Device Architecture). CUDA has become the *de facto* industry standard, generating network effects and creating substantial switching costs for NVIDIA's AI chip customers.[19]

Apart from major digital corporations and a handful of companies with extensive in-house data centres, cloud services are the primary means of access to the computational power required for training AI models. CSPs provide developers with infrastructure and platform services tailored to their needs while

[16] B. Martens (2024) Why artificial intelligence is creating fundamental challenges for competition policy, Bruegel Policy Brief, No. 16/2024, Bruegel.

[17] Besides Intel and AMD, all Amazon, Apple, Google, Meta and Microsoft are actively developing proprietary chips for AI model training. Google leads this effort with its Tensor Processing Units (TPUs) which serve as an alternative to Nvidia's GPUs for training its FM including Gemini. In addition, Qualcomm developing AI-specialised mobile chipsets powerful enough to run FMs.

[18] It is worth noting that NVIDIA is a GPU designer and producer, but not a chip manufacturer, meaning it is not a "foundry", i.e., a semiconductor fabrication plant that manufactures integrated circuits. NVIDIA designs its own chips (notably the A100 and H100 GPUs), including architecture, logic, and firmware; while the actual fabrication of the chips is outsourced to third-party foundries, most notably TSMC (Taiwan Semiconductor Manufacturing Company). Therefore, NVIDIA does not operate its own chip foundries, but it controls the product pipeline and branding, making it the producer in commercial and functional terms. Moreover, NVIDIA may collaborate in packaging and assembly phases (e.g., with Foxconn or Amkor), but again it owns the intellectual property (IP) and controls how the product is marketed, sold, and supported.

[19] Competing GPU providers, such as AMD and Intel, face significant challenges in gaining market traction, as most machine learning frameworks are deeply integrated with CUDA and require major engineering effort to port; however, Big techs are investing in order to support interoperability between AI chips, and reduce NVIDIA entrenched market position: for example, AWS provides Neuron, a software development kit, to help customers switch back and forth between third-party AI chips and AWS's AI chips.

eliminating the need for substantial upfront investment in IT infrastructure.[20] As well known, the three largest cloud service providers in terms of market share are Amazon AWS, Microsoft Azure and Google Cloud Platform, commonly referred to as hyperscalers, which are present at all levels of the cloud service value chain (IaaS, PaaS, SaaS)[21]. In 2022, Amazon AWS and Microsoft Azure had market shares between 35% and 40% in the European Union, while Google Cloud Platform had market shares between 5% and 10%.

Data, as described, plays a crucial role in the training phase, since its volume is a key factor to develop high-performance models. Most of this data comes from publicly available sources (e.g., web-scraped info), however, public sources could become insufficient in the future, thus creating significant bottlenecks. [22] This could make proprietary datasets, controlled by a small number of major players, essential for the FM performance. As a result, in principle, FMs developed by big tech firms may gain a competitive advantage due to their exclusive access to large-scale training data generated within their ecosystems of applications.[23]

All these competitive concerns are well-founded. However, in cases of overt exclusionary conduct or denial of access to essential inputs, such behaviours would constitute a conventional exercise of market power, for which a relatively straightforward theory of harm could be developed under existing competition law.[24]

[20] As described in the next section, cloud serves also as a key channel for distributing and deploying FMs downstream.

[21] Infrastructure-as-a-Service (IaaS) is the basic level of service that includes access to IT infrastructure. Platform-as-a-Service (PaaS) represents an intermediate level in the value chain where, compared to the previous configuration, middleware is added. Software-as-a-Service (SaaS) indicates the highest level in the value chain and consists of specific applications that users typically access via a web browser. See A. Manganelli, D. Schnurr (2024) Competition and Regulation of Cloud Computing Services: Economic Analysis and Review of EU Policies - CERRE Report.

[22] See A. Ribera Martínez (2024) Generative AI in Check: Gatekeeper Power and Policy Under the DMA, available at SSRN: https://ssrn.com/abstract=5025742

[23] In this regard, in April 2024, Meta announced to train generative AI models using publicly shared content from users in the European Union, including posts, comments, and interactions with AI systems. Meta emphasized that private messages and data from users under 18 would have been excluded, and that EU users are offered an opt-out mechanism. However, the decision to rely on "legitimate interest" as the legal basis for data processing—under Article 6(1)(f) of the General Data Protection Regulation (GDPR)—has been criticized as this choice may circumvent the requirement for explicit opt-in consent required by Article 6(1)(a) for sensitive personal data. Further concerns are about the complexity of the opt-out process, which may not comply with GDPR transparency and user control rules.

[24] Nevertheless, as previously mentioned, strategic partnerships and mergers needs a particular scrutiny. Likewise, the dual dominance by NVIDIA in both hardware design and software tooling creates strong network effects and introduces substantial barriers

More profound concerns stem from a broader, systemic observation: that technology markets - particularly those governed by network effects and data feedback loops - exhibit a strong propensity to "tip" toward winner-takes-all outcomes, allowing big techs to build entrenched market power positions in adjacent digital markets, shielding themselves from inter-platform competition. [25]

This assumption is grounded in the experience of the past two decades, during which competition authorities often failed to intervene early or robustly enough to prevent the emergence of entrenched positions by dominant digital platforms. As a result, regulators now seek to apply the "lessons learnt over the last 10 to 15 years" [26] from the Web 2.0 era to the evolving generative AI ecosystem. The central competition policy question, therefore, is whether and under what conditions the competitive dynamics of the genAI value chain may replicate those of previous digital markets.

In the first place, it seems unlikely that market power in highly innovative markets could be derived and entrenched on a lasting basis by controlling an upstream technological input, because persisting strong innovation dynamic in all these segments. Recent advancements support this argument from two perspectives.

First, the emergence of Deepseek and similar FMs show the potential for a smaller reliance on extensive computational resources. Second, there is a growing trend toward the development of smaller models that require fewer resources for training and deployment, in terms of both data and computing power. [27] Moreover, the shift toward smaller FMs - along with production of specialised AI chipsets from companies like Qualcomm and Intel - is enabling on-device computing and deployment of FMs on consumers' devices at the edge. [28] This

to entry for rival chipmakers. See J Vipra, S Myers West (2023) Computational Power and AI, AI Now Institute.

