Ber

m

S

brue el

1 Artificial intelligence competition concerns

The artificial intelligence industry is booming, powering stock market valuations[1]. From 2022 to 2023, investment in generative AI in the United States by big tech companies and private venture capital in AI start-ups increased from less than €1 billion to over €20 billion (Madiega and Ilnicki, 2024). In the EU, it increased from almost zero to nearly €4 billion over the same period.

However, a series of high-profile collaboration agreements between AI start-ups and big-tech firms, combined with emerging bottlenecks in AI input markets, have drawn the attention of competition authorities. There is suspicion that big tech companies are jostling to strengthen their market positions in the AI industry and that they use these agreements to stifle competition and increase the dependency of, and restrict the room for manoeuvre for, start-ups. Investigations into these agreements have been started in France, Portugal, Hungary, the United Kingdom and the US[2]. The US Department of Justice in June 2024 announced "*urgent*" scrutiny of big tech's control of AI[3], in particular choke points in the supply chain, including access to computing power, data and dedicated AI processors. Moreover, acquisitions of entire teams of AI engineers from other firms, such as Microsoft's hiring of Inflection AI staff in early 2024[4], are perceived as a way to circumvent merger regulation.

Such competition checks are considered urgent because competition authorities do not want to be caught short a second time by big-tech firms, as happened over the past decade when a few online platforms grew very fast and managed to carve out dominant market positions. Classic slow-grinding competition policy procedures were unable to catch up with them. A major justification for the EU Digital Markets Act (DMA, Regulation (EU) 2022/1925) – the main EU competition policy tool for very large digital platforms – was precisely to create a fast *ex-ante* instrument so that policymakers would no longer have to wait for a competition problem to occur before they could intervene.

The French Autorité de la Concurrence (2024) issued an opinion in June 2024 on competition bottlenecks in the AI value chain. The UK Competition and Markets Authority (2024) noted that the growing presence of a few big-tech firms that underpin AI by providing computing resources, expertise and monetisation channels, might shape AI-related markets to the detriment of fair, open and effective competition. The European Commission as the European Union's competition authority is reviewing collaboration agreements and called at the start of 2024 for contributions on competition in generative AI[5]. AI systems are already integrated in services operated by the hard-to-avoid 'gatekeeper' platforms that are monitored under the DMA.

This Policy Brief examines competition-reducing market-entry barriers in each segment in the AI value chain, from upstream markets for model training inputs, intermediate markets that match AI model developers with deployers of AI-driven services and downstream markets that match deployers and end users. To structure the debate, it proposes a collaboration and competition or "*co-opetition*" rationale (Brandenburger and Nalebuff, 1996) for collaboration agreements between AI start-ups and big-tech firms. It examines possible policy responses to these bottlenecks and shows how, in some parts of the value chain, there are no clear-cut solutions because competition authorities are caught between tive text input. Hundreds of thousands of specialised AI model applications run like apps in an app store on top of FMs. By the end of April 2024, the OpenAI GPT applications store contained 159,000 specialised GPT applications created on top of the baseline ChatGPT AI model[9], which provides the equivalent of an 'operating system' for these AI apps, similar to the way Google Android is the operating system for apps running on an Android smartphone.

However, a rst important quali cation to this overall view of the market for AI models is that models should be distinguished according to their availability to users. About two thirds of all FMs are fully open source, available publicly for free commercial and non-commercial use. Open models reduce AI market entry barriers and increase competition between users, giving them a wide range of models to choose from. Depending on the degree of openness, model users can create their own versions of the model (Solaiman, 2023). Once they are released, developers lose control of their open models, which may have consequences for safety and responsible use. Anyone can modify the models for harmful purposes, such as producing fake news, disturbing videos or racist speech.

However, the performance of open models is somewhat below that of closed models (Maslej *et al*, 2024, p 146). e best FM models are not released for open use[10]. at reduces the competitiveness of rms that deploy freely available open models, compared to rms that have access to closed but better performing models – if the underlying AI model is less good, the apps built on top of it will also be less good. is nding contradicts the often-heard claim – copied from the open software movement – that open models accelerate AI-driven innovation because many parties can use them.

