FRIEDRICH NAUMANN FOUNDATION For Freedom.
Global Innovation Hub

Europe's Path to Competitiveness in the Global AI Race

Anselm Küsters
Pauline Weil
Christophe Carugati

ANALYSIS

Imprint

Publisher

Global Innovation Hub
Friedrich Naumann Foundation for Freedom
15F.-6, No. 171, Songde Road,
Xinyi District, Taipei City 110030
Taiwan

/freiheit.org/taiwan
/FNFGIHUB
/FNFGIHUB
/FNFGIHUB

Author

Anselm Küsters
Pauline Weil
Christophe Carugati

Editor

Global Innovation Hub of the Friedrich Naumann Foundation for Freedom

Contact

E-Mail: globalinnovation@freiheit.org

As of

November 2024

Notes on using this publication

License

Disclaimer

Table of Contents

1. Introduction: The Global AI race – where does Europe stand?

Dr. Anselm Küsters
Centre for European Policy (cep), Berlin
Céline Nauer
Friedrich Naumann Foundation for Freedom Global Innovation Hub

Artificial Intelligence (AI) has long moved beyond science fiction – it is already here, transforming economies, societies, and even global power dynamics in real time. The development of Large Language Models (LLMs) has driven this change, allowing machines to take on tasks that require a level of language understanding once thought to be exclusive to humans. **This transformation has sparked a global race to develop frontier AI models, which in turn has intensified geopolitical rivalries.** As nations vie for AI dominance, they are increasingly implementing legal hurdles, sanctions, or even export controls to protect their technological advances and prevent rivals from gaining access to critical AI technologies. For example, the United States (US) has implemented export controls aimed at restricting China's access to advanced AI and semiconductor technologies. The economic, political, and security significance of generative AI today can hardly be overstated – it has the potential to drive economic growth and solve complex problems across a range of academic disciplines, but it also poses challenges in terms of market concentration, and international relations. As a result, countries are competing for technological supremacy and economic advantage, leading to new tensions and strategic manoeuvres on the international stage.

At the moment, however, Europe seems to be falling behind in this race for AI leadership. Despite having strong academic institutions, a robust regulatory system, and a commitment to ethical AI, the European Union (EU) and its Member States are struggling to keep up with the speed and scale of AI development in the US and China. The US is leveraging its home-grown tech giants such as Google, Microsoft, and OpenAI, which have been at the forefront of many AI breakthroughs and continue to push the boundaries of what is possible. Meanwhile, China's AI surge is being driven by huge government investment, vast amounts of data from its large population, and an aggressive plan for technological independence. Among experts, Chinese AI labs are known for achieving impressive results through the efficient use of computing power. In an era of rising geopolitical tensions and fragmented trade, sensitive AI-related data and strategic assets will be increasingly exposed to external influences and regulatory frameworks. To remain competitive and uphold its democratic and liberal values, Europe needs to understand and adapt to the approaches of these AI superpowers. The question is: **What steps must Europe take now to ensure that its AI industry remains competitive and independent in the face of the dominance of a few external players? With this series of publications, we aim to contribute to the rationalisation of the debate and present liberal solutions to foster an informed and constructive discussion.**

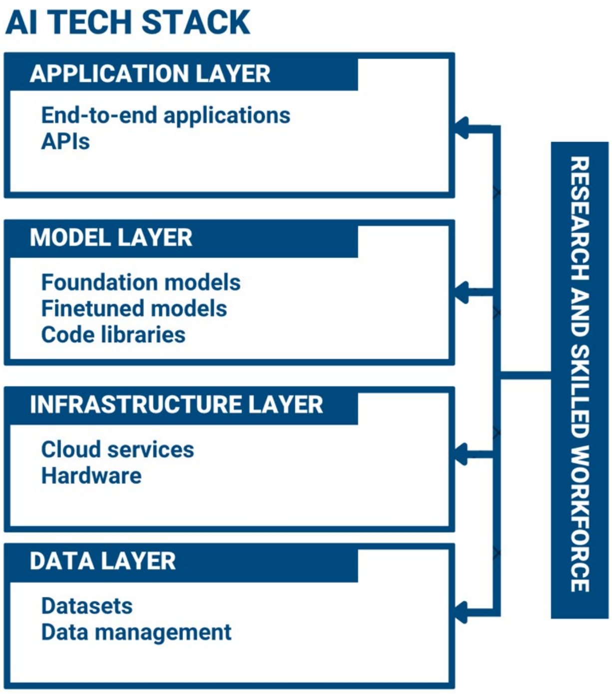

Figure 1: The AI tech stack visualized

To identify where Europe can strengthen its position, it is helpful to visualise the typical value chain behind state-of-the-art generative AI applications (Figure). The value chain starts with data acquisition and management, as high-quality and diverse datasets are essential for training effective AI models. Increasingly, synthetic data generated by AI or proprietary datasets that can be licensed play a key role. At this initial stage, the focus is on the **computing infrastructure** that provides the necessary processing power for this data to train frontier AI models. This requires specialised hardware such as certain chips (called graphics processing units, or GPUs, and tensor processing units, or TPUs) or access to cloud services and high-performance computing facilities. The next critical step in the value chain is **model development**, that is the refinement of algorithms that enable machines to learn from data. It covers both foundation models – broad, general-purpose models – and fine-tuned models that are specialised for specific tasks. Here, significant expertise is required in areas such as natural language processing (NLP) and software engineering, e.g. for the handling of code libraries. Once the models are trained, the next step is **deployment and integration**, which involves embedding these AI models into practical, end-to-end applications. For example, Application Programming Interfaces (APIs) allow companies to integrate AI into their systems. Finally, it is important to keep in mind that **research plays a crucial role at every level of this value chain.** At the infrastructure layer, research is driving advances in computing technologies and optimising cloud services. In the model layer, the development of new algorithms and techniques enables the construction of more efficient and accurate AI models. Finally, in the application layer, research is essential to improve scalability and ensure compliance with emerging regulatory standards.

Europe's challenge is to engage effectively at each stage of this value chain. While it excels in data protection, certain areas of research and ethical considerations, it often lags behind in hardware capabilities and suffers from market and regulatory fragmentation. These gaps lead to dependencies on external actors and hinder Europe's ability to fully realise the benefits of generative AI in geopolitically turbulent times. The incoming European Commission, led by Ursula von der Leyen for the 2024-2029 term, has recognised these shortcomings and plans to strengthen Europe's position in AI through initiatives such as the creation of AI Factories and a European AI Research Council, which aim to pool resources and foster collaboration across Member States – but more needs to be done. In this context, this **set of policy briefs analyses Europe's position in the generative AI value chain and provides recommendations in each area** that align with a liberal market-based order and leverage the continent's unique strengths.

The first **paper by Dr Anselm Küsters (Centre for European Policy) focuses on the algorithm and model development stage of the value chain**, assessing the key drivers of generative AI and evaluating Europe's capabilities in terms of model inputs. By analysing a wide range of data sources and reviewing recent literature on AI, the paper examines model development trends, academic and industrial output, and policy frameworks in Europe, the US, and China. It argues that AI development will increasingly be hardware-enabled but software-defined. Europe should therefore leverage its unique advantages in software inputs – such as strong academic research, a skilled workforce, and a culture of open source collaboration – rather than engage in costly subsidy races focused on hardware such as the most advanced chips and data centres. The paper concludes with recommendations for more targeted and better coordinated investments in AI research, strengthening decentralised innovation through fair competition, and fostering a skilled workforce to support an open source ecosystem.

The second **paper by Pauline Weil (independent economist, formerly Bruegel) looks at the computing infrastructure stage**, evaluating Europe's approach to hardware dependencies and infrastructure needs. By examining the global distribution of AI firms and high-end computing resources, it highlights Europe's limited foothold compared to the US and China. As AI firms increasingly rely on cloud computing services dominated by a few US tech giants, the paper argues that the EU should aim to attract investment for more computing resources on European soil, with a priority on ensuring access to high-end computing for European public and private actors to develop AI models and applications. In particular, the EU should focus on creating an ecosystem that fosters investment, optimises resources, and integrates AI into a broader industrial policy. Recommendations for policymakers include making Europe attractive for computing investments by focusing on resilient energy infrastructure, integrating AI into a comprehensive industrial strategy, and pushing for international cooperation to ensure a global level playing field and agreements on balancing AI risks.

The final **paper by Dr Christophe Carugati (Digital Competition) examines the market and competitive dynamics phase**, focusing on the rapidly evolving regulatory landscape in generative AI and the uncertainties it creates for developers at both the development and deployment stages. It highlights the importance of competition authorities in Europe closely monitoring the generative AI sector to keep pace with its rapid changes. The paper advocates cooperation between regulators to ensure a consistent approach, suggesting that authorities coordinate their efforts in frameworks such as the Digital Markets Act High Level Group. It recommends adopting a co-regulatory approach by engaging with industry stakeholders to better understand ongoing developments and address novel and complex cross-regulatory issues. By providing guidance to companies on their practices prior to implementation, authorities can foster a more supportive and innovative regulatory environment and maintain vibrant competition in Europe's generative AI market.

Overall, the intersection of AI technology development, hardware infrastructure, and market competition, as described in these three policy briefs, will define Europe's future position in the global AI landscape. Our series of papers highlights that **Europe's key opportunities lie in leveraging its strengths in software and research, making strategic but targeted investments in critical infrastructure, and fostering a fair market environment**. At the same time, these measures will also help Europe reduce dependencies and exert greater control over its AI trajectory. However, the global nature of AI means that **complete sovereignty is likely not achievable – or even desirable**. Maintaining open trade relationships with like-minded countries such as the US and ensuring vibrant competition through close monitoring and cooperation between authorities will remain crucial to ensure that AI market remains open to innovation and alternative providers. Through coordinated efforts and a holistic strategy, Europe can increase its autonomy and competitiveness in the dawning era of generative AI.