[25] About competition concerns for "traditional" digital markets and platforms, see A. Manganelli, A. Nicita (2022) Regulating digital markets: the EU approach.

[26] Remarks by Sarah Cardell, CEO of the CMA, delivered during the 72nd Antitrust Law Spring Meeting. Washington DC, USA.

[27] Beside Deepseek, Mistral AI is an example of an existing model which is small but effective for certain use cases. See Mistral AI, Mistral NeMo: our new best small model (July 18, 2024). Moreover, Microsoft's 'small language model' Phi-2. See also Microsoft Research Blog, Phi-2: The surprising power of small language models (Dec. 12, 2023); WSJ, For AI Giants, Smaller is Sometimes Better (July 6, 2024). Other examples of small FMs released recently include Google's Gemma 7B, Hugging Face's Zephyr 7B

[28] See B. Edwars (2024) Apple releases eight small AI language models aimed at on-device use (Ars Technica, April 25, 2024)

transformation is further reducing dependence on cloud-based computing infrastructure.

Furthermore, it is crucial to consider that a large part of Nvidia's GPU is currently bought by hyperscalers, giving them a significant countervailing buying power. Furthermore, hyperscalers will continue to develop their own AI chips in order to strategically reduce their dependence on Nvidia and try to commoditize the AI hardware markets. When they are able to start supplying hardware components to other FM developers, and not any for self-consumption, this will develop even a greater competitive pressure.

As for data, big tech companies may have strong limitations in fully exploiting their proprietary datasets and gain a competitive advantage. There is no strong evidence that Google's Gemini or Meta's LLaMA models outperform OpenAI's GPT or Anthropic's models because of data availability. This is because the volume and variety of data is not always a determinative factor for a model performance, when there are highly effective models trained on relatively smaller data sets. Moreover, in the near future, more models will be developed for specific industries like healthcare and manufacturing or for the specific use case and will need industry-specific or case-specific data.[29]

In addition, the add value of proprietary data for FM training is not certain, as a few alternative and third-party data sources can provide the same function. In many cases, training and fine-tuning datasets come from publicly available sources, including web-scraped data and open-source datasets. Moreover, data partnerships,[30] paid licensing agreements, or the use of synthetic data[31] can achieve similar training outcome compared to proprietary data. Finally, big tech firms are increasingly constrained by regulatory frameworks such as the General Data

[29] See A. Chandrasekaran (2024) *3 Bold and Actionable Predictions for the Future of GenAI* (Gartner, Apr. 12, 2024) predicting that by 2027, more than 50% of generative AI models will be specific to either an industry or business function, up from approximately 1% in 2023.

[30] For example, OpenAI partnered with publishing house Axel Springer to give ChatGPT users access to real-time summaries of Axel Springer content.

[31] Synthetic data is data which has been generated artificially, for example data prodcued with simulations or using existing AI models to generate new data sets. Synthetic data can complement real-world data for improving AI models. For example, Amazon used synthetic data to train Amazon One (a biometric payment system introduced by Amazon in 2020 in USA that allows users to pay for purchases or verify their identity by scanning their palm). Since Amazon only had a small amount of palm data, they used genAI to create millions of synthetic images of palms. See *How generative AI helped train Amazon One to recognize your palm* (September 1, 2023). See also *What is Synthetic Data?* (on AWS's website) and AWS's *Synthetic Data Specialty Practice*

Protection Regulation (GDPR) [32] and the Digital Markets Act (DMA), [33] which impose restrictions on how they can lawfully collect, process, and exploit user data, further limiting their ability to exploit data as an exclusive competitive asset.[34]

Overall, the upstream segments of the AI value chain do exhibit a tendency toward market concentration,[35] yet there is an ongoing conglomerate competition among major tech firms operating in these segments. This is very different from the situation in the web 2.0 where a much net segmentation of the digital sphere has taken place, i.e., devices: Apple; search: Google; e-commerce: Amazon; comms: Meta; productivity: Microsoft.

Whereas all big techs have actively invested in almost all AI-related tech markets, often with significant overlaps, alongside other firms that specialize in specific AI segments, and none has established dominant control over a specific segment. As a result, competition persists across different segments of the genAI value chain, with incumbent firms strategically working to limit their rivals' ability to expand and consolidate market power. This interplay between large tech firms fosters multi-market interactions, [36] which may, in turn, dampen the intensity of "competition for the market"[37] and mitigate the winner-takes-all dynamics that were prevalent during the Web 2.0 era.

[32] Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L119/1 (GDPR).

[33] Regulation (EU) 2022/1925 of 14 September 2022 on contestable and fair markets in the digital sector and amending Directives (EU) 2019/1937 and (EU) 2020/1828 (Digital Markets Act)

[34] A valid counter-point here to consider is that regulatory compliance creates a chilling effect for smaller actors, while Big Tech could bare the risk to challenge regulation or anyway try to twist it/interpret it to their own favour. Likewise, the tech giants are also better positioned than smaller rivals to negotiate licensing agreements with copyright owners, or where this fails, to resolve, evade, or absorb the legal and financial consequences of inappropriately exploiting copyright-protected material to train AI models. Few companies other than the tech giants are willing and financially capable of taking on so much legal risk. See M. von Thun, D. Hanley (2024) Stopping Big Tech from Becoming Big AI: A Roadmap for Using Competition Policy to Keep Artificial Intelligence Open for All - Open Markets Institute 2024, Available at SSRN: https://ssrn.com/abstract=4990780

[35] See A. Korinek, J Vipra (2025) Concentrating intelligence: scaling and market structure in artificial intelligence, Economic Policy, Volume 40, Issue 121, January 2025, Pages 225–256

[36] Multimarket contact refers to the situation in which more than two firms simultaneously compete in multiple products markets. Most studies on multimarket contact have explored how the market overlap creates "mutual forbearance". This lessens the intensity of rivalry by establishing cooperation/coopetition dynamics. See B. D. Bernheim, M. Whinston (1990) Multimarket Contact and Collusive Behavior, in The RAND Journal of Economics 21, no. 1 (1990): 1–26.