2.2 Fi ed computing costs as a barrier to market entr

A crude measure of the capability of AI models is the number of tokens they can handle. Tokens are created by slicing training data into fragments – for example parts of words or sentences – enabling analysis at a very granular level. Since the launch of the rst Google Transformer model (Vashwani *et al*, 2017), AI models have grown exponentially, from millions to trillions of tokens in the latest models (Figure 2). e processing of more tokens require more computing capacity and training data, and so computing capacity requirements and the associated costs have increased accordingly[11] (Maslej *et al*, 2024, p 49-51).

e red line in Figure 2 shows that models are still on a linear growth path on the logarithmic scale of computing needs and costs. ere are no signs yet of diminishing returns to scale, not even across this wide range of orders of magnitude of increase in inputs. Since models are pre-trained, prior to use, training is a xed cost, independent of the intensity of use of the model. High xed training costs are the main AI market entry barrier. is favours large rms with the nancial capacity to cover these costs.

It explains why AI start-ups seek collabora1p5osoins r(-u)1 (lej)c-ups stee orts.

Figure 2: Estimate training costs and computing requirements for selected AI models

Source: Maslej *et al* (2024). Note: * see footnote 11.

Model training requires expensive dedicated AI processors. Nvidia is currently the market leader but competition is picking up, from established chip producers like AMD and Intel, and from big-tech rms building their own AI processors. Other input market segments in the chips supply chain are dominated by a small number of rms. For example Netherlands-based ASML is a market leader in lithography machines for chip production.

To train and run the largest FMs, hundreds of thousands of dedicated AI processors need to be assembled in cloud server farms. Only the largest big-tech players – essentially Google, Amazon, Microsoft and Meta, and Apple and Nvidia, collectively known as the GAMMANs – have the required cloud infrastructure and computing capacity to cater to the training needs of the largest AI models. Meta is reportedly developing a state-of-the art AI computing system, which with 350,000 Nvidia H100 processors[12], each costing around $30,000, is estimated to cost more than $10 billion. Extrapolating the red line in Figure 2 by a few years, it is easy to see that the trillion-dollar AI processing farm would soon be reached. is market entry barrier is completely out of reach of public funding; only the very largest rms can aspire to this, and even

Figure 3: Collaboration agreements between big tech and smaller AI start-ups

Source: CMA (2024).

2.3 Access to model training data

FMs and LLMs are pre-trained on very large text datasets taken from books, documents, Wikipedia and webpages. FMs are essentially statistical models that predict sequences of tokens. They need many occurrences of sequences of tokens in order to make robust predictions. That, in turn, requires large volumes of data. Likewise, FMs that produce images and audio require large volume inputs of audio-visual material.

AI developers have already reached the limits of available high-quality human-edited text data to train their models (Maslej *et al*, 2024, p 52). The supply of lower-quality text from social media, or voice-to-text conversion is less constrained and could be sufficient until the early 2030s on current model size trends. However, low-quality input reduces the quality of model outputs. Using synthetic training data may cause model 'collapse', meaning a dramatic drop in model output quality. This will become an important issue as more digital media content is generated synthetically by AI models. It would trigger a negative synthetic data feedback loop in AI model training. Picture and audio data are sufficiently ubiquitous to keep audio-visual model training going for another two decades.

Many if not most of the text, audio and picture AI training datasets are subject to copyright. Authors can in principle charge license fees for use of their works. That would trigger an additional cost-induced shrinkage in the supply of training data (Gans, 2024). It would also increase the cost of training of models and reduce competition between model developers. The EU AI Act (finalised in May 2024) requires model developers to respect EU copyright law as set out in the Copyright Directive (Directive (EU) 2019/790), in particular Article 4, which grants an exception to copyright for commercial research but allows the copyright holder to opt-out of this exception. The AI Act requires AI model developers to be transparent about the data they use, including with regard to this opt-out. Many copyright holders and their collecting societies have become aware of uses of their creative content for AI training purposes and are introducing explicit opt-outs[14].