2. How to strengthen the EU's position in the AI industry: The importance of software and research

Dr. Anselm Küsters
Centre for European Policy (cep), Berlin

This policy brief examines Europe's strategic positioning in generative AI research, in comparison to the United States and China. Using a range of data sources and reviewing the literature on language models, it analyses model development, academic and industrial output, and policy frameworks. The findings reveal significant disparities between the EU and its two main competitors in the race for advanced AI models, particularly in terms of computing power and human resources. Europe should focus on software inputs to increase computational and algorithmic efficiency rather than entering a costly subsidy race focused on hardware inputs such as frontier chips and data centers.

Introduction

This policy brief is part of a larger study analyzing Europe's reliance on other countries regarding Artificial Intelligence (AI) technologies compared to the United States (US) and China, with a focus on research into generative AI models. Generative AI, which includes advanced machine learning (ML), natural language processing and image generation technologies, will be key for innovation and welfare in the next decade. However, the future availability of these technologies is constrained by Europe's reliance on external hardware and software inputs and the vulnerabilities inherent in global technology supply chains. This study provides recommendations for European policymakers to secure AI sovereignty without over-stretching public finances.

The current geopolitical landscape, characterized by competition between the US and China and the lingering effects of the COVID-19 pandemic, highlights technological dependencies. The US is increasingly viewing AI through a national security lens, leading to trade restrictions. Notably, the US Department of Commerce's revision of semiconductor export controls in October 2022 and President Biden's executive order in August 2023 have constrained China's semiconductor industry but have also led China to increase domestic investment in advanced chips (Shivakumar et al., 2024). In May 2024, US lawmakers proposed significant funding for domestic AI research and development, signaling a commitment to maintaining global technology leadership and addressing national security concerns related to AI advances.

This homeshoring trend is also evident in Europe, where the EU has initiated a "geoeconomic turn". On 3 October 2023, the European Commission adopted a recommendation on critical technology areas for the EU's economic security, identifying advanced semiconductors and AI as key areas. Despite several initiatives aimed at boosting AI investment and adoption, there continues to be a large AI investment gap between the US and the EU (Maslej et al., 2024). The new Commission's plan to create European "AI factories" with access to supercomputing capacity, to develop an "Apply AI Strategy" to stimulate new applications, and to establish a European AI Research Council to pool resources (von der Leyen, 2024), depend on a diversified and secure supply of AI components.

In this context, this policy brief assesses the key drivers of generative AI development and evaluates capabilities related to model inputs. By drawing on a wide range of data sources and reviews of the current literature on large language models (LLMs), it analyzes model development, academic and industrial output, and policy frameworks for Europe, the US, and China. The core argument is that AI development will increasingly be hardware-enabled but software-defined, with software absorbing much of the growing sophistication of generative AI models. Europe should therefore leverage its unique advantages in software inputs, rather than engaging in a costly subsidy race focused on hardware inputs such as chips and data centers.

Comparative quantitative analysis of AI development, innovation, and policy

To understand the disparities and dynamics of AI devel-opment among global leaders, this section quantitatively compares the US, China, and Europe in terms of AI model capabilities, innovation outputs, and policy initiatives. Based on data from Epoch AI, which tracks over 800 his-torically significant AI models (Epoch AI, 2024), Figure 1 shows a sharp increase in the number of frontier AI models in the US, especially in the last decade, outpacing Europe and China. This indicates a strong and growing AI development ecosystem in the US, driven by both aca-demic and industrial efforts. However, the increase in the number of AI models in China in recent years suggests an emerging, competitive AI ecosystem, while Europe has maintained a modest pace. To better capture the leading capabilities in AI development, the other panels of Figure 1 show the top 10 models per region based on three key metrics: the number of parameters, the number of training calculations, and the size of the training dataset. Among the top 10 models included in the Epoch AI dataset, the US and China dominate with models that have a higher number of parameters compared to Europe. Large mod-els, measured in billions of parameters, reflect greater computational investment and, by storing more informa-tion, greater ability to handle complex AI tasks. US models tend to have the highest number of parameters, followed closely by Chinese models. Europe, however, lags behind in terms of size. Computing requirements (Floating Point Operations, short: FLOP) for the top models in the US are consistently higher than those in China and Europe, which may indicate a lack of advanced chips or a reliance on US cloud providers. FLOP is a critical metric that reflects the computational intensity involved in training these mod-els, which is why both the White House Executive Orders on AI Safety and the EU AI Act codify the use of FLOP to identify more powerful systems (Hooker, 2024). Dataset size is another critical factor for training large AI mod-els, with the US and China again leading in terms of data volume due to extensive data collection efforts. Europe may be constrained by data protection regulations (such as GDPR), which limit the types of data that can be used for AI development. All of the metrics depicted in Figure 1 drive the so-called scaling laws behind the current AI research trend in deep learning (Kaplan et al., 2020). Scaling laws refer to empirical relationships that show how AI model performance improves predictably as model size, computational resources, and data volume increase. As these factors are scaled up, models generally exhibit improved capabilities *and* accuracy.

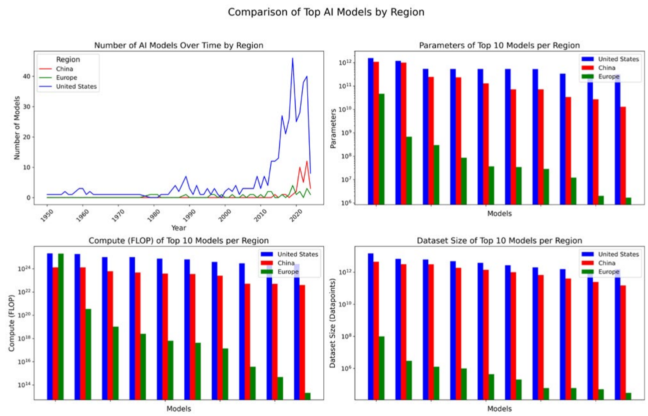

Figure 1. Notes: Data source: Epoch AI dataset, which tracks historically significant AI models. Panel 1: Number of frontier AI models developed over time in the US, China, and Europe. Panels 2-4: Top 10 AI models per region based on number of parameters, training computations (FLOPs), and training dataset size. Note that the log scale allows for a clear comparison of the top 10 mod-els per region across metrics.

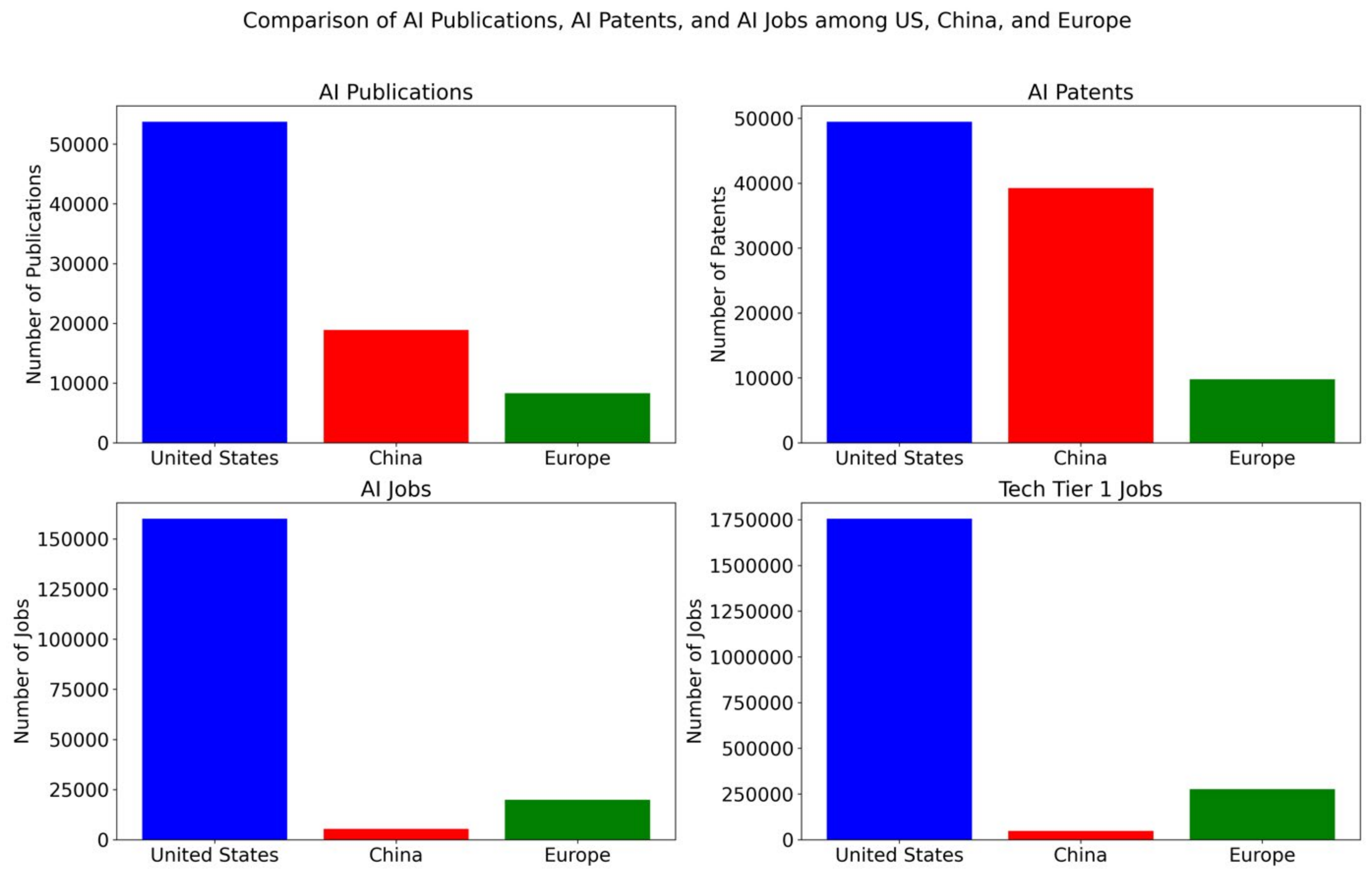

Figure 2. Notes: Data Source: Emerging Technology Observatory's Private Sector AI-Related Activity Tracker (PARAT). The axes indicate the quantity for each metric (e.g., number of publications). Note that the figure focuses on private-sector output from leading firms.