[37] Competition for a market refers to definition of new dominant standards or business models tending to a monopoly market structure and usually associated with the process of innovation that brings new displacing technologies to market. See P.

As a result, the competitive dynamics in the upstream AI value chain appear to differ largely from those observed in "traditional" digital markets, such as search engines, web browsers, and social media, and may lead to very distinct market outcomes. Notably, in such markets tipping effects and entrenched positions of major tech companies have largely been sustained by consumer lock-ins, which were in turns driven by data feedback loops and network effects,[38] rather than by any inherent superiority in efficiency or quality arising from exclusive access to upstream resources.

Similarly, the most significant competitive risks in the genAI landscape are likely to emerge downstream, particularly in the distribution and deployment of FMs via genAI applications that directly interact with end users.

Furthermore, it is important to underline that the expansion of genAI systems may have a disruptive effect on "traditional" digital markets currently dominated by Big Tech firms. This is particularly true in the application layer of the AI value chain, where GenAI could facilitate the emergence of substitute services for incumbent digital platforms. In some cases, GenAI may also enable new modes of delivering existing services, therefore potentially completely bypassing traditional gatekeepers.

A clear example of this dynamic can be found in the search market, where Google have had a long-standing dominant position. Some applications based on FM start delivering search results by providing real-time web access enriched by sources and conversational interactions. One example is Perplexity AI, which is challenging the way of working of traditional search engines, as well as their effectiveness based on a static list of links. Perplexity does not develop its own FM but instead

Geroski (2003) Competition in Markets and Competition for Markets, in Journal of Industry, Competition and Trade, 2003, vol. 3, issue 3.

[38] A data feedback loop is a self-reinforcing process made of quality enhancements (objective and/or subjective) based on customer data. If the quality enhancement is subjective, i.e., for a specific user, the relevant data depend on that customer's usage of that service: therefore, a customer, by using a service, enhances its quality under a subjective perspective, and could be individually locked-in. If the quality is objectively enhanced, i.e., for all users, the relevant data depends on the overall usage of the service (and thus on the number of users): therefore, above a certain threshold, a winner-takes-it-all dynamics may take place (collective lock-in). Indeed, objective feedback loop are network externalities based on data, where the value for each user to belong to a certain network depends on the overall number of network's users and the overall amount of data provided. See A. Hagiu, J. Wright (2023) Data-enabled learning, network effects and competitive advantage. RAND Journal of Economics 54 (4), 638–667.

integrates multiple third-party models, including those from OpenAI and Anthropic.[39]

These kinds of innovations reflect the inherently pro-competitive potential of generative AI, particularly in facilitating innovative and potentially disruptive market entry. However, this same dynamic may incentivize dominant firms to adopt defensive leveraging strategies aimed at preserving their incumbency. In such cases, downstream anticompetitive conducts may be even more strongly motivated and driven by strategic efforts to pre-empt or neutralize the disruptive effects of genAI applications.

Some competition authorities have begun to recognize potential competitive risks emerging downstream, however, for the moment such assessments have often remained high-level and primarily focused on the gatekeeping roles of Big Tech firms. However, a more granular approach seems necessary to differentiates between various AI-enabled downstream services and business models. This would allow to identify downstream market contexts where tipping dynamics, consumer lock-in, and foreclosure risks are most likely to arise.

Furthermore, recognizing these differences is essential for ascertaining whether and to what extent existing regulatory frameworks, such as the EU Digital Markets Act (DMA), may already be applicable. Indeed, the DMA’s ex-ante obligations are designed to preserve contestability and fairness in digital markets and may be particularly relevant where GenAI is applied by gatekeepers in their core platform services, reinforcing existing structural advantages.

2.2. Interplay between FMs and GenAI applications

Describing the complex interplay between FMs and GenAI applications is not straightforward. FMs are typically general-purpose, meaning they can be applied across a wide range of use cases and integrated into numerous applications. However, more specialized FMs can be developed on top of these general models, by a fine-tuning process, enhancing performance for specific tasks. In some cases, GenAI applications can be built using multiple FMs.

A key distinction - regulatory and economic rather than purely technical - lies in the modes a GenAI application is deployed. At one end, GenAI functionalities may be integrated into existing

[39] Other players, such as You.com, combine conventional search results with AI-generated summaries to deliver context-aware, conversational outputs. Similarly, Net.com offers customizable search preferences, allowing users to prioritize sources (e.g., Reddit, scholarly papers, news media) and interact via a chat-based interface.

digital services, such as search engines (e.g., Bing integrating GPT-4), productivity platforms (Microsoft Copilot in Word and Excel), or social networks (Snapchat's My AI chatbot). At the other end, stand-alone, AI-native applications are emerging as independent platforms built entirely around generative capabilities - examples include ChatGPT or Perplexity AI.

Within this multifaceted landscape, downstream firms have several strategic options for accessing and utilizing FMs:

- Developing an in-house FM: this approach offers maximum control over model architecture, data governance, and deployment. It has been pursued by Google (Gemini), Meta (LLaMA series), Mistral, and xAI (Grok)[40]. This is obviously resource-intensive, requiring massive compute infrastructure and proprietary or licensed training datasets.
- Fine-tuning a third-party FM: FM developers provide access to open-weight models[41] (such as OpenLLaMA by Meta, GPT Models by OpenAI, Falcon LLM by Hugging Face, or Bloom by TII). Downstream firms can fine-tune those models for specialized use cases, enabling customization and differentiation, but switching costs may emerge.
- API-based access to a third-party FM: many downstream companies rely on API access to FM models hosted by major CSPs. For example, OpenAI's GPT-4 is accessible via Azure, Anthropic's Claude via Amazon Bedrock, and Google's Gemini via Vertex AI. These access modes enable downstream firms to integrate GenAI features into their apps without training or hosting the models, thus providing cost-effective scalability but introducing dependency on upstream providers.
- Using FM plug-ins or extensions: platforms such as OpenAI's ChatGPT plug-in ecosystem or Zapier AI provide software components or modules that allow to

[40] Grok is a large language model (LLM) and AI assistant developed by xAI, the artificial intelligence company founded by Elon Musk in 2023. It is closely integrated with X (formerly Twitter and is marketed as a direct competitor to models like OpenAI's ChatGPT, Anthropic's Claude, and Google's Gemini. Key Features of Grok is its ability to pull real-time data from the X platform, giving it more up-to-date responses than many competitors, which rely on static training data.

