Competition Between AI Foundation Models: Dynamics and Policy Recommendations

Authors:
**Thibault Schrepel &
Alex ‘Sandy’ Pentland**

Publication Date:
June 2023

**MIT Connection Science
Working Paper**

Competition Between AI Foundation Models: Dynamics and Policy Recommendations

Thibault Schrepel* & Alex 'Sandy' Pentland**

Amsterdam Law & Technology Institute
(ALTI) Working Paper 3-2023

MIT Connection Science
Working Paper 1-2023

Version: June 28, 2023

Abstract

Generative AI is set to become a critical technology for our modern economies. If we are currently experiencing a strong, dynamic competition between the underlying foundation models, legal institutions have an important role to play in ensuring that the spring of foundation models does not turn into a winter with an ecosystem frozen by a handful of players.

* Associate Professor at the Vrije Universiteit Amsterdam; Faculty Affiliate at Stanford University CodeX Center; Invited Professor at Sciences Po Paris & University of Paris 1 Panthéon-Sorbonne. Corresponding author: t.schrepel@vu.nl

** Professor of Media Arts and Sciences at the MIT; Director of the MIT Connection Science.

1. Introduction

Spring has finally arrived.[1] Recent advances in deep learning have given rise to foundation models that underpin an infinite number of generative AI applications.[2] The growth we are currently witnessing is exponential. OpenAI's ChatGPT reportedly reached 100 million users in just two months;[3] venture capitalists are increasing their investments in startups from $408 million in 2018 to $4.5 billion by 2022;[4] major technology companies such as Google, Facebook and Microsoft are multiplying new product announcements;[5] models such as Anthropic's Claude processes 100,000 tokens of text per minute (approximately 75,000 words) in May 2023, up from 9,000 tokens in March 2023.[6] Competition in the space is highly dynamic, despite initial fears that AI would be monopolized before it became mainstream.[7] But dynamic competition is not a given. The nature of foundation

[1] Nils J. Nilsson, *The Quest for Artificial Intelligence* (Cambridge University Press, 2009): 408-409 (exploring the reasons for the AI winter).

[2] Foundation models are versatile models capable of being customized for various downstream tasks, *see* Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein et al. "On the opportunities and risks of foundation models." *arXiv preprint arXiv:2108.07258* (2021) (popularizing the term "foundation model"); Jakob Mökander, Jonas Schuett, Hannah Rose Kirk and Luciano Floridi, "Auditing large language models: a three-layered approach." *AI Ethics* (2023). Large language models ("LLMs") are a subset of foundation models. One can define LLMs as models that can process large amounts of unstructured text and learn the relationships between words or parts of words (called tokens). Generative AI refers to a group of technologies that automatically generate content based on prompts, *see* Nanna Inie, Jeanette Falk, and Steve Tanimoto, "Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI," *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023.

[3] Krystal Hu, "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note," *Reuters*, February 2, 2023, https://perma.cc/Z4XU-AFBX.

[4] Kyle Wiggers, "VCs Continue to Pour Dollars into Generative AI," TechCrunch, March 28, 2023, https://perma.cc/8DBX-HLEZ.

[5] Meta AI, "Introducing Make-A-Video: An AI System That Generates Videos from Text," ai.facebook.com, September 29, 2022, https://perma.cc/M93E-4HZ4; Meta AI, "Greater Creative Control for AI Image Generation," ai.facebook.com, July 14, 2022, https://perma.cc/3LMP-FPD3; Katie Paul and Sheila Dang, "Facebook Owner Meta Announces Tests of Generative AI Ads Tool," *Reuters*, May 11, 2023, https://perma.cc/3XMG-9W57; Ivan Mehta, "Meta Wants to Use Generative AI to Create Ads," TechCrunch, April 5, 2023, https://perma.cc/UNL2-TN2X; Johanna Voolich Wright, "Announcing New Generative AI Experiences in Google Workspace," Google Workspace Blog, March 14, 2023, https://perma.cc/QP6F-2BQ9; David Pierce, "The AI Takeover of Google Search Starts Now," The Verge, May 10, 2023, https://perma.cc/5BNS-JCAY; Kyle Wiggers, "Hands on with Google's AI-Powered Music Generator," TechCrunch, May 11, 2023, https://perma.cc/Z26M-CVSN; Lisa Eadicicco, "Google Will Use AI to Rewrite Your Texts and Generate Android Wallpapers," CNET, May 10, 2023, https://perma.cc/S76D-ZM9Z.

[6] Anthropic PBC, "Introducing Claude," March 14, 2023, https://perma.cc/93RV-VDRZ; Anthropic PBC, "Introducing 100K Context Windows," May 11, 2023, https://perma.cc/A5A5-4CZS.

[7] *See* Bruno Lasserre and Andreas Mundt, "Competition Law and Big Data: The Enforcers' View," *Italian Antitrust Review*, no. 1 (2017); OECD, "Summary of Discussion of the Hearing on Big Data -

models, coupled with increasing returns, could rapidly consolidate the entire space around a handful of strategic, but not necessarily highly innovative players.

If one accepts the assertion that innovation and new technologies are "what separates us from the Middle Ages,"[8] then policymakers and regulators have an important role to play in ensuring that the field of generative AI remains innovation-intensive. Early estimates predict a productivity boost from generative AI, with low-skilled workers benefiting the most,[9] leading to a 7% increase in global GDP and an acceleration of innovation.[10] The economic impulse generated by foundation models means that more wealth is being created.[11] Foundation models should therefore be seen as a key infrastructure for our economies and our societies as a whole. The more dynamic and competitive the field remains, the more foundation models accelerate wealth creation.

Ensuring competitive dynamism in the space requires a clear understanding of the competitive forces at play. Foundation models are commonly observed by policymakers and social scientists at the species level (i.e., "foundation model" as a class), but these lenses fail to see the inherent diversity within the species. Worst of all, designing the same regulation for all foundation models leads to ineffective and harmful proposals. Against this background, we present a new taxonomy of foundation models (**2**). We explore the competitive dynamics between and within

Annex to the Summary Record of the 126th Meeting of the Competition Committee Held on 29-30 November 2016", April 26, 2017: 2; Jan Wolfe and Dave Michaels, "FTC Chair Lina Khan Vows to Protect Competition in AI Market," WSJ, March 27, 2023, https://perma.cc/8YMH-N7QQ; Bhaskar Chakravorti, "Big Tech's Stranglehold on Artificial Intelligence Must Be Regulated," Foreign Policy, August 11, 2021, https://perma.cc/J686-Y777; Vinod Iyengar, "Why AI Consolidation Will Create the Worst Monopoly in US History," TechCrunch, August 25, 2016, https://perma.cc/2HFR-9BZJ.