To better understand the drivers behind these divergent model capabilities, one can turn to the underlying academic and industrial inputs for AI model development. Based on data from the Emerging Technology Observatory's Private-sector AI-Related Activity Tracker (PARAT), Figure 2 shows the number of AI publications, patents, and jobs in the US, China, and Europe for hundreds of leading companies around the world.[1] The US leads in both AI publications and patents among these leading companies, reflecting its strong academic-industry collaboration and robust innovation ecosystem. China follows closely in patents but lags in publications within the private sector, suggesting a public sector driven strategy with focus on commercialization and applied research, not least in the military domain. Europe, in comparison, ranks lower in both publications and patents, evidencing weaknesses in academic-industry linkages. The US clearly dominates the AI labor market, highlighting its ability to attract and retain a skilled AI workforce. According to alternative data from the Center for Security and Emerging Technology (Maslej et al., 2024), which covers whole geographical areas and not just the top companies per region, China leads the world in AI patents (although by a much smaller margin if only granted patents are considered). However, focusing on the concrete resources of leading technology companies is a more helpful lens for strategic analysis, as this is where the next generation of state-of-the-art LLMs will be developed. It is therefore important to note that while recent data-such as from SCOPUS-suggest that China may now lead in overall AI publications, the analysis in Figure 2 focuses specifically on private sector output from leading firms. Publications and patents associated with technology companies are likely to be of higher quality and have immediate practical applicability, whereas general "AI publications" may include non-peer-reviewed work. Moreover, the strategic resources that drive the development of frontier AI models are more likely to come from the private sector, as the example of OpenAI shows, so the difference in top jobs in these firms does not bode well for the EU's aspiration to reach AI sovereignty.

1 See: https://eto.tech/tool-docs/parat/, last accessed: 24.09.2024.

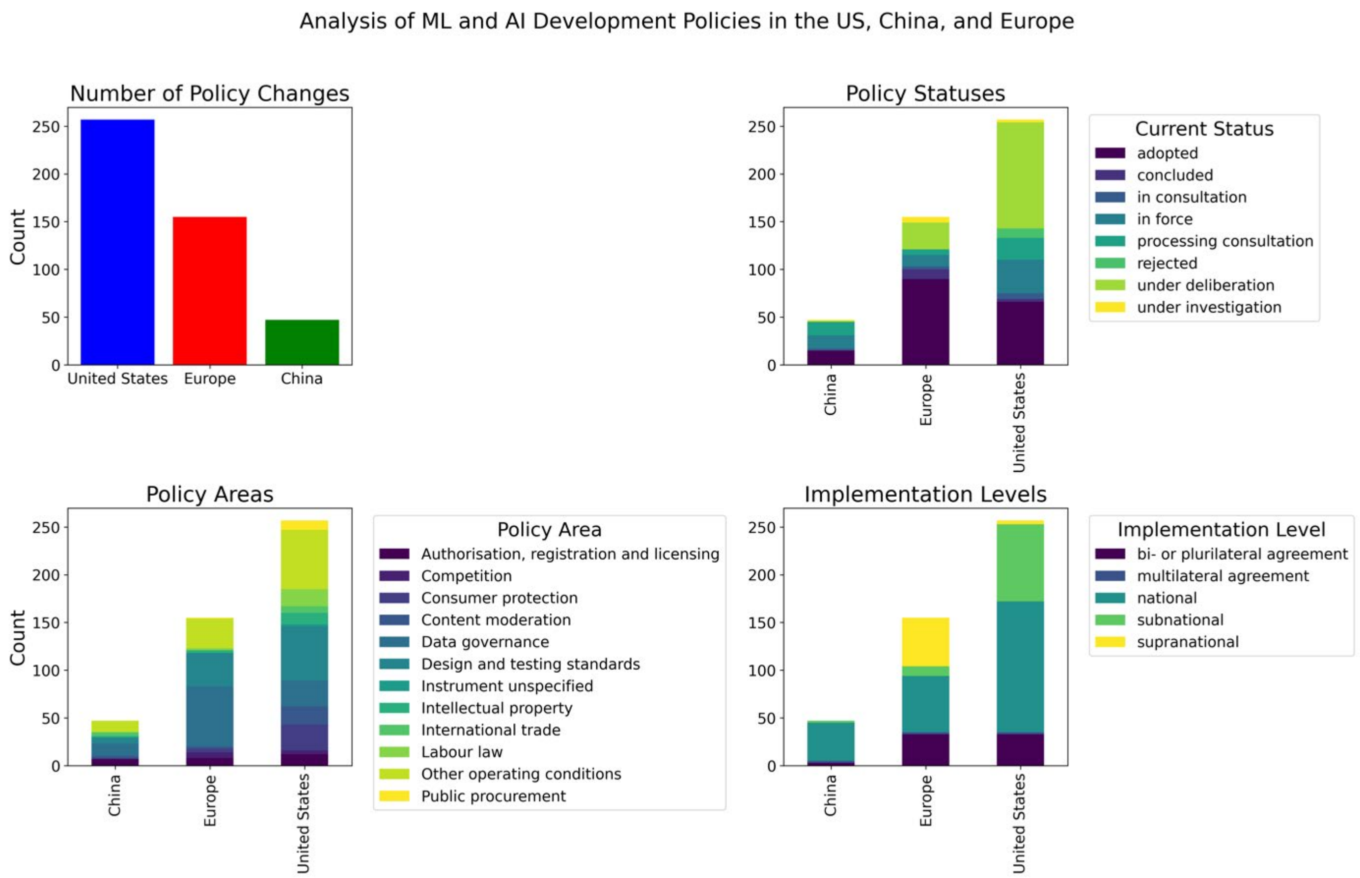

Figure 3. Notes: Data source: DPA Activity Tracker. The data is filtered for all policies related to the category "ML and AI development". According to the DPA Handbook, a specific policy is classified under "ML and AI development" if it directly affects companies specializing in ML and AI, such as OpenAI or Anthropic. Number of policy changes: Total number of AI-related policy initiatives proposed, adopted, or rejected. Policy Status: Classification of policies by their current status (e.g., proposed, adopted). Policy Areas: Specific areas of focus, such as competition or content moderation. Implementation Levels: Whether policies are implemented at the national, regional, or local level.

Figure 3 tracks the number of AI-related policy changes, policy status, policy areas, and implementation levels in China, Europe, and the US based on aggregating data from the DPA Activity Tracker, which monitors developments in more than 50 jurisdictions worldwide.[2] The data is filtered for all policies related to AI development and differentiated by status, policy area, and implementation level. The US leads in the number of policies proposed, but many of these initiatives in areas such as licensing, competition, consumer protection, and intellectual property were ultimately rejected. While China has seen fewer policy changes than the US, there has been significant activity to ensure the security of the Communist Party regime in the face of rapid technological advances. Europe stands out as having adopted the highest number of policies, often focusing on heavy-handed regulatory measures such as data governance or content moderation.

2 See: https://digitalpolicyalert.org/activity-tracker?offset=0&limit=10&period=2020-01-01,2024-07-23, last accessed: 24.09.2024.

Current bottleneck and strategic opportunities

The rapid progress of AI has been driven by the LLM "scal-ing law" (Kaplan et al., 2020; Villalobos et al., 2022). As noted above, this correlation suggests that performance improvements are proportional to increases in model size, data volume, and computing power. As a result, key inputs such as advanced chips, rich data, and innovative algo-rithms have become strategically important. What are the current bottlenecks and strategic opportunities for Europe in these areas?

Lack of computational resources

To begin with, the exponential growth in AI has led to a critical shortage of Graphics processing units (GPUs), which are, at least for now, essential to large-scale AI training because of their ability to perform complex math-ematical computations in parallel and thus efficiently. In this way, GPUs accelerate the processing of the big data required to train deep learning models. This shortage is exacerbated by the concentration of AI training in large data centers, high-end GPUs that are sometimes idle, and fragmented pricing structures.[3] While there are several cloud service providers, the market is oligopolistic and dominated by non-European companies. Promoting the decentralization of computing resources and fostering collaboration to optimize the use of GPUs is therefore a strategic imperative. Distributed training can cost-effi-ciently train LLMs with up to 1.3 billion parameters (Sani et al., 2024). Similarly, there are open training methods for LLMs that allow model training to be distributed across multiple, remote computer clusters in different countries (Jaghouar et al., 2024). Accordingly, the EU should sup-port research in federated and decentralized learning to leverage its distributed computing power, such as the Eu-roHPC supercomputers.