[41] Not all foundation models (FMs) available for fine-tuning are open-source. Some are open-weight models, while others are fully open-source models, and there is a significant distinction between the two: open-source models are released under permissive licenses (e.g., Apache 2.0, MIT) that provide unrestricted access to the model's code, architecture, and weights. Users can inspect, modify, redistribute, and fine-tune the model as they see fit. Open-weight models provide access to the pre-trained weights but are not fully open-source, as they user access is limited to pre-trained weights, while code is often unavailable. Moreover, they usually come with restrictive licenses that limit how they can be used (e.g., non-commercial use only) or prohibit redistribution.

embed FM capabilities into existing apps or platforms. This enables downstream companies to enhance existing workflows or platforms without managing FM in any way.

Downstream actors may operate either as end-users of FMs - i.e., by using them for internal applications and functionalities or as business users, i.e., by embedding FM capabilities into consumer-facing applications.

Finally, on the consumer side, GenAI services can be accessed through various modalities: (i) via dedicated apps or websites; (ii) through specific products or services that are bundled with GenAI services or embed GenAI functionalities; or (iii) by downloading GenAI-native applications from app stores. In many cases, the same FM is made available to users across multiple - or even all - of these access modes.

As highlighted earlier, the downstream segment of the AI value chain - where FMs are deployed, and their inference capabilities are leveraged by GenAI applications for end-users - is particularly critical from a market contestability and fairness perspective. However, this segment has received limited scrutiny due to its complexity and the lack of clear insights into how Big techs might replicate the competitive and market dynamics observed in "traditional" digital markets.

Given the diversity of GenAI applications and services that utilize general-purpose FMs - across various sectors, markets and services - it may not be straightforward for dominant firms to extend their market power from the downstream application layer to control the FM landscape itself. Therefore, despite many important digital markets and services are dominated by big-tech companies, it appears unlikely to leverage market power from such a composite downstream application layer into the FM.

Moreover, differently from what happen in "traditional" digital markets, data feedback loops within the AI value chain and across its segments may be not that strong.[42] Primarily, it has been observed that user feedback data is not automatically fed back into the model and most of time would be very expensive to do so.[43] Moreover, genAI applications do not always originate so strong and accurate users' signals to create meaningful feedback.[44] Finally, as highlighted, multiple successful and competing general-purpose FMs exist, making it unlikely that

[42] A. Hagiu, J. Wright (2025) Artificial intelligence and competition policy, International Journal of Industrial Organization.

[43] CMA (2023) AI Fundation Models: initial Report.

[44] For example, a chatbot's answer could not receive any feedback or a thumb up which is much weaker feedback compared, for example, to a choice of a specific link/product/feed among many proposed by any kind of algorithmic recommendation system.

user data collected from a single application - even within big tech’s vast customer base - would provide a significant competitive advantage in the FM market.

Likewise, indirect network effects do not appear to be always major factor in the downstream FM market. Unlike traditional multi-sided digital platforms, where user choices influence other market participants, GenAI applications built on FMs operate in a linear pipeline structure. This means that end-users select a GenAI-powered application independently, without their choice directly affecting other users or market-side dynamics in the same way multi-sided platforms do.

Nevertheless, under certain conditions AI downstream segments can be subject to “platformisation” ,i.e., replicate a platforms’ economic and business models. This occurs particularly when genAI applications and devices are commercially and technically integrated in a way that makes them FM-specific. In such scenarios, users may indirectly become reliant on a specific FM - not by actively selecting the FM itself, but because all GenAI applications they access are necessarily based on that model.

This situation may arise when a FM is integrated within an entire ecosystem, for instance an Operating System (OS) or a device, restricting end-users’ ability to engage with alternative AI models outside of that ecosystem. Such integrations can be either structural, where a single company develops and provides both the FM and the OS/device, or commercial, where some forms of exclusive agreements exist between OS or handset manufacturers and FM developers. In this scenario, the FM value chain, which is mostly linear, begins to have some characteristics typical of platforms, i.e., develop cross-network externalities and an increased likelihood of market tipping.

OS/device ecosystems may significantly influence FM deployment, by prioritizing their own FM services through seamless integration, preferential accessibility, and enhanced compatibility. Therefore, they can potentially restrict consumer choice, creating artificial barriers that make it impossible or strongly discourage users from switching to applications based on alternative FM.

Another form of "platformization" of GenAI downstream value chain is related to cloud computing services. While cloud computing plays a crucial role as an upstream input for training FMs, it also serves as a key distribution channel in the downstream market. Specifically, CSPs act as intermediaries between FM developers, FM deployers, and end-users, allowing both FM service deployment and inference.

At the downstream level, most firms rely on cloud computing to fine-tune FM and to run FM services and often they access those services through “platforms” configured as “FM marketplace”, managed by the largest CSPs - Amazon, Microsoft, and Google. For example, Microsoft offers OpenAI’s foundation models via the Azure OpenAI Service, as well as within its existing enterprise cloud solutions, including Dynamics and Power Platform; additionally, businesses can deploy FMs from other providers, such as Hugging Face, using Azure’s infrastructure. Google provides access to FMs through its Google Cloud Platform. Amazon’s AI platform, Bedrock, provides access to leading foundation models, (including AI21 Labs, Anthropic, Cohere, Meta, Mistral AI, Stability AI, Deepseek and Amazon itself). Most Amazon Bedrock’s customers use more than one model, combining advanced models for complex tasks with simpler models for basic, quick tasks. Moreover, Amazon SageMaker provides access to proprietary and open-source models that customers can incrementally train and fine-tune.