In the US, the fair-use and transformative-use exceptions to copyright may apply to AI training data. But this is still subject to legal uncertainty, with several court cases pending. AI investors run the risk of punitive statutory damages if courts decide against the fair-use exception. To avoid this, the largest AI firms have signed data-licensing deals with large media companies. For example, OpenAI signed with the New York Times, the Bertelsmann media group and the Reddit news platform. These deals give access to high-quality human-edited text for model training and to recent information that can be 'grounded' into models without having to go through costly re-training. But even the largest AI firms are unlikely to sign licensing deals with all copyright holders. Strict enforcement of copyright law will increase the price of access to training data. That is likely to reduce the volume of copyright-protected

14 See for example Brad Spitz 'AI data mining: French music collecting society Sacem opts out', *Kluwer Copyright Blog*, 25 January 2024, https://copyrightblog.kluweriplaw.com/2024/01/25/ai-data-mining-french-music-

training data that model developers will access, thus weakening model performance. Smaller AI developers and start-ups may not have the financial resources to pay for copyright licenses and may be pushed out of the market altogether. Smaller models with less data and tokens can still be trained, but will not be at the frontier of AI model performance.

Global differences in copyright regimes may distort the geographic level playing field for AI developers. They may choose to move model training to countries with more favourable copyright regimes. The EU AI Act requires that all models deployed in the EU respect EU copyright law. This 'Brussels effect' may force AI firms based outside the EU to comply with EU copyright law, even though copyright law is essentially territorial. Alternatively, a reversal of the Brussels effect might keep the best models out of the EU market, at the expense of EU consumer welfare and businesses productivity[15].

2.4 Access to intermediate model deployers and end users

AI model developers need business channels to generate revenue to cover the costs of training and running their models. Some start-ups try to build their own business models from scratch. For example, OpenAI reached more than 100 million users within a year of the launch of ChatGPT. It has created a GPT app store which downstream model deployers have populated with hundreds of thousands of specialised ChatGPT applications. Application developers have access to the open ChatGPT model but pay an app-store entry fee. OpenAI also charges subscription fees for users of the professional version of ChatGPT. The app store generates network effects that make ChatGPT even more attractive: more users attract more application developers, and vice versa. Responsibility for compliance with AI Act standards shifts from model developers to deployers.

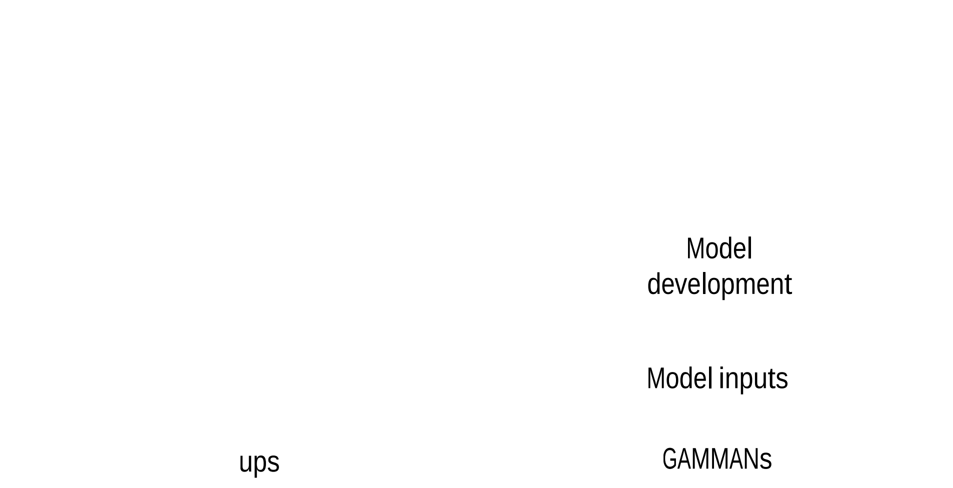
However, starting a business model from scratch is hard for less-successful AI start-ups with weaker or no network effects. An easier route to revenue is to collaborate with the GAMMANs and embed AI models into their well-established business models. For example, Microsoft is embedding its own and third-party AI models into all its productivity software and the Bing search engine. It charges premium prices for access to some of its AI-driven services. Google is doing the same for its search engine and other services. Meta has substantially increased its advertising revenue with the help of AI systems[16].