[8] W. Brian Arthur, *The Nature of Technology : What It Is and How It Evolves* (New York: Free Press, 2009): 10.

[9] Shakked Noy and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," *SSRN*, March 1, 2023 ("Exposure to ChatGPT increases job satisfaction"); Erik Brynjolfsson, Danielle Li, and Lindsey Raymond, "Generative AI at Work," *arXiv.org*, April 23, 2023 ("gains of over 30% for the least experienced workers")

[10] Bryce Elder, "Surrender Your Desk Job to the AI Productivity Miracle, Says Goldman Sachs," *Financial Times*, March 27, 2023, https://perma.cc/TX2W-DRUS (generative AI could raise global GDP by 7%); Martin Neil Baily Korinek Erik Brynjolfsson, and Anton, "Machines of Mind: The Case for an AI-Powered Productivity Boom," Brookings, May 10, 2023, https://perma.cc/6D8P-8BPY.

[11] Michael Chui et al., "The Economic Potential of Generative AI: The Next Productivity Frontier," McKinsey & Company, June 2023, https://perma.cc/MJE7-GAWN, ("the automation of individual work activities enabled by these technologies could provide the global economy with an annual productivity boost of 0.2 to 3.3 percent from 2023 to 2040 depending on the rate of automation"); Tyna Eloundou et al., "GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," March 27, 2023, https://arxiv.org/pdf/2303.10130.pdf..

different types of foundation models (**3**) and formulate effective policy recommendations aimed at maintaining a thriving ecosystem (**4**).

2. Proposing a Taxonomy

We distinguish between three types of foundation models: general public foundation models (**2.1**), ecosystem foundation models (**2.2**), and personal foundation models (**2.3**). As we shall see, the main difference between these foundation models relates to access and is inferred by the training data.

2.1. General Public Foundation Models

General public foundation models can be accessed by any Internet user. Depending on their training data, these foundation models can be of two types. They serve a general purpose when they are trained on a large variety of data with the aim of performing tasks in all possible domains (e.g., business decisions, cooking recipes, writing, etc.). ChatGPT and Google Bard are good examples of such general-purpose foundation models.[12] They are to foundation models what Google Search is to search. But general public foundation models can also be domain-specific, if they can perform tasks in defined domains such as specialized topic models (e.g., BloombergGPT[13]), specialized production models (e.g., GitHub copilot[14]), or highly specialized process enhancement models that aid the "invisible" everyday operations (e.g., routing phone calls, driving instructions, work-scheduling, etc.).

2.2. Ecosystem Foundation Models

Ecosystem foundation models are only accessible to specific user groups. These foundation models are typically fine-tuned, at least in part, on a dataset that is not available on the Internet.[15] Users can provide data that already pre-exists the

[12] "Domain-Specific LLMs | Technology Radar," Thoughtworks, April 26, 2023, https://perma.cc/T4TC-HLXH.

[13] Shijie Wu et al., "BloombergGPT: A Large Language Model for Finance," March 30, 2023, https://perma.cc/ZDU3-56SM.

[14] "GitHub Copilot · Your AI Pair Programmer," GitHub, https://perma.cc/88L9-5JJ9; Sara Verdi, "Inside GitHub: Working with the LLMs behind GitHub Copilot," The GitHub Blog, May 17, 2023, https://perma.cc/GY3U-W76C.

[15] *See* Sixing Yu, J. Pablo Muñoz, and Ali Jannesari, "Federated Foundation Models: Privacy-Preserving and Collaborative Learning for Large Models," arXiv.org, May 18, 2023; Suchin Gururangan et al., "Don't Stop Pretraining: Adapt Language Models to Domains and Tasks," *ArXiv* April 23, 2020.

training of foundation models, such as private financial information, meeting minutes, etc. Users can also provide data collected for the purpose of training foundation models. They can create or rely on shared databases and data cooperatives for this purpose.[16] Lastly, the model behind ecosystem foundation models can be built from scratch, but they typically rely on fine-tuned general public foundation models.[17]

An example of an ecosystem foundation model would be one trained on data from different companies in the same industry.[18] These companies would use the foundation model to uncover typical behaviors in given situations.[19] Ecosystem foundation models are less likely to hallucinate than general public foundation models because they are trained on smaller datasets and have fewer possible use cases, which means the output can be more easily be tested and controlled.

2.3. Personal Foundation Models

Personal foundation models are only accessible to one user, be it an individual, a company, a government, etc. Personal foundation models are typically pre-trained on large data sets and fine-tuned on the individual's private data.[20]

[16] Tobin South et al., "Secure Community Transformers: Private Pooled Data for LLMs," https://perma.cc/NR25-F65P; Thomas Hardjono and Alex Pentland, "Data Cooperatives: Towards a Foundation for Decentralized Personal Data Management," arXiv.org, May 21, 2019.

[17] Stratos Tsesmetzis, "A Comprehensive Guide to Fine-Tuning a GPT-3 Model," Bare Square, April 13, 2023, https://perma.cc/DM5H-M6KA; Uwais Iqbal, "From Knowledge Management to Intelligence Engineering -A Practical Approach to Building AI inside the Law-Firm Using Open-Source Large Language Models," in *Proceedings of the Third International Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023)* (CEUR Workshop Proceedings , 2023), https://perma.cc/KXD2-UUYN.

[18] E.g., Michael Moor et al., "Foundation Models for Generalist Medical Artificial Intelligence," *Nature* 616, no. 7956 (April 1, 2023): 259–65, https://perma.cc/YGX2-LPUB. (describing a foundation model trained on "data from imaging, electronic health records, laboratory results, genomics, graphs or medical text"; Gengchen Mai et al., "On the Opportunities and Challenges of Foundation Models for Geospatial Artificial Intelligence," In ACM, New York, USA, https://perma.cc/J9TR-8AFB (describing a foundation model trained on "text, images (e.g., remote sensing or street view images), trajectory data, knowledge graphs, and geospatial vector data (e.g., map layers from OpenStreetMap), all of which contain important geospatial information (e.g., geometric and semantic information)").