Currently, marketplace allow deployers to choose from a broad range of platforms, thus reducing transaction costs and enhancing the market well-functioning and AI models diversity. However, in these environments, self-preferencing strategies by major CSPs may eventually emerge, leading to: (i) preferential treatment of their own proprietary or partnered FMs over competing models;[45] (ii) favouring their own downstream FM services over third-party alternatives.

Indeed, Google, AWS, and Microsoft not only supply essential computing resources but also compete directly in the downstream FM services market, offering their own user-facing AI applications. In addition, hyperscalers play a crucial role in enabling end-user access to FM services, as well as facilitating real-time data retrieval for inference.

Therefore, CSP-integrated genAI models can induce cumulative network effects, where user feedback is leveraged to implement models’ improvements, and introduce new services.[46] In turn,

[45] As for Amazon Bedrock, currently FM developers should be able to port models built or trained on AWS to another cloud provider or on-premises. For example, a developer can start training a model on AWS (e.g., with Amazon SageMaker), take the model to another IT services provider, and continue training on that other provider from a checkpoint they have set in the model.

[46] See T. Schrepel, A. Pentland (2024) Competition between AI foundation models: dynamics and policy recommendations, in Industrial and Corporate Change, 2024, 1–19.

From this perspective, the deployment of FMs directly on devices (on-device AI), or a hybrid approach that combines on-device and cloud-powered AI, offers several advantages. Besides improving performance, resilience, and security, this approach enhances privacy by reducing reliance on centralized data collection. Consequently, this evolution could mitigate competitive concerns related to the data feedback loop within CSP-controlled ecosystems, promoting a more balanced AI landscape.

data feedback loops may have a much greater impact, as these systems directly benefit from continuous user interactions, allowing to fine-tune and customize models more efficiently than independent FM developers.

These competitive dynamics should be viewed as an "ecosystem effect" rather than isolated market behaviours. Therefore, CSPs that also control OS can significantly expand their dominance across the AI value chain, reinforcing their market power.

3. Competitive issues for GenAI applications in the mobile ecosystems

3.1. Integration of FMs and OSs

Big Tech firms can integrate FMs into their own existing products to enhance functionalities. Additionally, they can leverage their vast customer bases to promote stand-alone or AI-native applications developed on top of their proprietary FMs. Their ability to do so at scale - considering their conglomerate dimension[47] - is unmatched by other players and is likely to result in significantly greater increasing returns to scale compared to non-integrated firms.

As previously noted, such conditions do not indicate anticompetitive leveraging per sè. Nonetheless, the relative risk becomes more pronounced within vertically integrated digital ecosystems, particularly when platform-like dynamics and self-preferencing incentives emerge. This is especially relevant in mobile ecosystems, which have been the subject of a few competition law actions, including market studies[48], antitrust enforcement actions[49], and legislative interventions (e.g., the EU's Digital Markets Act) aimed at curbing exclusionary practices and preserving inter- and infra-platform contestability.

As is well established, mobile end-users in the EU - and globally - can primarily choose between two dominant mobile OS: Apple

[47] M. Burreau, A. De Streel (2019) Digital Conglomerates and EU Competition policy – CERRE Report.

[48] See, e.g., Japanese Federal Trade Commission, 'Competition Assessment of the Mobile Ecosystem', (2023); U.S. Department of Commerce, 'Competition in Mobile Application Ecosystem', (2023); UK Competition and Markets Authority, 'Mobile Ecosystems Market Study', (2022); Australian Competition and Consumer Commission, 'Digital Platform Services Inquiry – App Marketplaces', (2021) .

[49]For the most recent antitrust decision, see European Commission, 4 March 2024, Case AT.40437, *Apple – App Store Practices (music streaming)*.

iOS and Google Android, which respectively underpin the two main mobile ecosystems. Mobile OSs are pre-installed, system-level software that are tightly integrated with the underlying hardware. These "bundles" reflect different modes of integration: in Apple's case, iOS is exclusive to Apple devices and fully controlled within a vertically integrated hardware-software stack; by contrast, Google's Android OS is distributed on an open-source basis, but its implementation across most commercial devices is governed by contractual and financial arrangements.[50] Consequently, when consumers purchase a mobile device, they simultaneously make a non-separable decision to enter either Apple's or Google's ecosystem. This includes not only the OS, but also a suite of pre-installed core applications - such as app stores, browsers, and search engines.

As the exclusive providers of the two dominant mobile OSs, Apple and Google exercise significant control over access conditions for downstream service providers. This control extends to decisions on which applications are pre-installed, how they are positioned on user interfaces (e.g., default settings), and which app stores or search engines serve as default intermediaries between users and online content. Mobile OSs can also place limits or restrictions on the channels through which software and applications can be downloaded onto the device. In other words, once that initial purchasing decision is made, subsequent choices become intertwined with it. As a result, once a customer adopts a variety of services within an ecosystem, the exit option can be costly.

Therefore, despite its composite architecture, the global mobile ecosystem is an effective duopoly, where Apple and Google are the two dominant gatekeepers. Each company maintains end-to-end control over its respective ecosystem: Apple through iOS, the App Store, and Safari; Google through Android, Google Play, and Chrome. Their control spans both infrastructure and

[50] Google's control over the Android ecosystem, despite the operating system's open-source core, is largely exercised through two contractual and compliance mechanisms: the Android Compatibility Program (ACP) and the Mobile Application Distribution Agreements (MADAs). Namely, MADAs are commercial agreements that mandate the pre-installation and preferential placement of a suite of Google apps—such as Google Search, Chrome, YouTube, and the Play Store—on certified Android devices. These mechanisms explain why most Android smartphones come preloaded with Google services, a fact that has been central to antitrust scrutiny and platform regulation. In 2018 Android decision, the European Commission found that MADAs constituted an abuse of dominance under Article 102 TFEU, as foreclosing rival search engines and browsers by tying key Google services to Play Store access. Consequently, the Commission required Google to unbundle its apps within the EU market. Decision C(2018) 4761 final, 18 July 2018 (Case AT.40099 – Google Android). These same concerns underpin several provisions in the EU Digital Markets Act (DMA), which seeks to limit self-preferencing, tying, and bundling practices by designated digital gatekeepers

application layers, supported by vertical integration across hardware, software, and services. (figure2)