Data availability constraints

The availability of public human data is a further potential bottleneck for the growth of LLMs in Europe. Projections suggest that LLMs could exhaust the stock of public human text data between 2026 and 2032 (Villalobos et al., 2022). Moreover, web domain consent protocols are becoming increasingly restrictive, affecting the diversity of new AI training data and making it more difficult for Europeans to catch-up with leading American and Chi-nese companies, which probably backed up all this data before access became more restricted (Longpre et al., 2024). Finally, the strict GDPR has led to EU firms decreasing data storage by 26% and data processing by 15% relative to comparable US firms and thereby becoming less "data-intensive" (Demirer et al., 2024). To mitigate this bottleneck, EU firms can turn to large-scale synthetic data creation (Chan et al., 2024). Research shows that AI models trained on heterogeneous synthetic datasets outperform those trained on limited or homogeneous real-world datasets (Zhang et al., 2024). However, although synthetic data can augment training resources, an over-reliance on it can compromise model quality (i.e. the quality of the model's output), a phenomenon known as "model collapse" (Shumailov et al., 2024). This occurs when models produce outputs that are too uniform or lack diversity because they have not learned from sufficiently diverse real-world data. Moderate use of synthetic data is beneficial, and new techniques are emerging to safely integrate more synthetic data without sacrificing model performance (Wu et al., 2024).

Importance of algorithmic advances

Often forgotten, algorithmic advances are as crucial to the development of AI as computing power and depend not only on experimental computing but also on research effort (Aschenbrenner, 2024, p. 22ff.). For instance, the advances seen in GPT-2, the predecessor to OpenAI's flagship model "ChatGPT", were not only the result of increased computational power, but also of significant algorithmic modifications (Minaee et al., 2024, p. 6). Innovations such as DeepMind's JEST (Evans et al., 2024) and FlashAttention-3 (Shah et al., 2024) demonstrate improvements in computational efficiency and model performance with reduced computing power. Accordingly, current EU policies, such as the EU Chips Act, likely focus too much on expensive semiconductors. In short, investing in algorithmic efficiency can accelerate AI progress in Europe by optimizing domestic resource use.

Exploration of alternative AI architectures

Similarly, transformative advances in AI architecture, such as test-time training (Sun et al., 2024) and state-space models (Wiggers, 2024), offer promising alternatives to traditional transformers, which stand behind the current paradigm in deep learning as performed by OpenAI and others. In plain English, these architectures improve computational efficiency and can handle large datasets more effectively. Similarly, so-called small language models

3 See: https://www.primeintellect.ai/blog/compute, last accessed: 24.09.2024.

offer advantages in terms of cost, accessibility, and operational versatility. With algorithmic advances and partially synthetic data, it is now possible to achieve nearly the performance of a 2022 model, which at the time cost over $100,000, for around $2,000 (Sehwag et al., 2024).

Optimize existing AI infrastructure

Finally, significant progress can also be made by optimizing existing AI infrastructure. For instance, grounding LLMs with task-specific tools such as debuggers and scripting environments improves their ability to perform vulnerability research.[4] Likewise, advanced reasoning problems can be solved by combining a language model with a symbolic engine, which uses formal languages such as logic to represent knowledge.[5] By enhancing existing technologies, the EU can thus achieve AI gains without relying solely on cutting-edge developments. EU-supported research could follow the program outlined by AI researcher Gary Marcus, who has long advocated using more hybrid neuro-symbolic architectures (Marcus, 2020, p. 53).

Conclusion

Based on the quantitative and qualitative results, this policy brief concludes by outlining three core recommendations for the EU and its member states:

Greater, more targeted, and better coordinated investment in AI:

Despite the comprehensive AI plans for 2018 and 2021, the framework for coordinating and regulating AI investments remains underdeveloped, as shown by a recent special report by the European Court of Auditors (ECA) (European Court of Auditors, 2024). Similarly, the current EU investment targets are vague and outdated, according to the ECA report, as the Commission has not significantly increased private co-financing or ensured the commercialization of EU-funded AI research results. Based on the above analysis of the drivers of current LLM development, the next wave of financial and infrastructural support should focus on federated learning, small language models, and novel AI architectures.

Strengthening decentralized innovation through fair competition:

Recent proposals for the next Commission, for example from France and Germany, to undermine current competition law and state aid rules by supporting "national champions" and merging them to become even bigger, are misguided. This approach will increase tail-end risks and undermine innovation because it creates further chokepoints and potential bottlenecks. Moreover, from the point of view of the innovation pathways described in this brief, such as decentralized training, concentrated providers are not strictly necessary. Instead, improved access to funding through a capital markets union and targeted exemptions for smaller AI players will facilitate domestic growth.

Foster a skilled workforce and open-source ecosystem:

Leading AI researchers have noted that open-source projects reduce dependence on proprietary technologies from other regions without increasing AI risks (Gray Widder et al., 2023; Kapoor et al., 2024). Rather than aligning Europe's financial commitments to computing infrastructure and data resources with spending levels in the US and China, there should be a greater focus on research, human skills, and open-source projects, as this will help accelerate and apply the software-focused research highlighted in this brief.

4 See: https://googleprojectzero.blogspot.com/2024/06/project-naptime.html, last accessed: 24.09.2024.
5 See: https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/, last accessed: 24.09.2024.

References

- Aschenbrenner, L. (2024). Situational Awareness. The Decade Ahead. https://situational-awareness.ai
- Chan, X., Wang, X., Yu, D., Mi, H., & Yu, D. (2024). Scaling Synthetic Data Creation with 1,000,000,000 Personas (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2406.20094
- Demirer, M., Hernández, D. J. J., Li, D., & Peng, S. (2024). Data, Privacy Laws and Firm Production: Evidence from the GDPR (w32146; p. w32146). National Bureau of Economic Research. https://doi.org/10.3386/w32146
- Epoch AI. (2024). Data on Notable AI Models. https://epochai.org/data/notable-ai-models
- European Court of Auditors. (2024). EU Artificial intelligence ambition. Stronger governance and increased, more focused investment essential going forward (Special Report 08). https://www.eca.europa.eu/ECAPublications/SR-2024-08/SR-2024-08_EN.pdf
- Evans, T., Parthasarathy, N., Merzic, H., & Henaff, O. J. (2024). Data curation via joint example selection further accelerates multimodal learning (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2406.17711
- Gray Widder, D., West, S., & Whittaker, M. (2023). Open (For Business): Big Tech, Concentrated Power, and the Political Economy of Open AI. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4543807
- Hooker, S. (2024). On the Limitations of Compute Thresholds as a Governance Strategy (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.05694
- Jaghouar, S., Ong, J. M., & Hagemann, J. (2024). OpenDiLoCo: An Open-Source Framework for Globally Distributed Low-Communication Training (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.07852
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling Laws for Neural Language Models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2001.08361
- Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., Hopkins, A., Bankston, K., Biderman, S., Bogen, M., Chowdhury, R., Engler, A., Henderson, P., Jernite, Y., Lazar, S., Maffulli, S., Nelson, A., Pineau, J., Skowron, A., … Narayanan, A. (2024). On the Societal Impact of Open Foundation Models (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2403.07918
- Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., Saxena, N., Obeng-Marnu, N., South, T., Hunter, C., Klyman, K., Klamm, C., Schoelkopf, H., Singh, N., Cherep, M., Anis, A., Dinh, A., Chitongo, C., Yin, D., … Pentland, S. (2024). Consent in Crisis: The Rapid Decline of the AI Data Commons (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2407.14933
- Marcus, G. (2020). The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2002.06177
- Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. (2024). The AI Index 2024 Annual Report. AI Index Steering Committee, Institute for Human-Centered AI, Stanford University.
- Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey. arXiv. https://arxiv.org/abs/2402.06196
- Sani, L., Iacob, A., Cao, Z., Marino, B., Gao, Y., Paulik, T., Zhao, W., Shen, W. F., Aleksandrov, P., Qiu, X., & Lane, N. D. (2024). The Future of Large Language Model Pre-training is Federated (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2405.10853
- Sehwag, V., Kong, X., Li, J., Spranger, M., & Lyu, L. (2024). Stretching Each Dollar: Diffusion Training from Scratch on a Micro-Budget (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.15811
- Shah, J., Bikshandi, G., Zhang, Y., Thakkar, V., Ramani, P., & Dao, T. (2024). FlashAttention-3: Fast and Accurate Attention with Asynchrony and Low-precision (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2407.08608
- Shivakumar, S., Wessner, C., & Howell, T. (2024). Balancing the Ledger: Export Controls on U.S. Chip Technology to China. Center for Strategic & International Studies. https://www.csis.org/analysis/balancing-ledger-export-controls-us-chip-technology-china
- Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755–759. https://doi.org/10.1038/s41586-024-07566-y

- Sun, Y., Li, X., Dalal, K., Xu, J., Vikram, A., Zhang, G., Dubois, Y., Chen, X., Wang, X., Koyejo, S., Hashimoto, T., & Guestrin, C. (2024). Learning to (Learn at Test Time): RNNs with Expressive Hidden States (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2407.04620

- Villalobos, P., Ho, A., Sevilla, J., Besiroglu, T., Heim, L., & Hobbhahn, M. (2022). Will we run out of data? Limits of LLM scaling based on human-generated data (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2211.04325

- von der Leyen, U. (2024). Europe's Choice. Political Guidelines 2024-2029. https://commission.europa.eu/about-european-commission/president-elect-ursula-von-der-leyen_en

- Wiggers, K. (2024, March 28). AI21 Labs' new AI model can handle more context than most. Tech Crunch. https://techcrunch.com/2024/03/28/ai21-labs-new-text-generating-ai-model-is-more-efficient-than-most/?guccounter=1

- Wu, T., Yuan, W., Golovneva, O., Xu, J., Tian, Y., Jiao, J., Weston, J., & Sukhbaatar, S. (2024). Meta-Rewarding Language Models: Self-Improving Alignment with LLM-as-a-Meta-Judge (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2407.19594

- Zhang, E., Zhu, V., Saphra, N., Kleiman, A., Edelman, B. L., Tambe, M., Kakade, S. M., & Malach, E. (2024). Transcendence: Generative Models Can Outperform The Experts That Train Them (Version 3). arXiv. https://doi.org/10.48550/ARXIV.2406.11741

About the Author

Anselm Küsters is Head of the Department of Digitisation / New Technologies at the Centre for European Policy (cep) in Berlin. He works in the field of Digital Humanities as an affiliated researcher at the Max Planck Institute for Legal History and Legal Theory in Frankfurt am Main and as a post-doctoral researcher at the Humboldt University in Berlin. His current research uses Natural Language Processing to analyse and classify discourses on technology from a historical perspective. Anselm holds a doctorate from Goethe University Frankfurt am Main (Dr. Phil.) and studied Economic and Social History at the University of Oxford, UK (M.Phil.), European Law at the University of Würzburg (LL.M.), and History, German Literature, and Economics at the University of Heidelberg (B.A. and B.Sc.).