This results in "*co-opetition*" agreements (Brandenburger and Nalebuff, 1996) between GAMMANs and AI start-ups. Start-ups collaborate with GAMMANs to embed their AI models into existing GAMMAN user-facing services at the downstream end of the value chain, in return for reverse collaboration at the upstream end of the value chain, where GAMMANs grant start-ups access to computing infrastructure and possibly training data. In the middle of the value chain however, collaboration is replaced by competition: start-up AI models compete with the GAMMAN's own AI models. And the parties must bargain over how to divide the value chain they create together. GAMMANs may be vertically integrated along the entire AI value chain while start-ups cover mostly the input and intermediate parts of the value chain.

15 For more details on the AI and copyright debate, see Martens (2024).

16 Jo Arazi, 'Meta's Q1 Ad Revenue Soars 27% Amid AI-Driven Growth', *West Islin* [illegible]

Figure 4: The collaboration-competition model between AI start-ups and GAMMANs



Source: Bruegel.

3 Polic responses to bottlenecks in the AI value chain

is section examines competition policy concerns that may arise in this competitive symbiosis between GAMMANs and start-ups, and possible tools to deal with these concerns.

3.1 Bottlenecks in upstream AI model inputs markets

Hardware inputs

GAMMANs have their own hyperscale hardware facilities. Hardware is a rival good; it can only be used by one party at the time. It is di cult to nd a market-clearing mechanism that ensures open and fair access for all contenders to available AI processor chips and cloud computing capacity. First-come- rst-served is not a viable strategy because the rst could buy the entire production capacity and re-sell it later. Price auctions are also hard to conceive. Quota would be di cult to allocate on a fair basis: how much to whom? e essential facilities doctrine (Graef, 2019) under Article 102 TFEU is not applicable because there is no single unique source of hardware inputs[17].

is makes it di cult to issue a practical policy recommendation. An optimist might assume that hardware bottlenecks are just temporary transition problems as new AI processors from competing manufacturers, and additional cloud computing capacity from new market entrants, enter the market. ere is no guarantee however that this bottleneck will disappear soon. e UK CMA is prudent and advocates "*vigilance*" in this segment of the AI value chain. e French competition authority (Autorité de la Concurrence, 2024) suggests public sector investment in EU supercomputing capacity – in other words, government subsidies. Given the exponential growth in AI computing requirements (Figure 2), it is unlikely that the public sector has the nancial capacity to o er a viable alternative. Also, providing public computing capacity does not solve the fair-allocation problem. On the contrary, political preferences might a ect allocation decisions.

17

Model training data

With exponentially growing AI model size, and no diminishing returns in sight yet, maximum access to data is a necessary condition to maximise model performance. The good news is that, unlike hardware inputs, training data is non-rival and can be used by many at the same time. The bad news is that many sources of high-quality training data are subject to copyright and licensing fees. Strict copyright enforcement will increase the cost and shrink the supply of data. Model quality will decline, especially for smaller start-ups that cannot afford copyright licensing fees, and in smaller language zones where the supply of data is intrinsically limited. This reduces competition and innovation (Azoulay *et al*, 2024; Korinek and Vipra, 2024). It may also distort the geographical level playing field if copyright regimes differ between countries.

Competition authorities are aware of these potential negative effects. France's Autorité de la Concurrence (2024) recognises that provisions in the EU AI Act might disproportionately affect start-ups. It recommends collective licensing to reduce copyright licensing transaction costs, and price differentiation according to the value of the data. It is not clear how that would work in practice. In our view, competition and innovation would be best served with the elimination of the opt-out clause for AI training data from Article 4 in the EU Copyright Directive. If US courts would confirm the application of fair use and transformative use for AI model training data, that would have a similar impact. It would not affect traditional revenue channels for copyright holders, while they would benefit from productivity increases through better AI models.