[19] Ecosystem foundation models trigger specific legal challenges. The sharing of business secrets in the database can lead companies to coordinate market behaviors. Companies could also infringe data protection laws if they share the personal information of their customers without their consent. On the subject of privacy, *see* Qiang Yang et al., "Federated Machine Learning," *ACM Transactions on Intelligent Systems and Technology* 10, no. 2 (February 28, 2019): 1–19.

[20] Hannah Rose Kirk et al., "Personalisation within Bounds: A Risk Taxonomy and Policy Framework for the Alignment of Large Language Models with Personalised Feedback," ArXiv (Cornell University), March 9, 2023; Personal AI, "Your True Personal AI," https://perma.cc/UT4P-2A86;

Personal foundation models experience slow adoption curves as standalone products. However, recent private initiatives such as BloombergGPT are emerging.[21] Microsoft and others are also starting to offer new products for businesses and individuals to combine OpenAI models with private data.[22]

3. Exploring Competitive Dynamics

Competition between foundation models is driven by several factors. We first outline the current competitive landscape of foundation models (**3.1**), before focusing on what will define competition in space (**3.2**): their design and ability to improve the learning curve.

3.1. Current Landscape

Foundation models are currently being developed and offered under different conditions, from closed source to open-source.[23] The proliferation of open-source foundation models – starting with Google BERT[24] – is giving rise to a thriving ecosystem, now led by Hugging Face[25] (BigScience[26] and BigCode[27]) and EleutherAI.[28] HuggingFace hosts an Open LLM Leaderboard listing dozens of open-access models.[29] Meta has also joined the open-source movement in recent weeks by releasing the weights of its LLaMA model.[30] Open-access foundation models, where

PrivateGPT, https://perma.cc/A7K4-8MSG; Kyle Wiggers, "LlamaIndex Adds Private Data to Large Language Models," TechCrunch, June 6, 2023, https://perma.cc/66C3-ZGHF.

[21] Shijie Wu et al., "BloombergGPT: A Large Language Model for Finance," March 30, 2023, https://perma.cc/ZDU3-56SM.

[22] Andy Beatman, "Introducing Azure OpenAI Service on Your Data in Public Preview," Tech Community MICROSOFT, June 19, 2023, https://perma.cc/KH76-6YTU.

[23] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," arXiv.org, February 5, 2023 ("We propose a framework to assess six levels of access to generative AI systems: fully closed; gradual or staged access; hosted access; cloud-based or API access; downloadable access; and fully open").

[24] Jacob Devlin and Ming-Wei Chang, "Open Sourcing BERT: State-of-The-Art Pre-Training for Natural Language Processing," Google AI Blog, November 2, 2018, https://perma.cc/6G5E-T63B.

[25] Hugging Face, "Hugging Face - on a Mission to Solve NLP, One Commit at a Time.," https://perma.cc/6VG6-BGSP.

[26] "BigScience Research Workshop," https://perma.cc/T6X7-9EYQ.

[27] BigCode "Open and Responsible Development and Use of LLMs for Code," https://perma.cc/2562-8SAA.

[28] EleutherAI, https://perma.cc/24LB-2ADX.

[29] "Open LLM Leaderboard - a Hugging Face Space by HuggingFaceH4," Hugging Face, https://perma.cc/84Y6-A7C6.

[30] Cade Metz and Mike Isaac, "In Battle over A.I., Meta Decides to Give Away Its Crown Jewels," *The New York Times*, May 18, 2023, https://perma.cc/5QAX-4KGA.

the company releases the API but not the model or training data, are also emerging. OpenAI is one such open-access foundation model.[31]

Should they continue to exert competitive pressure, the existence of these open-source and open-access models would make a notable difference compared to the early days of the Web2 giants' core services, such as search, social media, etc. The presence of more competitors from the outset means competition in terms of use.. There is also competition to fine-tune of each foundation model to differentiate itself from the others and thus increase its chances of survival. Looking ahead, the question is whether open-source solutions have a viable path to offer attractive features and continue to improve, or whether a handful of private companies is likely to take over.

3.2. Competitive Forces

i. Design

The initial quality of foundation models – i.e., when they are first made available – plays an important role in defining competition in the field. This is easy to understand: if a foundation model is clearly inferior to others, chances are that it will not survive. We therefore propose to study the variables that play a critical role in the creation of superior foundation models.

The **first** variable relates to the ability of the model to learn from a dataset. General public foundation models are trained on large datasets. In recent months, the size of these datasets has grown exponentially. OpenAI's GPT-2 was trained on 1.5 billion parameters; GPT-3 on 175 billion parameters; and GPT-4 seemingly even more (numbers not disclosed).[32] However, these models see diminishing returns to scale in the use of data, which means that large datasets are a necessary but not sufficient condition to achieve great results. This led Sam Altam, CEO of OpenAI, to observe that "we're at the end of the (...) [era of] giant models."[33]

[31] Irene Solaiman, "The Gradient of Generative AI Release: Methods and Considerations," arXiv.org, February 5, 2023. "While they can allow for more feedback than a closed system, because people outside the host organization can interact with the model, those outsiders have limited information and cannot robustly research the system by, for example, evaluating the training data or the model itself."

[32] Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," Wired, April 17, 2023, https://perma.cc/VZ8E-MGQE.

[33] Will Knight, "OpenAI's CEO Says the Age of Giant AI Models Is Already Over," Wired, April 17, 2023, https://perma.cc/VZ8E-MGQE.

Moreover, advances in computer science and analytics are making the amount of data less relevant every day. In recent months, important technological advances have allowed companies with small data sets to compete with larger ones.[34] These advances are lowering the cost of training basic models. We list some promising avenues:

- In April 2023, the Berkeley Artificial Intelligence Research Lab (BAIR) at UC Berkeley showed that 'small' models can rely on high quality data to compensate for a lack of quantity. The Lab released a 13 billion open model, built on top of Meta's LLaMA, that competes with ChatGPT in terms of quality of result by learning from high-quality datasets.[35]
- In February 2022, a team at DeepMind introduced "Retrieval-Enhanced Transformer."[36] Train language models compare what the machine writes in real time with existing databases, such as Wikipedia and other websites. The team claims its transformer matches the quality of models 25 times larger.[37]
- In June 2021, a team of researchers from Microsoft introduced "low-rank adaptation (LoRA)" which "freezes the pre-trained model weights and injects trainable rank decomposition matrices into each layer of the Transformer architecture."[38] The team claims LoRA can "reduce the number of trainable parameters by 10,000 times and the GPU memory requirement by 3 times."[39]
- In September 2020, a team of researchers from the University of Waterloo introduced "less than one-shot algorithms."[40] As they show, a machine can be trained to distinguish between two classes—say, cats and dogs—and later add tigers to the list of classes it can recognise, without being provided with images of tigers. The researchers have already used this method with hierarchical soft-label classification algorithms.