3. How to strengthen the EU's position in the AI industry: The importance of computing power

Pauline Weil
Independent Economist

As digitalization unfolds, companies need more and more computing power. Artificial intelligence is an especially compute hungry domain, as requires extensive computation on specialized hardware such as Graphics Processing Units (GPUs). Instead of purchasing in house computer resources they increasingly turn to providers offering remote access to computers: this is the cloud computing market. As a most compute hungry sector, artificial intelligence (AI) relies on this service. Currently, the EU does not have homegrown companies offering high end computing hardware and cloud computing at scale. The topic of this paper is to discuss what this entails for the EU's position in the AI space. This paper argues that the EU should aim to attract investments for more computational resources on EU soils. But the priority should be on ensuring access to the most high-end compute for European actors – public and private – to develop AI models and applications. This goal must be pursued in line with a broader strategy of AI's potential for EU industry. In parallel, the EU should also engage in international cooperation to secure a global level playing field and agreements on balancing AI risks.

AI firms are mostly from the United States

The European Union has a limited foothold in the Artificial Intelligence (AI) industry. In 2024, 43% of large-scale AI systems were from the United States (United States), 28% from China and only 6% from the European Union (EU)[1].

This state of play stems from different focuses of industrial strategies, with the EU focusing historically more on medium tech sectors (Terzi et. al. 2022). In comparison, the US and China have a built an ecosystem more conducive to the rise of tech giants. In addition to policy, these tech giants owe their enduring growth to three main factors: big national market offering scalability opportunities, a fit-for-purpose financial sector, which also supported, high research and development (R&D) investments over the years (European Commission 2022). The EU lags in both private and public investments compared to the US and this investment gap is notably big in the information, communication and technology (ICT) sector[2].

The EU's still fragmented capital market is not well suited to finance the risky investments of start-ups nor does it offer many options for initial public offerings. These investments are much better conducted by venture capital than banks and the EU has a much smaller venture capital ecosystem than that of the US (Demertzis et. al. 2021). The IMF finds that over the past ten years, European venture capital investments stood at 0.3% of GDP, three times less than that of the US (Arnold et. al. 2024). Specifically in the AI sector, the OECD finds that over the past five years (2019-2023) venture capital investments in the US amounted to USD 314 billion[3]. This is over twice the USD 136 billion recorded in China and over four times the EU's USD 71 billion.

The main features of the AI sector stem from those of its native land: the high technology sector. Concentration is constitutive of the AI industry as big incumbent concentrate the resources needed to build and power AI models (Kaltheuner and Saari 2024). Although there is currently no monopoly, there is a clear stronghold of a handful of big tech actors which translate into major infrastructure and ecosystem dependencies (Martins 2024). Indeed, to enter the Generative AI (GenAI) market, that is to train a large-scale AI model such as ChatGPT or Llama, start-ups need access to high-end computer power that is currently concentrated in the biggest American technology firms to the likes of Amazon, Microsoft, and Alphabet. In addition to capacity for investments, large US and Chinese-based tech firms also concentrate access to talent, data and consumers.

1 See : Our World in Data (2024) "Cumulative number of large-scale AI systems by country since 2017" Available at: https://ourworldindata.org/grapher/cumulative-number-of-large-scale-ai-systems-by-country
2 European Central Bank (2024) "From laggard to leader? Closing the euro area's technology gap", speech by I. Schnabel,. Available at: https://www.ecb.europa.eu/press/key/date/2024/html/ecb.sp240216~df6f8d9c31.en.html
3 OECD AI Policy Observatory. Available at: https://www.oecd.org/en/topics/policy-issues/artificial-intelligence.html#related-policy-issues

The high-end computing industry is concentrated around a few US firms

Compute refers to the computer hardware needed to store, process or transfer data. In practice, computers are a gathering of different units performing either calculation or memory tasks. These units, commonly referred to as chips or semiconductors, have become ubiquitous com-modities with the advent of digitalization. Shocks to pro-duction and trade logistics, notably during the Covid-19 pandemic, and ever-increasing demand, have shed light on bottlenecks in this highly concentrated industry. Be-cause of the importance of chips as inputs to production across industries, a lot of countries have set up industrial strategies to ensure sustainable supply since the Covid-19 shortages (Poitiers and Weil 2021).

Some high-end chips have emerged as best performers for the AI industry, namely the last generations of Graphic Processing Units (GPUs). Originally designed for image processing notably for the gaming industry, these are very good at processing many small tasks simultaneously. Currently, only three companies design them: Nvidia (US) dominates by far, followed by ARM (United Kingdom) and Intel (US). And only one company, TSMC (Taiwan), fabricates them. The demand for these powerful state-of-the-art chips is currently growing much faster than the industry can supply, which creates problematic bottle-necks and contributes to sky-rocketing prices. Nvidia's H100, the chip deemed most efficient for AI, are sold for $15,000-40,000 each depending on size order[4]. Training a model the size of ChatGPT requires a supercomputer made of around 8,000 H100 chips – which means upfront compute costs from $120,000 to $320,000, not account-ing for operating costs[5]. Incidently, Nvidia is on trajectory to surpass both Microsoft and Apple to become the most valuable company[6]. If AI model sizes continue growing along current trajectory, estimates consider that compute costs could exceed the entire US GDP by 2037 (Vipra and West 2023).

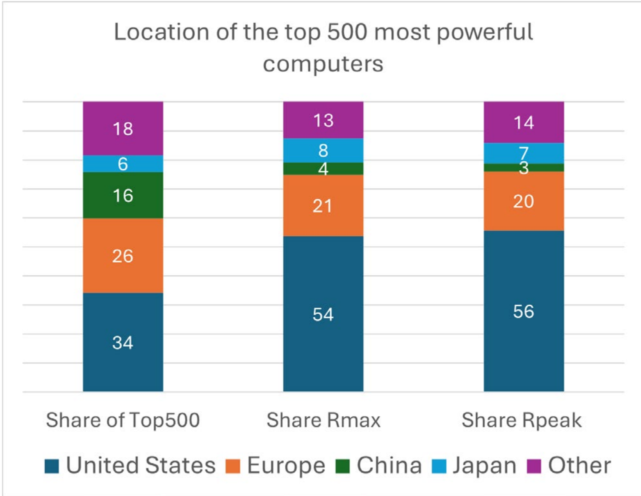

Figure 1: Location of the top 500 most powerful computers

Source: Top500

Currently, as shown in Figure 1, Europe lags far behind the US in terms of its share in the most powerful computers globally, and even more when looking at their perfor-mance (Rmax is the maximum potential performance a computer could reach and Rpeak is its best reviewed per-formance). China's ambitions have so far been relatively curtailed by US efforts to block its access to the last gen-eration of chips. In the EU only 4 countries have compute representing above 1% of the five hundred most powerful computers: Germany, France and Italy each hold 4% and the Netherlands 1%[7].

4 New York Times (2024) "Nvidia, Powered by A.I. Boom, Reports Soaring Revenue and Profits". Available at: https://www.nytimes.com/2024/05/22/technology/nvidia-quarterly-earnings-results.html
5 The Guardian (2024) "Nvidia: what's so good about the tech firm's new AI superchip?". Available at: https://www.theguardian.com/business/2024/mar/19/nvidia-tech-ai-superchip-artificial-intelligence-humanoid-robots
6 Financial Times (2024) "Nvidia vaults past Apple and Microsoft to become world's most valuable company". Available at: https://www.ft.com/content/1ec5523a-80ec-47ba-9f71-8765b4ae4577
7 Top500 Statistics. Available at: https://top500.org/statistics/list/

AI firms rely on cloud computing

Companies that own vast amounts of computer power rent them out to third companies: this is the data center or cloud computing market. Instead of purchasing com-puters, AI firms tend to rent computing power because acquiring compute is prohibitively expensive and less efficient[8]. In the EU, 41% of companies use cloud ser-vices, the figure goes up to 76% in the information and communication sector[9]. In the AI sector, firms turn even more to cloud service because demand is often for imme-diate, punctual and large needs. Computational needs are measured in terms of time – duration of training – and compute – processing power and memory needed - which are measured in FLOPs (or floating point operations per seconds).

Indeed, the advent of the AI industry, along with deep learning, has fueled exponential demand growth for high-end compute: the amount of compute needed to train models has been doubling every six to ten months (Vipra and West 2023)[10]. The large language models that power generative AI (GenAI) are the most compute demanding.

Already pinned as a sector of importance for European strategic autonomy, the **cloud computing sector** is cen-tral to shaping the AI industry: it is a determinant in the strength of AI models and the main cost. Model perfor-mance is found to be predicted by the compute power on which models where trained (Khan 2020). Training costs for the last generation of GenAI models are growing ex-ponentially: in 2022, Google's PaLM cost $ 12.4 million to train, in 2023 OpenAI's Chat GPT-4 cost $78.4 million and in 2024, Google's Gemini Ultra cost $ 191.4 million – over ten times more than its model two years earlier[11]. Compa-nies are spending up to 80% of their total capital on com-pute resources[12].