3.2 Bottlenecks in downstream markets for model deployers and users

Co-opetition agreements between start-ups and GAMMANs point to potential competition bottlenecks in downstream AI markets on two levels: the market for models between AI start-up model developers and GAMMAN deployers, and the AI services market between GAMMAN deployers and end users. If start-ups directly deploy their models to end users, or via their own application stores, there is no competitive bottleneck because the market for start-ups is very competitive. Except for Nvidia, all GAMMANs have been designated as gatekeepers in their core platform services (CPS) for business users and end users under the EU DMA[18]. Most of these CPS already run on embedded AI models. The DMA is in principle technology-agnostic: relevant DMA obligations for these CPS apply, irrespective of the technology used to deliver the service.

Some obligations in the DMA's Article 6 are relevant for co-opetition relationships between independent AI model developers that deploy their services through established GAMMAN business models and intermediation platforms. For example, Microsoft Windows, applications from the EU market[19]. Google Search has been designated as a CPS. DMA Article 6(11) mandates sharing of search engine ranking, query, click and view data. Would that apply to queries and outputs of AI-driven services embedded in the search engine?

Some start-ups grow so fast that they may become DMA gatekeepers in their own right. For example, OpenAI ChatGPT is coming close to fulfilling the quantitative DMA threshold criteria to become a gatekeeper, and the ChatGPT app store is close to being a core platform service: over 100 million users, estimated capital value over $80 billion and hundreds of thousands of business users that develop specialised ChatGPT applications. If designated, would DMA app store obligations apply and would they be relevant to an AI model app store, as compared to a smartphone app store for which the obligations were designed? ChatGPT could be considered as the 'operating system' on which specialised AI model applications run. As long as access to the app store remains open, not subject to discrimination or preferential treatment of OpenAI's own apps, that should satisfy the DMA objective of contestable markets.

The Autorité de la Concurrence (2024) recommended that competition policymakers should pay particular attention to this developer-deployer market segment for AI models-as-a-service (MaaS). The UK CMA (2024) recommended that co-opetition agreements should not contain vertical restraints, exclusive deals or tying or bundling, either in upstream or downstream markets in the AI value chain. For these recommendations to work in practice, the technical layers of this intermediate developer-deployer market side, as well as the deployer-end user market side, should be unbundled. This is technologically complex and economically difficult. DMA authorities are currently trying to cope with the challenges of the designated CPS and are nowhere near to contemplating further unbundling of these CPS. Competition authorities can explore vertical restraints and exclusive deals but have no practical means to deal with them yet. That leaves a lot of regulatory uncertainty and investor hesitations hanging over the EU market for AI model services.

4 Conclusions

Exponentially growing model training costs, with no end in sight, are the biggest AI model market entry barrier, at least at the technology frontier. For firms with smaller models below the technology frontier, competition is likely to be intense, suggesting a more limited role for competition policy. AI start-ups that want to stay at the technology frontier need to sign co-opetition agreements with GAMMANs to overcome the training cost barrier. Competition authorities are looking at these deals with suspicion. They fear that co-opetition agreements may become Trojan horses for GAMMANs to exercise leverage over, and reduce competition from, AI start-ups. But several investigations by competition authorities have so far not produced any smoking guns.

These investigations have to some extent diverted attention away from the main market entry barrier – model training costs – and directed it to potential competition problems in different market segments in the AI supply chain, while suggesting ways to keep these segments open and contestable. The UK CMA (2024) offers guidance on rules of behaviour that it expects the GAMMANs to follow in their business dealings with AI start-ups. This includes

Concurrence (2024) proposed similar principles in its opinion.

ese principles are also in line with the obligations that the EU DMA imposes on a set of designated gatekeeper platforms that largely overlaps with the GAMMANs. However, they may be hard to implement in practice as they may require a considerable degree of technically complex and economically costly unbundling and untying of AI-driven services and the underlying AI models that power these services. is may do more harm than good to consumers and business users. Co-opetition agreements are necessary to enable AI start-ups to access computing infrastructure. ey come with privileged access for big tech to the latest AI models. Licensing fees for copyright-protected data further tighten the already scarce supply of a ordable data.