[34] For a first overview, *see* Thibault Schrepel, "Alternatives to Data Sharing," The Regulatory Review, February 21, 2022, https://perma.cc/NGH8-SN7T.

[35] Xinyang Geng et al., "Koala: A Dialogue Model for Academic Research," The Berkeley Artificial Intelligence Research Blog, April 3, 2023, https://perma.cc/9HUC-K9KC.

[36] Sebastian Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," in *International Conference on Machine Learning* (PMLR, 2022), 2206–40.

[37] Sebastian Borgeaud et al., "Improving Language Models by Retrieving from Trillions of Tokens," in *International Conference on Machine Learning* (PMLR, 2022), 2206–40.

[38] Edward Hu et al., "LoRA: Low-Rank Adaptation Of Large Language Models," June 17, 2021, https://perma.cc/L2E9-BXK7.

[39] Edward Hu et al., "LoRA: Low-Rank Adaptation Of Large Language Models," June 17, 2021, https://perma.cc/L2E9-BXK7.

[40] Ilia Sucholutsky and Matthias Schonlau, "`Less than One'-Shot Learning: Learning N Classes from M < N Samples," *Proceedings of the AAAI Conference on Artificial Intelligence* 35, no. 11 (May 18, 2021): 9739–46.

- In February 2020, a team of researchers introduced "dataset distillation."[41] Data distillation has given rise to new data processing techniques that allow computers to extract multiple features from a single data point.
- In April 2017, Google unveiled a Transformer that "can be trained significantly faster than architectures based on recurrent or convolutional layers" by "relying entirely on an attention mechanism to draw global dependencies between input and output."[42]
- Although the concept of synthetic data first appeared in the 1970s, it is now being used to train AI models.[43] Synthetic data is a technique used to generate new data points based on the patterns and characteristics of an existing dataset.[44] This technique can be used to compensate for a smaller dataset by creating additional data points that can be used to train machine learning models.

All in all, recent technical developments are increasing the importance of the efficiency of AI models, while proportionally decreasing the importance of ever-larger data sets. The ability to improve models to work with less data is at the center of attention.

But if access to ever-larger datasets is not a decisive competitive factor, access to *unique* data sets is critical. There are two reasons for this. First, access to unique datasets may be necessary to provide the specific answer that users of foundation models are looking for (e.g., users may want to know what are the most-read articles on a particular website). Second, these datasets may play a critical role in the overall training of foundation models. A company like Google can prohibit access to YouTube's video transcripts, comments, etc. This allows Google to train foundation models with data that no other company has, to understand trends, fashions, how

[41] Tongzhou Wang et al., "Dataset Distillation," arXiv.org, February 24, 2020.

[42] Ashish Vaswani et al., "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017, 30.

[43] Richard J. Chen et al., "Synthetic Data in Machine Learning for Medicine and Healthcare," *Nature Biomedical Engineering* 5, no. 6 (June 2021): 493–97, https://perma.cc/HK4J-TJKP; Sergey I Nikolenko, *Synthetic Data for Deep Learning, Springer Optimization and Its Applications* (Springer International Publishing, 2021).

[44] Ian Goodfellow et al., "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, vol. 27 (Curran Associates, Inc., 2014), 2672–80; Connor Shorten and Taghi M. Khoshgoftaar, "A Survey on Image Data Augmentation for Deep Learning," *Journal of Big Data* 6, no. 1 (July 6, 2019); - Antreas Antoniou, Amos Storkey, and Harrison Edwards, "Data Augmentation Generative Adversarial Networks," *ArXiv:1711.04340*, March 21, 2018; Sergey I Nikolenko, *Synthetic Data for Deep Learning, Springer Optimization and Its Applications* (Springer International Publishing, 2021).

users talk online, how they interact, what kind of videos they watch and react to depending on time of day, gender, identity, and so on. This information can help answer not only video-related questions, but all questions related to culture. In short, companies with an existing ecosystem to exclusively train their model on have a significant advantage.

The **second** variable that (currently) determines the level of concentration in the industry is the cost of training and running models. OpenAI spent $540 million on the development of GPT-4 in 2022 alone (training costs, salaries, etc.).[45] It reportedly cost Google $20 million to train the Pathways Language Model (PaLM) from scratch.[46] Even fine-tuning a model with access to a pre-trained model is expensive. EleutherAI trained its larger model using the GPT-3 dataset. It took the company three and a half months and cost $400,000.[47] Moreover, foundation models are expensive to run. OpenAI reportedly spends $700,000 per day running ChatGPT (mostly to compute all the prompts).[48] Operating costs "far exceed training costs."[49]

That being said, companies – including chip makers – are competing to lower these costs. Nvidia, whose stocks surged 200% in June 2023 since, the beginning of the year,[50] claims that using its GPU cut the price of training LLMs from $10 million down to just $400,000.[51] There are also newly-available collections of model compression and algorithms that make designing and running foundation AI an order an magnitue cheaper.[52] In June 2023, for example, researchers from the UC Berkeley Sky Computing introduced vLLM which deploys an attention algorithm

[45] Erin Woo and Amir Efrati , "OpenAI's Losses Doubled to $540 Million as It Developed ChatGPT," The Information, May 4, 2023, https://perma.cc/GUE5-2UK4.

[46] AI Now Institute, "ChatGPT and More: Large Scale AI Models Entrench Big Tech Power," April 11, 2023, https://perma.cc/A82W-3A78.

[47] Will Douglas Heaven, "The Open-Source AI Boom Is Built on Big Tech's Handouts. How Long Will It Last?," MIT Technology Review, May 12, 2023, https://perma.cc/2SPD-2UTY.