As explained, the market is dominated by a few American and Chinese giants. 80% of the cloud computing markets is held by 8 companies, including Amazon (31% market share), Microsoft (25%) and Alphabet (11%)[13]. This mar-ket concentration will linger as those same companies currently dominating are those investing the most in new chips (Boakye et. al. 2023).

The leverage of big tech firms is further strengthened as they are embracing integration strategies. These big tech firms compete in the AI industry by building their own in-house AI models or investing in new models. Upstream, Amazon, Meta and Microsoft, aim to move to the compute design sector. Downstream, they ambition to enter the software markets. Existing market dominance and inte-gration ambitions create strong interdependencies and competition concerns (see Christophe Carugati's paper). For instance, the leader in computer chips design, Nvidia, has chosen to sell more to smaller cloud companies that do not ambition to enter the design market (Vipra and West 2023). On the other hand, cloud computing firms could push certain AI products in markets that they al-ready dominate. They could also impose the use of spe-cific software when licensing hardware. AI Now reports that some AI companies are concerned with cloud provid-ers' capacity for market control (Vipra and West 2023).

Responding to concerns of those actors, and supporting the growth of EU actors, needs state support that includes antitrust investigations but also market intervention to en-sure access to capital, computers and know-how. In turn, it could foster a diversification of the ecosystem benefi-cial to innovation, competition and consumer protection.

Current European industrial policy for computing power

At the European level, the main goal of the compute strat-egy are to ensure access, by pooling compute resources. Sharing compute power at the EU level supports an up-take of European ethical approach and strengthening of the Brussels effects by making access to this infrastruc-ture reliant on joining the AI Pact. More importantly, pool-ing EU resources to provide compute infrastructure allows to bypass the prohibitive high entry cost and defuses of some of the market power of foreign hardware key play-ers.

8 Venture Beat (2024) "The Cloud wins the AI infrastructure debate by default". Available at: https://venturebeat.com/data-infrastructure/the-cloud-wins-the-ai-infrastructure-debate-by-default/

9 OECD (2024) Share of businesses purchasing cloud services. Available at: https://goingdigital.oecd.org/en/indicator/25

10 Since 2015 however, trends in compute growth have split into two: the amount of compute used in large-scale models has been doubling in roughly 9.9 months, while the amount of compute used in regular scale models has been doubling in only about 5.7 months. But large scale do use 2-3 times more in the first place. See: Epoch AI (2022) "Compute Trends across three Eras of Machine Learning". Available at: https://epochai.org/blog/compute-trends

11 VisualCapitalist (2024) Visualizing the training costs of AI models over time. Available at: https://www.visualcapitalist.com/training-costs-of-ai-models-over-time/

12 Andreessen Horowitz (2023) Navigating the High Cost of AI Compute. Available at: https://a16z.com/navigating-the-high-cost-of-ai-compute/

13 Statista (2024) Amazon Maintains Cloud Lead as Microsoft Edges Closer. Available at: https://www.statista.com/chart/18819/worldwide-market-share-of-leading-cloud-infrastructure-service-providers/

In 2024, the European High Performance Computing Joint Undertaking (EuroHPC JU), which consisted of €7 billion of funding for 2021-2027 for EU competitiveness in supercomputing, was amended to include AI specific measures. It now aims to create "AI Factories", AI specific European compute resources readily accessible to home-grown start-ups and academics[14]. Nine computers have been built so far, including Finnish Lumi, the fifth most powerful supercomputer in the world, and Italian Leonardo, the sixth.

The European strategy is also considering the broader supply chain of the industry as it includes plans for a European single market for data and it set up in parallel with European cloud and chips ambitions[15]. Several EU countries have also developed their own strategies, to benefit and contribute to EU funding (Von Thun 2024). Differences among national strategies reflect the difficulty in forging a common European approach. While the EU's AI Act pressed the interest of safety, some countries, notably France, are making European AI competitiveness the priority. Looking forward, this paper argues that Europe should focus on its competitive advantages and push forward a vision that takes stock of its current position in the supply chain.

Policy recommendations

1. Make Europe attractive to compute investments with a focus on resilient power

Fiscal resources should not be used to compete with investments from private big tech actors. Instead, public investments should aim to create an ecosystem that fosters investments. Broadly, this includes pillars such as reforms of the capital market to better finance innovation and ensuring a qualified labor force. Focusing on compute for AI, this entails providing the infrastructure needed to install and run high end compute resources.
Compute represents a sustainability challenge: it is hungry for mineral resources, land and energy. While latency – or internet speed – was expected to rein in, power is increasingly front and center for attracting investments[16]. In the past decade, data centers have accounted for 1 to 1,5% of global electricity use[17]. A middle-ground scenario estimates that in 2027, new AI servers sold in one year could use 85 to 134 terawatt hours (Twh) annually – which represents the annual consumption of Argentina, the Netherland or Sweden[18]. Concerns on how to grapple with this surge in energy demand are already prominent in localities dense with data centers in the US. The expected continued growth of this energy hungry industry creates competition for electricity use between residents and industry, and is especially challenging in a context of transition to green energy sources[19]. Anticipating this surging demand by investing in the resilience of power grids should be a priority for a long term EU AI strategy.
Further, given the fragmentation and inefficiencies of the compute market, some resources sit idle for 20-50% of the time[20]. Optimization of resources through compute specialization, collaboration and decentralization should be incentivized[21].

2. Integrate AI into a broader industrial policy

Ultimately, the goal should be for the EU to not miss out on AI's productivity gains, which will be driven by AI applications more than AI models. This reality check should allow for more targeted investments to curtail excessive public spending and rushing to compete with subsidies given out in competing countries[22]. In 2009, France had rolled out a strategy for a sovereign cloud, which ended with a failed investment of at least €75 million[23]. Since then, France has rallied European Gaia-X which is more fitted to European strengths: a governance body is charge of enforcing coordination for security standards, interoperability and data sharing. An AI strategy should skip the investments in services that the market already provide to focus on access for European actors and enforcement of EU vision. The European HPC initiative keep a narrow focus of providing access to strategic actors, either private entities needing support, or academia. Indeed, access to compute will be continously more important in advancing basic research. Meanwhile, a strategic forsight should focus on integrating AI into government and industrial processes.

14 Euro HPC JU (2024) The 'AI Factories' Amendment to the EuroHPC JU Regulation Enters Into Force. Available at: https://eurohpc-ju.europa.eu/ai-factories-amendment-eurohpc-ju-regulation-enters-force-2024-07-09_en
15 European Commission 'Common European Data Spaces'. Available at: https://digital-strategy.ec.europa.eu/en/policies/data-spaces
16 LightReading (2024) Power, not latency, to guide computing at the edge. Available at: https://www.lightreading.com/the-edge-network/power-not-latency-to-guide-computing-at-the-edge
17 International Energy Agency 'Data Centres and Data Transmission Networks'. Available at: https://www.iea.org/energy-system/buildings/data-centres-and-data-transmission-networks
18 Washington Post (2024) Tech ginats fight plan to make them pay more for electric grip upgrades. Available at: https://www.washingtonpost.com/technology/2024/09/13/data-centers-power-grid-ohio/
19 Washington Post (2024) Amid explosive demand, America running out of power. Available at: https://www.washingtonpost.com/business/2024/03/07/ai-data-centers-power/
20 AIIA, FuriosaAI and ClearML (2024) The State of AI Infrastructure at Scale 2024. Available at: https://ai-infrastructure.org/wp-content/uploads/2024/03/The-State-of-AI-Infrastructure-at-Scale-2024.pdf
21 The Economist (2023) Can computing clean up its act?. Available at: https://www.economist.com/science-and-technology/2023/08/16/can-computing-clean-up-its-act
22 An IMF study showed tit-for-tat dynamic in introducing new subsidies for the same products between the EU, US and China after six months of the new policy. The US and the EU tend to react faster than China. After 2 years, most subsidies have been matched. See Evenett et. al. 2024
23 Sciences Po (2020) Le cloud souverain est de retour: généalogie d'une ambition emblématique de la souveraineté numérique en France. Available at: https://www.sciencespo.fr/public/chaire-numerique/2020/07/20/cloud-souverain-genealogie-ambition-emblematique-souverainete-numerique/

3. Push for international cooperation

AI entails vast opportunities and risks going forward. To ensure a global level playing field and international cooperation to agree on standards, an international body to oversee its development should be set up. Such an organization should draw from experience from other sectors such as climate change (the International Panel on Climate Change), the aviation industry (the International Civil Aviation Organization), or atomic power (the International Atomic Energy Agency or the European Organization for Nuclear Research). All of these endeavours take into account the potential magnitude of risks of these sectors while recognising that their development represents a global public good. Major roles to be endorsed are: setting global standards, provide research on safety, creating expertise hubs outside of the private sector and overall finance prohitively expensive projects. These missions would rely on three major pillars: a social and organisational infrastructure, technical infrastructure (including compute) to support reseach and access to AI models and data. An international organisation has the advantage of being shielded from economic and political pressures allowing to impose standards without the temptation of watering them down to face international competition. Europe should lead the diplomatic effort to build such an organisation. In doing so, Europe should push for building collaborative compute resources and efforts to foster AI as a global public good.