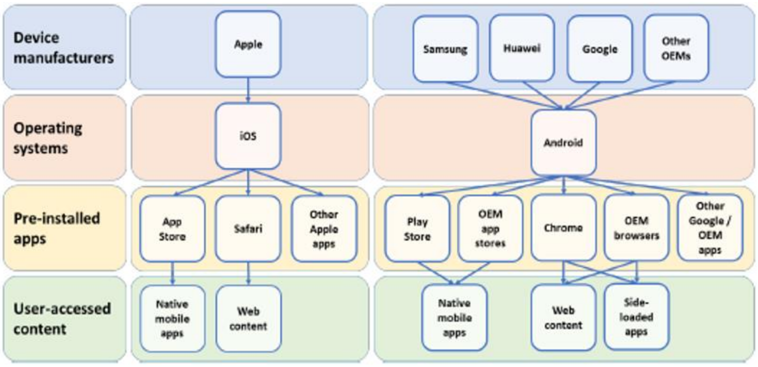

Fig. 2- Apple and Google mobile nested ecosystems, revolving around their OS;
source: CMA (2022) Mobile Ecosystems Market Study

Based on their strategic positions, Apple and Google possess the ability to define access conditions for both end users and business users (e.g., app developers, content providers). This includes setting technical and contractual terms for app distribution, payment systems, and data access. Importantly, both companies also compete directly with business users that rely on their platforms - creating a dual role as platform operator and market participant, which raises structural concerns related to self-preferencing, access discrimination.

For these reasons, the DMA introduces obligations designed to constrain those gatekeepers' behaviours that may result in unfair outcomes for business users and end-users, and to preserve or restore market contestability. However, in the context of mobile ecosystems, the DMA primarily fosters intra-platform competition - that is, competition among third-party developers and service providers operating within a dominant platform - rather than promoting meaningful inter-platform competition between mobile OSs themselves. Indeed, despite the presence of a duopoly, both Apple and Google continue to hold substantial and entrenched market power in mobile operating systems, as the degree of effective competition between the two ecosystems is limited.[51]

[51] This is mainly due to (i) the supply of mobile devices and operating systems has segmented into broadly two groups – higher-priced (Apple's iOS devices) and lower-priced devices (Android devices); (ii) users rarely switch between iOS and Android devices – with material perceived barriers to switching. Moreover, when the entry point is a physical product (the smartphone) inter-platform competition is per se much more difficult (compared to a fully virtual ecosystem) as multihoming is not feasible, likewise a "classical" network industry. CMA (2022) Mobile ecosystem market study.

In this context, the vertical integration of FMs into mobile OSs - whether through structural bundling or exclusive agreements - raises significant competition concerns. As previously discussed, this integration mirrors earlier competition issues stemming from the bundling of OSs with downstream applications such as app stores, browsers, and search engines. By embedding FMs directly into the OS layer, gatekeepers could extend their market power from the operating system into the genAI layer, potentially requiring developers of GenAI applications to rely on the gatekeeper's proprietary FM when they want to access users within that ecosystem.

More broadly, if dominant OS providers were to fully integrate FMs into core system functionalities, this could undermine not only inter-platform competition, but also intra-platform competition. Indeed, two primary risks would emerge: (i) the reinforcement of self-preferencing practices throughout the mobile ecosystem, thereby diminishing the visibility and competitiveness of third-party app developers, independent app stores, and alternative AI service providers; and (ii) the restriction of user choice, as consumers may face barriers to selecting and using applications based on FMs independent of their device's OS. These dynamics would further consolidate the gatekeepers' positions and frustrate the objectives of the DMA.

On the contrary, to ensure fair and contestable downstream markets, it is essential to maintain a diverse landscape of FMs. Independent firms and consumers should have the ability to choose freely and switch between different FMs without being locked into a single provider or ecosystem, regardless of the OS, device (or other key access points and distribution channels for FM deployment, like for example, productivity software).[52]

3.2. GenAI agents

Recent developments in AI systems have shown a growing trend toward the integration of FMs within mobile ecosystems, encompassing operating systems, applications, and digital content layers. In parallel, a proliferation of native GenAI applications has emerged to enhance application functionalities, streamline user interfaces, and enable more adaptive and predictive interactions.

These technological advances are also driving a shift toward the development of autonomous AI systems, in which genAI applications can execute variety of multi-step tasks and decision-

[52] See CMA (2024) AI Foundation Models: Technical update report.

making sequences with minimal human supervision. Such systems are often referred to as “AI agentic systems” or simply “AI agents”. [53]

GenAI agents are increasingly used in diverse applications, ranging from customer service and personalized recommendations to complex problem-solving and creative content generation. These AI agent systems could be applied both for working situations and in consumers’ environment,[54] and are expected to have incredible impact on how end-users act and interact with digital application services and digital devices. Indeed, the ultimate objective is to streamline work and enhance productivity, by reducing the human activity yet, at the same time, maintaining (or enhancing) end-users satisfaction level. This is done by reducing transaction and search costs, ultimately addressing the end-users’ “bounded rationality”[55] and establishing efficient data feedback loops, thus increasing the ability of AI agents to profoundly and quickly understand end-users’ preferences.

Just to give some examples, Anthropic introduced an AI agent capable of controlling user's browsers, executing “clicks”, and inputting text to automate a range of online tasks. Google developed a prototype for a "universal AI assistant", designed to seamlessly operate across multiple devices, such as smartphones and smart glasses, offering users real-time support in their daily activities. As an example of evolution of voice assistants, Amazon launched Alexa+ that, based on a variety of FMs with agentic capabilities, automatically connects services and devices at scale by surfing the web and executing tasks (not limiting its scope of action to companies having a ready-built set of externalized APIs). Deutsche Telekom is working on an app-less phone designed to replace traditional apps with an intelligent

[53] see OpenAI (2023) Practices for Governing Agentic AI Systems; see Microsoft (2024) Agent AI - Microsoft Research: Overview; see Department for Science, Innovation and Technology (2024), A pro-innovation approach to AI regulation: government response.