[48] Aaron Mok, "ChatGPT Could Cost over $700,000 per Day to Operate. Microsoft Is Reportedly Trying to Make It Cheaper.," Business Insider, April 20, 2023, https://perma.cc/NY9H-2CCA.

[49] Aaron Mok, "ChatGPT Could Cost over $700,000 per Day to Operate. Microsoft Is Reportedly Trying to Make It Cheaper.," Business Insider, April 20, 2023, https://perma.cc/NY9H-2CCA.

[50] Rachel Pupazzoni, "Nvidia's share price has surged almost 200 per cent on an AI boom. But will it stay there?," ABC News, June 26, 2023, https://perma.cc/4BMG-2A4X.

[51] Urian B., "NVIDIA Announces $9.6M Drop in Cost When Using Its GPUs for AI LLM Training," Tech Times, May 29, 2023, https://perma.cc/M8NR-K9B5; Usman Pirzada, "NVIDIA: Reduce The Cost Of CPU-Training An LLM From $10 Million To Just $400,000 USD By Buying Our GPUs," WCCFTech, May 28, 2023, https://perma.cc/6RB6-QLXZ.

[52] Prakhar Ganesh et al., "Compressing Large-Scale Transformer-Based Models: A Case Study on BERT," *Transactions of the Association for Computational Linguistics* (2021) 9: 1061–1080; Victor Kolev et al., "Combining Improvements in the Compression of Large Language Models," *Stanford CS224N Custom Project* (2022).

called "PagedAttention" in order to manage attention keys and values. The team claims vLLM "delivers up to 24x higher throughput than HuggingFace Transformers" and enabled them "to cut the number of GPUs used for serving the above traffic by 50%."[53] New approaches are also developed to train 'good enough' LLMs (i.e. LLMs that do not produce the same results as other LLMs whose training is cash intensive, but whose results are good enough for most users[54]) on a single GPU in just a couple of hours, or even smartphones.[55]

All in all, small projects currently rely on companies with deep pockets to get free access to their new models, or their ability to raise capital to train foundation AI models.[56] However, new advances may reduce these costs and thus the need to rely on third parties. On this basis, we cannot currently conclude whether costs will contribute to market concentration in the future.

ii. Learning Curve

The dynamics between foundation models are defined not only by their design (*see* i), but also by their ability to grow the user base. The more users they attract, the better the training, which improves the quality of the fine-tuning, attracts new users, increases the capacity to afford expensive training, and so on. In other words, foundation models are subject to positive feedback loops (i.e., increasing returns).[57]

[53] Woosuk Kwon et al., "vLLM: Easy, Fast, and Cheap LLM Serving with PagedAttention," June 20, 2023, https://perma.cc/D2AJ-RSAU; Khushboo Gupta, Meet vLLM: An Open-Source LLM Inference And Serving Library That Accelerates HuggingFace Transformers By 24x, Marktechpost, June 24, 2023, https://perma.cc/893Z-2V7N.

[54] *See* "Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality," March 30, 2023, https://perma.cc/P57B-V9MW.

[55] Hugging Face, "Efficient Training on a Single GPU," https://perma.cc/FKJ9-TVWQ; Jonas Geiping et al., "Cramming: Training A Language Model On A Single Gpu In One Day," *Arvix Preprint*, December 28, 2022; Philipp Schmid et al., "Train a Large Language Model on a single Amazon SageMaker GPU with Hugging Face and LoRA," AWS Machine Learning Blog, June 5, 2023 https://perma.cc/3ET4-H4JJ; ColossalAI, GitHub, https://perma.cc/X5NP-ZWR4; Alpaca-lora, Github, https://perma.cc/S4WN-WLT4; Edward Hu et al., "LoRA: Low-Rank Adaptation Of Large Language Models," June 17, 2021, https://perma.cc/L2E9-BXK7. The multimodal ScienceQA SOTA was trained in an hour, *see* Renrui Zhang et al., "LLaMA-Adapter: Efficient Fine-tuning of Language Models with Zero-init Attention," arXiv:2303.16199v2, June 14, 2023.

[56] Will Douglas Heaven, "The Open-Source AI Boom Is Built on Big Tech's Handouts. How Long Will It Last?," MIT Technology Review, May 12, 2023, https://perma.cc/8BN8-EXEV. For example, a new foundation model startup called Mistral AI raised $113 million four weeks after launch, *see* Ingrid Lunden, "France's Mistral AI Blows in with a $113M Seed Round at a $260M Valuation to Take on OpenAI," TechCrunch, June 13, 2023, https://perma.cc/MW4U-UQ2U.

[57] W. Brian Arthur, "Increasing Returns and the New World of Business," Harvard Business Review, July 1, 1996: 100-109; W. Brian Arthur, "Competing Technologies, Increasing Returns, and Lock-in by Historical Events," *The Economic Journal* 99, no. 394 (March 1989): 116.

As a result, foundation models appear to compete *for* the market rather than competing *in* the market. However, a closer look reveals that the strength of the feedback loops differs depending on the type of foundation model.

	Strength of the returns	**Limits**
General public foundation models	Significant increasing returns	The 10th million users improve the fine-tuning less than the first[58]
Ecosystem foundation models	Moderate increasing returns	Increasing returns limited to the industry: the model is not easily transferable to another industry
Personal foundation models	Small increasing returns	The model cannot be perfectly adjusted by other users, what matters most is the fine tuning

When it first comes to general public foundation models, returns tend to increase rapidly for several reasons. First, the more a model is used, the more it can compute user inputs. This dynamic makes the model better over time, which can attract new users and thus increase the learning curve (defined as the relationship between the number of users and the ability to improve the service by learning from them)[59]. But the learning curve will flatten over time, knowing that the 10th million user will improve the model proportionally less than the first user. Second, the more users they have, the more general public foundation model providers can generate revenue and pay for access to exclusive databases. Knowing that several companies such as Reddit, Stack Overflow, Twitter and others have started licensing access to their database for the purpose of training foundation models, one can expect large foundation model players to pay high fees and integrate them.[60] Small players will

[58] OpenAI, "GPT-4 Technical Report," *ArXiv:2303.08774*, March 15, 2023.