References

- Arnold, N. G, Claveres G. and Frie J. (2024) Stepping Up Venture Capital to Finance Innovation in Europe. IMF Working Papers. International Monetary Fund
- Boakye B., Garson M., Macon-Cooney B., Westgarth T. and Zandermann K. L. (2023) State of Compute Access: How to Bridge the Digital Divide
- Demertzis, M., M. Domínguez-Jiménez and L. Guetta-Jeanrenaud (2021) 'Europe should not neglect its Capital Markets Union', *Policy Contribution* 13/2021, Bruegel
- European Commission (2022) Where the EU stands vis-à-vis the USA and China? Corporate R&D intensity gap and structural change. European Commission, *Science for Policy Brief*
- Evenett S., Jakubik A., Martin F. and R. Michele (2024) The Return of Industrial Policy. IMF Working Papers. International Monetary Fund
- Kaltheuner, F. and L Saari (2024) Public Interest AI for Europe? Shaping Europe's Nascent Industrial Policy. AI Now Institute
- Khan S. M. (2020) AI Chips: What They Are and Why They Matter. Center for Security and Emerging Technology Martins B. (2024) The European Union AI Act: premature or precocious regulation?. Bruegel
- Poiters N. and Weil P. (2021) A new direction for the European Union's half-hearted semiconductor strategy, *Policy Contribution 17/2021*, Bruegel
- Terzi A., Singh A. and M. Sherwood "Industrial Policy for the 21st Century: Lessons from the Past" (2022) European Commission, *Discussion Paper 157*
- Vipra J. and West S. M. (2023) Computational Power and AI. AI Now Institute
- Von Thun M. (2024) To Innovate or to Regulate? The False Dichotomy at the Heart of Europe's Industrial Approach

About the Author

Pauline Weil has worked for two years as a Research Analyst at Bruegel. She holds an MSc in International Trade and Finance from Sciences Po Lille and an MSc in Political Economy of Europe from the London School of Economics.

Her main research interests are high tech trade and agriculture policy. She focuses on macroeconomics and geopolitics, building from a dual expertise on the European Union and Chinese economy.

Prior to Bruegel, Pauline was an Economist for the credit insurer Coface where she provided country risk analysis on Europe, working from Paris, and then on Asia, from Hong Kong. She also worked on the European Commission's development aid policy, working for DG INTPA in the Directorate for Asia.

Pauline is fluent in French and English and has a good command of Spanish.

4. How to strengthen the EU's position in the AI industry: Ensuring vibrant competition

Dr. Christophe Carugati
Digital Competition

GenAI is rapidly evolving, with constant market and regulatory changes. To ensure a vibrant GenAI landscape in Europe, competition authorities in Europe should closely monitor developments, collaborate with other regulatory bodies, and actively engage with industry stakeholders.

Introduction

Generative Artificial Intelligence (GenAI) refers to AI that generates content such as text, videos, and images from large datasets. Since the release of the OpenAI-owned application ChatGPT in 2022, GenAI has been under scrutiny from industry, investors, and governments due to its potential to deliver substantial societal and economic benefits in the coming years (Carugati, 2023a).

However, it will only reach its full potential if competition remains vibrant. In this context, competition authorities in Europe and worldwide closely monitor market and regulatory developments to understand how GenAI works[1]. They have initiated market studies into the entire value chain (Competition and Markets Authority, 2024) and related markets where GenAI is deployed, such as general search engine services (Australian Competition and Consumer Commission, 2024). Investigations are also launched into partnerships between large cloud providers and model developers to assess whether they could negatively impact competition[2].

Despite several ongoing market studies and investigations, competition authorities have already expressed concerns about the risks of concentration in the hands of a few large online platforms and potential anticompetitive practices that might distort the competitive process[3]. Therefore, authorities are committed to intervening quickly before harm occurs, learning from the slow interventions in digital markets in the past[4].

Against this backdrop, this paper provides a comprehensive understanding of the competition dynamics in the GenAI market in Europe. It begins by outlining market developments, finding that GenAI is fast-moving, characterised by intense competition and innovation among industry players, with potential risks of anticompetitive practices. The paper then examines regulatory developments, noting that digital and competition regulations significantly impact market developments in Europe, potentially causing Europe to lag in GenAI. Finally, the paper concludes with policy recommendations for competent authorities to ensure robust competition in GenAI in Europe.

Market Developments

GenAI relies on three main components: models, data, and computing power. Research papers and market stud-ies have already highlighted these elements' crucial roles in GenAI development and deployment (Competition and Markets Authority, 2024). This section summarises the key findings and recent developments across both the de-velopment and deployment stages.

1 Christophe Carugati, GenAI and Competition Hub (accessed 16 April 2024). Available at: https://www.digital-competition.com/genaiandcompetitionhub

2 See, for instance, Competition and Markets Authority, Microsoft / OpenAI partnership merger inquiry (accessed 16 April 2024). Available at: https://www.gov.uk/cma-cases/microsoft-slash-openai-partnership-merger-inquiry

3 See, for instance, Margrethe Vestager et al., Joint Statement on Competition in Generative AI Foundation Models and AI Products, 23 July 2024 (accessed 16 April 2024). Available at: https://competition-policy.ec.europa.eu/document/download/79948846-4605-4c3a-94a6-044e344acc33_en?filename=20240723_competition_in_generative_AI_joint_statement_COMP-CMA-DOJ-FTC.pdf

4 Sarah Cardell, Opening Remarks at the American Bar Association (ABA) Chair's Showcase on AI Foundation Models, Competition and Markets Authority, 11 April 2024 (accessed 16 April 2024). Available at: https://www.gov.uk/government/speeches/opening-remarks-at-the-american-bar-association-aba-chairs-showcase-on-ai-foundation-models

Development Stage: Necessary Inputs to Develop Models

Models serve as the essential backbone of GenAI. Developers frequently release new foundation models in various sizes (large, small, and on-device), capacities (ranging from general-purpose to task-specific models), and levels of accessibility (from open-source to proprietary). In turn, developers use FMs to build a new version of the models or a fined-tuned one, such as Meta Llama 3.1[5], and Meta Code Llama for coding built from Meta Llama 2[6]. Application developers then use these models to create GenAI applications downstream, such as ChatGPT, which generates text, developed from various GPT models[7].

Access to large, varied and high-quality datasets is crucial for model development. In essence, data is the lifeblood of these models. Developers use public datasets from the web and public repositories, as well as proprietary datasets from first- and third-party sources, during the training phase. Additionally, AI-generated synthetic data is increasingly used to create new datasets from original data. Proprietary datasets are also employed during the fine-tuning phase and for generating content from external sources during the grounding phase.

However, access to data is becoming more challenging for model developers. Legitimate concerns regarding the necessity for rightsholders' consent are increasingly limiting data collection of third-party proprietary datasets. In response, model developers have started forming partnerships with publishers to secure access to these data while avoiding legal issues[8].

Finally, computing power is essential for model developers. Compute acts as the engine of GenAI models. Specialised chips are required to run these models, which need computing resources during both the training and inference phases when users call the models to generate outputs. Large models are significantly more compute-intensive than small and on-device models, which can operate directly on devices like smartphones. Compute-intensive models predominantly run on large cloud providers, which offer flexible, scalable computing resources equipped with up-to-date chips, eliminating the need for developers to invest in costly infrastructure.

However, access to chips is also a growing concern for model developers. The chip sector is currently experiencing a shortage due to high demand from AI applications and limited supply, exacerbated by factory closures during the COVID-19 pandemic. Moreover, only a few chip providers, including Nvidia and AMD, and cloud providers like Amazon, Microsoft, and Google, can supply chips for AI applications, with Nvidia being the dominant supplier (Carugati, 2023b). As a result, model developers have increasingly turned to partnerships with large cloud providers to secure the necessary computing resources.

These partnerships can be either exclusive or non-exclusive. In exclusive partnerships, like the *Microsoft/OpenAI* collaboration, models are hosted and run solely on the partner's infrastructure, with the cloud provider integrating the models into its flagship products and services, such as Microsoft Office. Non-exclusive partnerships, like the Microsoft/MistralAI partnership, allow models to be hosted and run on alternative cloud providers.

In Europe, the Commission determined that the *Microsoft/OpenAI* partnership does not constitute a concentration under EU merger law due to the lack of Microsoft's control over OpenAI. However, it continues to investigate potential competition issues related to exclusivity clauses in the partnership[9]. In Germany, the German competition authority found that the *Microsoft/OpenAI* partnership does constitute a concentration under German merger law due to Microsoft's material competitive influence, but it did not investigate further as the merger did not meet the national jurisdictional threshold that triggers a merger review[10]. Meanwhile, the UK competition authority is investigating several partnerships in the UK, including *Microsoft/OpenAI*[11], *Amazon/Anthropic*[12], *Google/Anthropic*[13] *Microsoft/InflectionAI*[14]. It closed the *Microsoft/MistralAI* partnership investigation as it did not meet the UK's merger review criteria because Mistral remains independent from Microsoft[15]. It also cleared the *Microsoft/Inflection AI* arrangement, as the deal is unlikely to have negative effects on competition for the development and supply of consumer chatbots and FMs[16]. Finally, it cleared the *Amazon/Anthropic* partnership, as the deal does not meet the UK merger control threshold[17].