[54] Likewise, in most of digital markets and services the traditional distinction between consumers and businesses is blurred: both are considered as end-users.

[55] This is the inability to have access to the relevant information, and to assess even their own (intertemporal) preferences to maximise their utility. Consumers have a limited ability to process information, and this often implies that the information disclosed (even under legal obligations) does not really allow consumers to take a well-informed decision. Even, too much information (so-called ‘information overload’) may be confusing, not allowing consumers to clearly identify and select the relevant aspects. Furthermore, consumers often do not factor in all the costs that a certain action implies, causing so-called “internalities” (namely, externalities that people impose on themselves). See H. Allcott, C. Sunstein (2015) Regulating internalities, in Journal of Policy Analysis and Management, Volume34, Issue3.

assistant that receive (verbal) users request and executes tasks dynamically in the background.[56]

From a competition law and economic perspective, genAI agents have the potential to act as highly disruptive technologies. By significantly reducing information asymmetries and compensating for bounded rationality, these systems represent strong and effective consumer empowering mechanisms. In doing so, genAI agents may also lower or eliminate various forms of transaction costs - such as search, switching and multihoming costs - which have constrained consumer mobility and market contestability in digital environments. Their deployment could thus foster greater inter-platform competition, even within markets currently characterized by high concentration and dominant incumbents. [57] Over time, such dynamics may contribute to the commoditization of specific digital services and functions.

However, this disruptive potential is not unambiguously pro-competitive. When genAI agents are developed and tightly integrated within dominant digital ecosystems, they may instead incredibly reinforce the market power of incumbent platforms, amplifying entry barriers, and potentially foreclosing competitive threats. This potential dual impact of genAI agents - simultaneously enabling disruption and entrenchment, on one side, explains why major technology firms prioritize investment aimed to develop such agentic systems, and, on the other side, create complex regulatory and competitive issues.

A particularly concerning scenario arises when genAI agents are designed to assume high levels of autonomy over consumers' decisions. In such cases, the efficiency gains achieved - especially reducing transaction and search costs - may be accompanied by a decline in consumer "sovereignty". [58] Specifically, when users delegate significant decision-making authority to AI agents, either explicitly or implicitly, a "choice

[56] It's unclear whether Deutsche Telekom plans to launch a specific hardware device/OS or integrate this technology into existing Android/iOS ecosystems.

[57] Furthermore, depending on the scale of the genAI agent, new network effects could be created by reducing coordination costs among a plethora of end-users for "collective bargaining" and "collecting switching" with/from different digital and non-digital services.

[58] Consumer sovereignty refers to the economic principle that consumers, through their preferences and purchasing decisions, exert influence over the allocation of resources and the types of goods and services produced. It rests on the assumption that individuals are the best judges of their own welfare. In competitive markets—where no firm possesses significant market power—consumer choice disciplines firm behavior by rewarding providers that offer superior price or quality and sanctioning underperforming ones through reduced demand. See J. Persky, (1993) Retrospectives: Consumer Sovereignty, in Journal of Economic Perspectives 7, no. 1 (1993): 183–191; W. Fellner, C. Spash (2015) The Role of Consumer Sovereignty in Sustaining the Market Economy, in Handbook on the Politics and Governance of Sustainable Development (Edited by L. Reisch, J. Thøgersendward).

gap" [59] emerges: the effective autonomy of users in market interactions is diminished, with decision pathways shaped or even pre-determined by the agent.

The legal and economic implications of this principal-agent relationship[60] depend on several factors, including: the degree and scope of delegation; the richness and exclusivity of shared users' data; and the extent to which users can revise, monitor, or override agent-driven decisions. [61] Depending on the design of this relationship, various forms of agency costs may arise, including misaligned incentives, information asymmetries, and moral hazard. These costs may result in a net disempowerment of users, counteracting the initial promise of genAI systems to enhance consumer freedom of choice.

Beyond the direct consumer impacts, there are also significant implications for market contestability. Delegation to genAI agents may result in user entrenchment within "digital aftermarkets" [62] - i.e., secondary environments where users retain some degree of residual choice but within parameters effectively set by the agent. This situation would eventually constraints users' ability to switch provider due to: (i) information lock-in, as agent has exclusive access to users' behaviours and preferences; (ii) network effects, reinforced by agent integration within dominant ecosystems; (iii) self-preferencing, wherein agents prioritize services or products within its ecosystem.

It is important to underline that the creation of aftermarkets is typical of most gatekeeping positions in digital markets,

[59] See N. Shchory, M. Gal (2022) Voice Shoppers: From Information Gaps to Choice Gaps in Consumer Markets, 88 Brook. L. Rev. 111, who descibe a novel form of market failure that arises from consumers voluntarily delegating product selection to AI-powered voice assistants like Alexa or Google Assistant. The authors argue that traditional consumer protection and antitrust frameworks are insufficient to address the choice gap. Instead, they propose applying agency law, treating the voice shopper as an agent of the consumer and imposing fiduciary and informational duties to better align voice-based decisions with users' best interests.

[60] The principal-agent relationship refers to a situation in which a person or entity (the agent) undertakes actions on behalf of another person or entity (the principal) by whom he is delegated. The agency problem arises when there is a mismatch of interests and information resulting in agency costs for the principal. Moreover, given the diversity of objective functions, the agent may perform 'hidden actions' to pursue its own interests to the detriment of those of the principal (moral hazard). The agency problem is aggravated when the principal does not have optimal means to control the actions of the agent and sanction the agent in case of "misbehaviour". For this reason, the principal has an interest in devising an appropriate incentive system to induce the agent to behave in a manner that is consistent with the principal's goals and interests.

[61] See N. Kolt (2025) Governing AI Agents, Notre Dame Law Review, Vol. 101.

[62] In economic theory, an "aftermarket" is a secondary market where consumers make their choices after having purchased a primary product or service, with an ex-post freedom of choice that is constrained by previous choices in the primary market (typical examples are complementary products and services, like maintenance, or consumable goods related to durable ones, like printer ink cartridges).

however, AI functionalities would increase the end-users' informational captures. Notably, the degree of user lock-in is proportional to both the intensity (extent of autonomy transferred to the agent) and scope (range of services under agent control) of the initial delegation. As for the latter, the higher the genAI agent is positioned within the digital value chain - for instance, at the operating system layer - the more expansive its control across a wide array of digital services (figure 3).