[59] Hal R Varian, "Artificial Intelligence, Economics, and Industrial Organization," in *The Economics of Artificial Intelligence: An Agenda*, ed. Ajay Agrawal, Joshua Gans, and Avi Goldfarb, 2019, 399–422. ("There is a concept that is circulating among lawyers and regulators called "data network effects". The model is that a firm with more customers can collect more data and use this data to improve its product. This is often true---the prospect of improving operations is that makes ML attractive---but it is hardly novel. And it is certainly not a network effect! This is essentially a supply-side effect known as 'learning by doing,' also known as the 'experience curve' or 'learning curve'.")

[60] KeyserSosa, "An Update Regarding Reddit's API," Reddit, April 18, 2023, https://perma.cc/48CC-HT8M; Staff, "News/Media Alliance AI Principles," News/Media Alliance, April 20, 2023,

not be able to get the same access, which means that big players will have a competitive advantage, attract more users, etc. Third, the more users a foundation model has, the greater its reputation and the opportunity to partner with user-facing products. For example, one could imagine a search engine – let us call it Bing – partnering with a widely used foundation model – let us call it ChatGPT. Fourth, the more users, the more money the model can generate, i.e., the more the company can advertise and acquire new users. Here, the ability of large tech companies to push their foundation models to billions of users should translate into important increasing returns. In general, companies with an existing user base will have an advantage if they can add a foundation model to their existing products. These companies are well-positioned to meet the distribution challenge that comes with the scalability of foundation models.

When it comes to ecosystem foundation models, increasing returns are limited to each use case. A model fine-tuned to help judges write court decisions cannot be used to help sporting goods companies with commercial strategies. Compared to general public foundation models, ecosystem foundation models are even more dependent on the quality and exclusivity of the data on which they are trained. But within each industry or use case, the more users, the better the model, which increases the incentive to use the model. A company that already provides a key service to an industry will be well positioned to push a foundation model to its users and benefit from positive feedback loops. Overall, we should expect more ecosystem foundation models to survive than general public foundation models (knowing that use case specificity will drive demand without easily transferable models across the space), with dominant foundation models per use case likely to emerge.

Personal foundation models benefit from relatively smaller increasing returns. Companies that already have a strong reputation – such as Apple – and access to online users – such as Google –will benefit from an early advantage. However, the models underlying personal foundation models cannot be well tuned for other users. Knowing that fine-tuning to each user remains key to their relevance and user experience, increasing returns remain smaller than they are for other types of foundation models. Moreover, individuals produce large collections data with clear copyrights attached, e.g., usage of services, personal data stored locally, etc. Personal foundation models can thus be easily trained to assist each individual in a space where quality matters. One might expect strong competition in this space, with relatively low barriers to entry. Given that the personal development industry

https://perma.cc/D98J-WNQG; Maria Diaz, "Stack Overflow Joins Reddit and Twitter in Charging AI Companies for Training Data," ZDNET, April 21, 2023, https://perma.cc/4BSC-GPYJ.

generated $41.81 billion in 2021 in the United States alone,[61] we see a strong incentive for new players to effectively enter the market and compete to provide individuals with tailored advice on health, work, financial decisions, learning experiences, leisure activities, etc. That might explain why Google, in a leaked memo, described "scalable Personal AI" as a major reason why the company is not positioned to win the foundation AI model race.[62]

iii. Expected dynamics

When it comes to competitive dynamics within each type of foundation model, the higher the increasing returns, the more likely the foundation model is to be dominated by a handful of firms. In the presence of high increasing returns, the quality of foundation models design remains central to the ability to retain users, but competition is not based on quality alone: history matters. Random events and initial competitive advantages could well translate into sustained dominance. There is a multiplicity of possible outcomes.

If our analysis is correct, the larger players in the general public foundation models space will initially improve their foundation models faster than smaller competitors thanks to positive feedback loops. They will acquire dominance that way. That being said, the feedback loops from which they benefit will increase less rapidly over time. Should they sustain dominance, their market position will not necessarily correlate with superior foundation models, as smaller players will also be able to benefit from sufficient increasing returns to achieve similar quality. The ability of large players to lock-in the ecosystem will need to be closely monitored, along with other variables such as network effects among developers, consumer inertia, dynamic capabilities, etc.[63] Conversely, the more limited the increasing returns, the less robust market shares will be. A company or open-source project with a significantly better model will be able to break the initial cycle of feedback loops in the short to medium term and regain competitive advantage. The competitive dynamics in the personal foundation model space are therefore likely to remain intense over time.

[61] https://www.grandviewresearch.com/industry-analysis/personal-development-market

[62] Maya Posch, "Leaked Internal Google Document Claims Open Source Ai Will Outcompete Google And OpenAI," Hackaday, May 5, 2023, https://perma.cc/YH5F-ZXE3; Dylan Patel and Afzal Ahmad, "Google 'We Have No Moat, And Neither Does OpenAI'," Semi Analysis, May 4, 2023 https://perma.cc/FE3V-2MUS.

[63] Aaron Holmes and Jon Victor, "OpenAI Considers Creating an App Store for AI Software," The Information, June 20, 2023, https://perma.cc/J9Z9-2PQZ (OpenAI is reportedly considering the creation of an app store. The creation of an app store *could* lead to a network effect that will make OpenAI's market position more robust. The failure of OpenAI to successfully deploy its API).

When it comes to competition between models, although they serve different purposes and are likely to coexist, we see a competitive dynamic to attract investment and achieve scalability. Venture capitalists are currently investing in the general purpose models such as ChatGPT and Bing. Investments in specialized and process improvement models such as BloombergGPT and GitHub Copilot are now emerging. The winner-take-all effect of these general purpose LLMs should quickly attract more investment. We anticipate a longer lead time for ecosystem and individual foundation models to attract investment and scale up. There are two reasons for this. First, lower increasing returns mean that investors have less hope of capturing the market (i.e. betting on the winning horse). Second, these types of foundation AI models require a change in behavior on the part of users, e.g., to provide private data, be willing to rely on their input, etc. In short, we think these two types of foundation AI models will attract investment away from general public foundation AI in the not-too-distant future.

4. Adapting Public Policy

For the foundation model ecosystem to remain competitive, public policy must first and foremost support innovation among foundation models providers. This requires (**4.1**) a clear vision of what regulation is intended to achieve in the space, (**4.2**) the creation of specific rules, and (**4.3**) enforcement measures.