Model training infrastructure and AI processor chips are rival physical inputs that can only be used by one party at the time. It is hard to see how a fair quota or price-based market-clearing access mechanism could be put in place. Training datasets are non-rival inputs that can be used by many parties at the same time. Public data pooling could be a fair access mechanism (Azoulay *et al*

References

Autorité de la Concurrence (2024) 'Avis sur le fonctionnement concurrentiel du secteur de l'intelligence arti cielle générative', *Avis* 24-A-05, available at https://www.autoritedelaconcurrence.fr/sites/default/ les/2024-06/Diapo_IA_conf_presse_VF_BC.pdf

Azoulay, P., J. Krieger and A. Nagaraj (2024) 'Old Moats for New Models: Openness, Control, and Competition in Generative AI', *Working Paper* 32474, National Bureau of Economic Research, available at https://www.nber.org/papers/w32474

Brandenburger, A. and B. Nalebu (1996) *Co-opetition: A Revolution Mindset that Combines Competition and Cooperation*, Crown Publishers, UK

CMA (2023) *Proposed principles to guide competitive AI markets and protect consumers*, UK Competition and Markets Authority, available at https://www.gov.uk/government/publications/ai-foundation-models-initial-report

CMA (2024) 'CMA AI strategic update', UK Competition and Markets Authority, 29 April, available at https://www.gov.uk/government/publications/cma-ai-strategic-update/cma-ai-strategic-update

Gans, J. (2024) 'Market power in Arti cial Intelligence', *Working Paper* 32270, National Bureau of Economic Research, available at https://www.nber.org/papers/w32270

Graef, I. (2019) 'Rethinking the Essential Facilities Doctrine for the EU Digital Economy', *TILEC Discussion Paper* DP2019-028, available at https://research.tilburguniversity.edu/en/publications/rethinking-the-essential-facilities-doctrine-for-the-eu-digital-e

Korinek, A. and J. Vipra (2024) 'Concentrating Intelligence: Scaling Laws and Market Structure in Generative AI', mimeo, available at https://www.dropbox.com/scl/ /3vmu5q8js8bcgvujbc9bg/Economic-Policy-Draft-R2-Concentrating-Intelligence.pdf

Madiega T. and R. Ilnicki (2024) 'AI investment: EU and global indicators', *At a Glance Digital Issues in Focus*, European Parliamentary Research Service, available at https://www.europarl.europa.eu/RegData/etudes/ATAG/2024/760392/EPRS_ATA(2024)760392_EN.pdf

Martens, B. (2024) 'Economic arguments in favour of reducing copyright protections on Generative AI model inputs and outputs', *Working Paper* 09/2024, Bruegel, available at https://www.bruegel.org/working-paper/economic-arguments-favour-reducing-copyright-protection-generative-ai-inputs-and

Maslej, N., L. Fattorini, R. Perrault, V. Parli, A. Reuel, E. Brynjolfsson … J. Clark (2024) *e AI Index 2024 Annual Report*, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, available at https://aiindex.stanford.edu/report/

Schae er, R., B. Miranda and S. Koyejo (2023) 'Are Emergent Abilities of Large Language Models a Mirage?' 37th Conference on Neural Information Processing Systems (NeurIPS 2023), available at https://proceedings.neurips.cc/paper les/paper/2023/ le/adc98a266f45005c403b8311ca7e8bd7-Paper-Conference.pdf

Solaiman, I. (2023) ' e Gradient of Generative AI Release: Methods and Considerations', mimeo, available at https://arxiv.org/abs/2302.04844

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser and I. Polosukhin (2017) 'Attention Is All You Need', mimeo, available at https://arxiv.org/abs/1706.03762

Zhao, X., S. Ouyang, Z. Yu, M. Wu and L. Li (2022) 'Pretrained language models can be fully zero-shot learners', mimeo, available at https://arxiv.org/abs/2212.06950