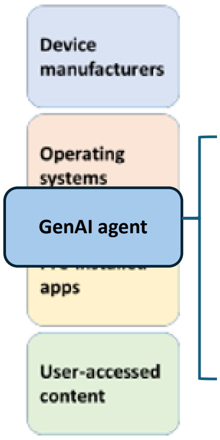

Fig. 3 – GenAI agent in the mobile ecosystem value chain

As a result, the generative AI agent itself may evolve into a multi-sided platform, mediating interactions not only between end-users and business-users, but also across entire ecosystems of business actors. In doing so, the agent may facilitate not merely horizontal expansion - by increasing the number of intermediated business users - but also incorporating additional intermediaries within the digital value chain. These intermediaries may include entities currently serving as key access points through which business users reach end-users - entities that, under the Digital Markets Act (DMA), may qualify as gatekeepers.

In this scenario, the rise of AI agents may not simply substitute one gatekeeper with another, but it could result in the emergence of a "meta-gatekeeper": an overarching, highly integrated intermediary able to orchestrate and govern multiple ecosystems and able to exert greater systemic influence than any individual platform, effectively concentrating decision-making power and reshaping the structure of intermediation.

To develop AI agents on mobile devices in a manner responsive to end-user, genAI applications providers must ensure tight interoperability with the underlying OS - enabling interaction with browsers, search engines, and a wide range of other digital services and device-level functionalities enabling a broad interaction with digital world.

In the medium term, AI agents may not only complement existing OSs but substitute them or both may evolve into a configuration that combine their functions into a single, integrated interface. In such a scenario, the agent could effectively and directly become the operative environment through which users interact with the digital world. Moreover, the role of the OS would shift - controlling not only access to hardware and core services but also shaping the pathways through which business users engage with end-users across a broad range of services.

If AI agents were to overlap or merge with OS-level functions, from the user's perspective, each platform ecosystem could become "the market" itself. This would incredibly increase lock-ins, both informational and functional, as well as raise substantial challenges for existing regulatory frameworks, including the Digital Markets Act and traditional competition law.

4. Conclusion

The rapid evolution of FM and GenAI and their integration within the mobile ecosystem present both opportunities and competitive concerns. To assess the latter, this paper has explored the structural complexities of the GenAI value chain, distinguishing between upstream and downstream dynamics and analysing the differentiated potential competitive risks. This analysis facilitates the identification of market contexts where tipping dynamics, consumer lock-in, and foreclosure risks are most likely to arise, and where the ex-post competition law tool could be less effective in addressing these emerging competition concerns, consequently necessitating of an ex-ante regulatory intervention.

At the upstream level, large technology firms have significant market power, particularly for cloud computing, AI chips, and proprietary datasets. While these firms have not achieved entrenched dominance in those segments, their control over key inputs creates a risk of leveraging strategies that could limit competition and innovation. However, both conglomerate competition dynamics and emerging technological trends seem to be able to counterbalance these risks and foster competition. Moreover, competition law tools seem to be able to prevent and, in case, remedy abusive exertion of upstream market power.

At the downstream level, the competitive risks seem more pronounced. The integration of FM into OS within mobile ecosystems, dominated by Google's Android and Apple's iOS, raises concerns about self-preferencing and restricted access for competing AI models and genAI applications. This vertical integration could replicate the gatekeeping effects observed in traditional digital markets, reinforcing lock-ins for consumers and reducing the ability of third-party developers to compete on fair terms. Moreover, the emergence of GenAI agents could further strengthen incumbents' control over consumer interactions while simultaneously introducing overarching layers of intermediation and ultimately building a multilayered ecosystem.

Given these dynamics, regulatory intervention seems necessary to ensure market fairness and contestability. From these perspectives, it is essential to ensure that these emerging AI-driven ecosystems - particularly the evolving relationship between OSs and AI agents - remain as open and interoperable as possible. The governance of such system should be guided by principles already embedded in the DMA.

In particular, ex-ante regulation should focus and address the following:

(i) consumer empowerment through "agent neutrality", i.e., end-users must retain the ability to make a genuine and informed choice of AI agent provider, irrespective of the OS environment in which they operate.[63] This principle reflects existing DMA obligations imposed on gatekeepers with respect to the freedom of choice for downstream applications and services;

(ii) data portability and access, i.e., users should have the right to transfer their data seamlessly between AI agents or authorize third-party access to their personal data, should they choose to switch providers. This is necessary to avoid lock-in effects that would otherwise undermine user autonomy and dynamic competition.

(iii) vertical interoperability across ecosystem layers[64], i.e., regulation must recognize that AI agents are not merely ancillary services embedded in OSs but may function as new intermediaries - positioned between end-users and a broad array of business users, including other platform intermediaries. Ensuring that AI agents can interoperate vertically - across

[63] F. Bostoen, J. Krämer (2024) AI Agents and Ecosystems contestability – CERRE Report

[64] Vertical interoperability promotes complementary innovation and the modular combination of services across the value chain allowing complementors to access the ecosystem and compete for end users by exchanging data and functionalities via application programming interfaces (APIs). See M. Bourreau, J. Krämer, M. Buiten 2022, Interoperability in digital markets – CERRE Report; G. Colangelo, A. Ribera Martinez (2025) Vertical interoperability in mobile ecosystems: Will the DMA deliver (what competition law could not)?, in International Review of Law and Economics Volume 83.

different layers of the mobile and digital ecosystem - will be critical for preserving multi-homing, and for limiting the emergence of meta-gatekeepers with the power to orchestrate vertically nested market positions.

The DMA provides a legal framework for addressing some of these competitive concerns, particularly by preventing exclusionary practices and ensuring interoperability between FM services and application and mobile OS ecosystem. However, an in-depth analysis is required to understand whether and how DMA is already applicable to the complex market situations where GenAI agents become more integrated into OS and the overall mobile ecosystems.