4.1 "Innovation First"

We suggest that policymakers follow an "innovation first" principle. Looking at the period from 1995 to 2013, the OECD estimates that "different components of innovation together often account for at least 50% of economic growth."[64] Moreover, "long-term trends suggest that innovation, productivity and job creation can go hand in hand."[65] Innovation is responsible for creating new opportunities for businesses and individuals, increasing productivity and competitiveness, and improving the overall quality of life for citizens.

In the foundation model ecosystem, innovation is the main driver of competiton.[66] Firms do not compete to make their foundation models slightly more efficient, but

[64] OECD, *The Innovation Imperative* (OECD Publishing Paris, 2015): 19.

[65] OECD, *The Innovation Imperative* (OECD Publishing Paris, 2015): 24.

[66] On that idea that innovation drives competition in highly dynamic ecosystems, *see* Damanpour, "Organizational innovation: A Meta-Analysis of Effects of Determinants and Moderators", Academy of Management Journal, Vol. 34, No. 3 (1991), 555-590 (innovation is positively related to firm

they compete to disrupt others through critical innovation. In this context, we propose an "innovation first" principle. The principle is not only pro-innovation,[67] but also implies that innovation should be given reasonable priority in the face of trade-offs. To be clear, we are not saying that innovation should never be restricted, but we are saying that policymakers and enforcers should only restrict innovation to address existing and documented risks. In other words, they should reject precautionary approaches and elevate the protection of innovation as a fundamental objective that can only be trumped by fundamental rights.

4.2. Regulatory and Policy Agenda

Implementing an "innovation first" principle requires not only a clear vision of what regulation should achieve, but also a concrete policy agenda. We set out several actionable points.

First, we recommend that new rules and standards in the space should only be enacted after the publication of an impact assessment documenting whether they lead to monopoly power.[68] The space should remain as permissionless as possible: rules and standards should not – unless strictly necessary – raise compliance costs to levels that small and medium-sized players cannot reasonably afford, force licensing, create unnecessary barriers to data access, etc. The first calls for regulation of generative AI are coming from the big players in the space, who may already be showing a desire to raise barriers to entry by increasing compliance costs.[69]

performance); Yang, Li, and Li, "Mechanism of Innovation and Standardization Driving Company Competitiveness in the Digital Economy" Journal of Business Economics and Management, Vol. 24, No. 1 (2023), 54-73 (the level of innovation and standardization of a company drives its competitiveness); Geroski, "Innovation as an Engine of Competition" in Mueller, Haid and Weigand (Ed.), Competition, Efficiency, and Welfare (Springer, 1991), pp. 13-26; Jorde and Teece, "Antitrust Policy and Innovation: Taking Account of Performance Competition and Competitor Cooperation", Journal of Institutional and Theoretical Economics, Vol. 147, No. 1 (1991), pp. 118-144; Organisation for Economic Co-operation and Development (2015), "The Innovation Imperative: Contributing to Productivity, Growth and Well-Being" (calling innovation a "key driver of economic growth and development"); Petit and Teece, Innovating Big Tech firms and competition policy: favoring dynamic over static competition, Industrial and Corporate Change, Vol. 30, No. 5 (2021), pp. 1168–1198.

[67] Which the UK defines as "enabling rather than stifling responsible innovation," *see* UK Government, "Pro-innovation Regulation of Technologies Review: Digital Technologies" (March 2023): 21, https://perma.cc/57T7-KUCP.

[68] Shikhar Singla, "Regulatory Costs and Market Power," Social Science Research Network (Rochester, NY, February 23, 2023).

[69] Matt O'Brien, "ChatGPT Chief Says Artificial Intelligence Should Be Regulated by a US or Global Agency," Alton Telegraph, May 16, 2023, https://perma.cc/ST6E-H3CL; Sam Altman, Greg Brockman, and Ilya Sutskever, "Governance of Superintelligence," openai.com, May 22, 2023, https://perma.cc/8JNA-ZRWL.

One way of ensuring that the regulatory burden is proportionately distributed is to focus regulation on large players and/or to impose heavy obligations on these players. The use of thresholds to define large players, as in the EU's Digital Services Directive, is a satisfactory solution. In the absence of such thresholds, the burden of regulatory compliance tends to fall disproportionately on small players, as has been observed with the GDPR.[70] Specifically, labeling all generative AI applications as "high risk" in the EU's AI Act,[71] regardless of the number of users, would penalize small companies that do not have the capacity to comply with data governance requirements, create an always-updated technical documentation, store all logs for long periods of time, create accessible instructions for using AI, have a dedicated employee to oversee the AI system, etc.[72]

To take just one example, Article 11 of the current draft of the EU AI Act requires companies of all sizes to create and publish a technical documentation, written in such a way as to demonstrate that their high-risk AI system is in line with the "values, fundamental rights and principle[s]" of the European Union.[73] Writing such a document would, at the very least, require technical and legal expertise that startups with a handful of employees do not have. If startups were to allocate the task of writing – and keeping up to date – the technical documentation to a minimum of two employees, some would be spending a large part of their human capital on this for a product that only a few users actually use. They would end up losing the innovation battle for lack of remaining resources. The same goes for imposing similar obligations without calling generative AI "high risk." The current draft of the EU AI Act requires that companies providing foundational models "demonstrate

[70] Chinchih Chen, Carl Benedikt Frey, and Giorgio Presidente. *Privacy regulation and firm performance: Estimating the GDPR effect globally*. No. 2022-1. The Oxford Martin Working Paper Series on Technological and Economic Change, 2022 ("Firms exposed to the GDP experienced an 8% decline in profits, and the decline in profits of small companies is almost double the average"); Rebecca Janssen, Reinhold Kesler, Michael E. Kummer, and Joel Waldfogel, *GDPR and The Lost Generation of Innovative Apps*. No. w30028. National Bureau of Economic Research, 2022 ("GDPR induced the exit of about a third of available apps; and in the quarters following implementation, entry of new apps fell by half"); Garrett A. Johnson, Scott K. Shriver, and Samuel G. Goldberg, "Privacy and market concentration: intended and unintended consequences of the GDPR." *Management Science* (2023) ("GDPR increased digital markets concentration").

[71] "There have been calls from outside and inside the Parliament for a ban or classifying ChatGPT as high-risk," MEP Svenja Hahn: Martin Coulter and Supantha Mukherjee, "Exclusive: Behind EU Lawmakers' Challenge to Rein in ChatGPT and Generative AI," *Reuters*, May 1, 2023, https://perma.cc/AJ5B-P9YN.