5 Meta, Introducing Meta Llama 3: The Most Capable Openly Available LLM to Date, 18 April 2024 (accessed 16 April 2024). Available at: https://ai.meta.com/blog/meta-llama-3/
6 Meta, Code Llama, A State-Of-The-Art Large Language Model for Coding (accessed 16 August 2024). Available at: https://llama.meta.com/code-llama/
7 Adobe Firefly (accessed 16 April 2024). Available at: https://www.adobe.com/products/firefly.html
8 Platforms and Publishers: AI Partnership Tracker (accessed 16 August 2024). Available at: https://petebrown.quarto.pub/pnp-ai-partnerships/
9 Margrethe Vestager, Speech by EVP Margrethe Vestager at the European Commission workshop on "Competition in Virtual Worlds and Generative AI", *European Commission*, 28 June 2024 (accessed 19 July 2024). Available at: https://ec.europa.eu/commission/presscorner/detail/en/speech_24_3550
10 Bundeskartellamt, Cooperation Between Microsoft and OpenAI Currently Not Subject to Merger Control, 15 November 2023 (accessed 19 August 2024). Available at: https://www.bundeskartellamt.de/SharedDocs/Meldung/EN/Pressemitteilungen/2023/15_11_2023_Microsoft_OpenAI.html
11 Competition and Markets Authority, *Microsoft / OpenAI* Partnership Merger Inquiry (accessed 19 August 2024). Available at: https://www.gov.uk/cma-cases/microsoft-slash-openai-partnership-merger-inquiry
12 Competition and Markets Authority, *Amazon / Anthropic* Partnership Merger Inquiry (accessed 19 August 2024). Available at: https://www.gov.uk/cma-cases/amazon-slash-anthropic-partnership-merger-inquiry
13 Competition and Markets Authority, *Alphabet Inc. (Google LLC) / Anthropic* Merger Inquiry (accessed 19 August 2024). Available at: https://www.gov.uk/cma-cases/alphabet-inc-google-llc-slash-anthropic-merger-inquiry
14 Competition and Markets Authority, *Microsoft / Inflection* Inquiry (accessed 19 August 2024). Available at: https://www.gov.uk/cma-cases/microsoft-slash-inflection-ai-inquiry
15 Competition and Markets Authority, *Microsoft / MistralAI Partnership* Merger Inquiry (accessed 19 August 2024). Available at: https://www.gov.uk/cma-cases/microsoft-slash-mistral-ai-partnership-merger-inquiry
16 Competition and Markets Authority, *Microsoft / Inflection* inquiry.
17 Competition and Markets Authority, *Amazon / Anthropic* Partnership Merger Inquiry.

Deployment Stage: Channels to Supply Applications

Application developers of all sizes use GenAI models to create applications across various sectors, from legal services to photography. As illustrated by the Microsoft/OpenAI partnership, some large online platforms are integrating GenAI applications into their flagship products and services. This raises competition concerns, particularly regarding potential leveraging issues through tying and self-preferencing, where a dominant firm uses its dominant position in one market to promote its services in another.

Furthermore, some large platforms have partnered with smartphone manufacturers to deploy GenAI applications. For example, the Apple/OpenAI partnership allows OpenAI to deploy its models and applications on certain iPhones, while Apple integrates these models into its Apple Intelligence service for AI features[18]. Similarly, the Samsung/Google partnership enables the pre-installation of Google's small language models on select Samsung devices, which is currently under the Commission's investigation[19].

Regulatory Developments

GenAI operates within a rapidly evolving and intricate cross-regulatory environment, creating new and complex policy and legal challenges. Regulations related to data protection, intellectual property rights, consumer protec-tion, competition, and digital markets significantly influ-ence GenAI development and deployment.

Development Stage: Issues Preventing Model Developments

Model developers often use personal data to develop their models, which requires strict adherence to data protection regulations, such as the General Data Protec-tion Regulation (GDPR) in Europe. As a result, developers face significant data protection challenges. For instance, the Italian Data Protection Authority (DPA) temporarily banned ChatGPT in Italy due to several concerns, includ-ing inadequate information disclosure, lack of a legal ba-sis, and insufficient age verification systems[20]. Moreover, Meta had to suspend the development of its models and related applications after the Irish DPA raised questions about the legal basis for training these models[21]. These regulatory developments directly impact model development in Europe, leading to delays and legal uncertainty and affecting competition in GenAI.

In addition to personal data, model developers use public and proprietary data, which may be subject to copyright protection. If copyright applies, developers must obtain consent from rightsholders and may need to compensate them for using their data. However, the scope of copyright protection in the context of GenAI is still under debate in Europe and globally. The European Union's Artificial Intelligence Act, which governs AI, mandates that providers of general-purpose AI models comply with copyright laws and provide a detailed summary of the training data used. This requirement aims to ensure that rightsholders can assert their rights. In France, the competition authority fined Google for failing to inform media publishers about the use of their content in its chatbot, Google Bard, violating previous commitments related to neighbouring rights[22]. Furthermore, it found in its report on GenAI that partnerships between model developers and publishers raise potential competition concerns over exclusivity or high compensation that could prevent other developers from securing similar partnerships (Autorité de la concurrence, 2024). This exclusionary effect might also result from dominant players refusing to provide access to copyrighted data, potentially constituting a refusal to deal (Graef, Tombal, and de Streel, 2019).

18 OpenAI, OpenAli and Apple Announce Partnership to Integrate ChatGPT into Apple Experiences, 10 June 2024 (accessed 19 August 2024). Available at: https://openai.com/index/openai-and-apple-announce-partnership/

19 Margrethe Vestager, Speech by EVP Margrethe Vestager at the European Commission workshop on "Competition in Virtual Worlds and Generative AI", European Commission, 28 June 2024 (accessed 19 July 2024).

20 Garante per la Protezione dei Dati Personali, Artificial Intelligence: Stop to ChatGPT by the Italian SA Personal Data is Collected Unlawfully, No Age Verification System is In Place for Children, 31 March 2023 (accessed 19 August 2024). Available at: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847#english

21 Stefano Fratta, Building AI Technology for Europeans in a Transparent and Responsible Way, Meta, 10 June 2024 (accessed 19 August 2024). Available at: https://about.fb.com/news/2024/06/building-ai-technology-for-europeans-in-a-transparent-and-responsible-way/

22 Autorité de la concurrence, Related Rights: The Autorité Fines Google €250 Million for Non-Compliance with Some of its Commitments Made in June 2022, 20 March 2024 (accessed 19 August 2024). Available at: https://www.autoritedelaconcurrence.fr/en/press-release/related-rights-autorite-fines-google-eu250-million-non-compliance-some-its

Deployment Stage: Issues Preventing Application Deployment

Some large online platforms integrate GenAI models and applications into their existing products and services. These practices raise competition concerns and are subject to competition and digital markets laws, including the Digital Markets Act (DMA), which aims to ensure contestability and fairness in Europe. However, the DMA's implementation has introduced uncertainties that may limit deployment. For instance, Apple is reportedly delaying the launch of its Apple Intelligence service due to concerns related to DMA provisions on access conditions to certain functionalities mandated by the Act[23]. Under the DMA, the High-Level Group (HLG) of the DMA allows authorities from different policy areas to work together. Composed of the European Regulators for electronic communications, data protection, consumer protection, competition, and audiovisual media, the HLG issued a joint statement on AI, outlining plans to coordinate efforts to ensure coherence in AI developments[24].

Policy Recommendations

The paper underscores the rapidly evolving market and regulatory landscape in GenAI, which has created uncertainty for model and application developers at both the development and deployment stages. In light of these challenges, competition authorities in Europe should consider the following recommendations to maintain vibrant competition in GenAI in Europe.

First, ***competition authorities should closely monitor the GenAI sector to keep up with its fast-paced changes***. This involves tracking market dynamics and regulatory developments as part of their ongoing oversight activities. Similar to the approach taken by the UK competition authority in its GenAI study, authorities should establish non-binding competition principles that encourage the openness of GenAI markets to alternative providers considering the current competitive market dynamics and lack of market failures. Formal investigations should be initiated only when necessary to address substantiated concerns.

Second, ***competition authorities should collaborate with other regulatory bodies to ensure a consistent approach across policy areas***. In Europe, this should involve coordinating efforts within the framework of the HLG of the DMA. It should establish working groups to address issues beyond the DMA and intersect with other areas, such as data protection and competition.

Third, ***competition authorities should adopt a co-regulatory approach in collaboration with the industry to support developing and deploying products and services in Europe***. By engaging with economic actors, authorities can better understand ongoing developments and address novel and complex cross-regulatory issues. This approach should allow firms to seek guidance on their practices before implementation, fostering a more supportive and innovative regulatory environment.

23 Foo Yun Chee, Apple to Delay Launch of AI-Powered Features in Europe, Blames EU Tech Rules, Reuters, 21 June 2024 (accessed 19 August 2024). Available at: https://www.reuters.com/technology/artificial-intelligence/apple-delay-launch-ai-powered-features-europe-blames-eu-tech-rules-2024-06-21/

24 European Commission, High-Level Group for the Digital Markets Act Public Statement on Artificial Intelligence, 22 May 2024 (accessed 6 August 2024). Available at: https://digital-markets-act.ec.europa.eu/high-level-group-digital-markets-act-public-statement-artificial-intelligence-2024-05-22_en

References

- Australian Competition and Consumer Commission (2024) 'Digital Platform Services Inquiry – March 2025 – Final Report'.
- Autorité de la concurrence (2024) 'Opinion 24-A-05 of June 28, 2024 on the Competitive Functioning of the Generative Artificial Intelligence Sector'.
- Carugati, C. (2023a) 'Competition in Generative Artificial Intelligence Foundation Models', *Bruegel.*
- Carugati, C. (2023b) 'The Competitive Relationship between Cloud Computing and Generative AI', *Bruegel.*
- Competition and Markets Authority (2024) AI Foundation Models: Technical Update Report.
- Graef, I., Tombal, T., and De Streel. A. (2019) 'Limits and Enablers of Data Sharing. An Analytical Framework for EU Competition, Data Protection and Consumer Law', *TILEC Discussion Paper.*

About the Author

Digital Competition (www.digital-competition.com) is a research and advisory firm dedicated to advancing open digital and competition policies that promote innovation. We provide our members and clients with impartial, forward-looking analyses on emerging global digital and competition issues, helping to shape policies that benefit everyone. We engage with stakeholders, offer insights, and make policy recommendations on complex and novel policy developments. This paper was conducted independently and received financial support from the Friedrich Naumann Foundation.

Dr. Christophe Carugati (christophe.carugati@digital-competition.com) is the founder of Digital Competition. He is a renowned and passionate expert on digital and competition issues with a strong reputation for doing impartial, high-quality research. After his PhD in law and economics on Big Data and Competition Law, he is an ex-affiliate fellow at the economic think-tank Bruegel and a lecturer in competition law and economics at Lille University.

freiheit.org