[72] For a view of these requirements, *see* Articles 9 to 15, European Union Artificial Intelligence Act, Draft European Commission, EUR-Lex - 52021PC0206 - EN, 2021, https://perma.cc/MP8U-V9FK.

[73] Page 1, European Union Artificial Intelligence Act, Draft European Commission, EUR-Lex - 52021PC0206 - EN, 2021, https://perma.cc/MP8U-V9FK.

through appropriate design, testing and analysis that the identification, the reduction and mitigation of reasonably foreseeable risks to health, safety, fundamental rights, the environment and democracy and the rule of law prior and throughout development."[74] Companies must also produce "extensive technical documentation and intelligible instructions for use."[75] These time-consuming obligations favor large companies that can afford to comply without impacting their innovative capacities.

Second, impact assessments should be conducted and regulations issued by regulators with specific AI expertise. We do not recommend the creation of a stand-alone AI regulator, as such a regulator can be more easily captured than multiple regulators working together on AI.[76] However, we do recommend the creation of an (informal) council with members from different regulatory agencies to coordinate their AI policies. Such a council would help avoid situations where different rules and standards contradict each other, for example, by forcing a database to be "representative" while privacy rules restrict the use of personal data that can help representativeness.[77] The UK Digital Regulation Cooperation Forum ("DRCF") that brings together the Information Commissioner's Office ("ICO"), the Competition and Markets Authority ("CMA"), the Office of Communications ("Ofcom") and the Financial Conduct Authority ("FCA") is a good example of an effective council between regulatory agencies.[78]

Third, we recommend the creation of exemptions from antitrust rules for the purpose to accelerate R&D by open source and open access companies in the

[74] Article 28b(2)(a), DRAFT Compromise Amendments, Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, KMB/DA/AS 9 May 2023, https://perma.cc/E5TH-KE24.

[75] Article 28b(2)(e), DRAFT Compromise Amendments, Proposal for a regulation of the European Parliament and of the Council on harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union Legislative Acts, KMB/DA/AS 9 May 2023,https://perma.cc/E5TH-KE24.

[76] Big companies with lobbying capacities tend to favor the creation of a single AI regulator, *see* for example Jeremy Kahn, "Microsoft Joins Calls for a New A.I. Regulator," Fortune, May 25, 2023, https://perma.cc/AKR9-2HHM; Gregory Schmidt, "A.I. Needs an International Watchdog, ChatGPT Creators Say," *The New York Times*, May 24, 2023, sec. Business, https://perma.cc/8CNG-PYGX.

[77] *See* Article 10 para 3 and 5, European Union Artificial Intelligence Act, Draft European Commission, EUR-Lex - 52021PC0206 - EN, 2021, https://perma.cc/MP8U-V9FK.

[78] For a critical analysis of the DRCF, *see* Aysem Diker Vanberg, "Coordinating Digital Regulation in the UK: Is the Digital Regulation Cooperation Forum (DRCF) up to the Task?," *International Review of Law, Computers & Technology* 37, no. 2 (March 27, 2023): 1-19, https://perma.cc/5VGY-5HA5; also, Philip Schlesinger, "The Neo-Regulation of Internet Platforms in the United Kingdom," *Policy & Internet* 14, no. 1 (March 3, 2022): 47-62.

ecosystem. Specifically, we suggest that open source and open access companies should be able to pool their resources in joint ventures without having to notify the operation to antitrust agencies.[79] We also recommend that strategic alliances of open source and open access companies should be allowed to form strategic alliances without risking cartel sanctions.[80] If open source and open access players were able to share up-front costs, marketing networks, and technical knowledge, they would be in a stronger position to compete with proprietary systems that capture more of the value they create.[81] We therefore propose to extend the scope of EU R&D exemption to antitrust law and to create a similar exemption in other jurisdictions. Currently, only R&D agreements between companies that together do not have more than 25% of the market can be exempted if "there are three or more competing R&D efforts in addition to and comparable with those of the parties to the R&D agreement."[82] We propose to remove the market share and competing efforts criteria for open source and open access companies.

In addition, and specifically with regard to personal foundation models, we recommend that small and medium-sized enterprises should be able to benefit from the R&D exemption when they pool resources and data for the purpose of training foundation models without fear of antitrust enforcement. Given that the design of personal foundation models is the main driver of competition in this area, allowing small and medium-sized enterprises to share costs will enable them to compete with larger enterprises that can afford to train and operate personal foundation models.

Fourth, we propose that policymakers and enforcers ensure a level playing field for all players in the general public foundation model ecosystem. Specifically, we suggest that foundation models be allowed to train on data that is publicly available but is proprietary and personal in nature. Japan has already announced that it will

[79] The creation of joint ventures performing on a lasting basis all the functions of an autonomous economic entity is currently considered a notifiable merger in European Competition Law, *see* Article 2, the European Council Merger Regulation, No 139/2004 of 20 January 2004, https://perma.cc/HY6J-HZ4J. Similarly, the formation of a for-profit joint venture may be subject to the HSR Act in the United States if it involves an acquisition of non-corporate interests or voting securities, *see* 16 CFR § 801.40 2005, https://perma.cc/A5WD-8W9P.

[80] Currently, "alliances" between competitors can be considered a collusion under EU and US antitrust law if it leads to coordinating market strategies.

[81] Arthur, "Positive Feedbacks in the Economy," Scientific American 262, no. 2 (1990): 92-99.

[82] *See* 6(2) and 6(3) DRAFT Revised Regulation on the application of Article 101(3) of the Treaty on the Functioning of the European Union to certain categories of research and development agreements, https://perma.cc/4AE6-P8G6.

follow this path.[83] As we have discussed, foundation models with the most users get preferential access to exclusive databases whose owners are willing to license in order to extract value. This means that if foundation models were only allowed to train freely on non-proprietary data, the rule would favor big players with the capacity to pay gigantic licensing fees for access to proprietary data. The same is true for non-personal data as small companies tend to have fewer personal data than large companies. We therefore recommend an exception to data ownership and privacy laws for the training of foundation models.

Fifth, "innovation first" requires complex adaptive regulations ("CAR") that respond to the effects they create – that is, regulations that adapt to continuously protect innovation. CAR requires defining what metrics policymakers should consider assessing whether innovation is being protected or not.[84] CAR then calls for the implementation of sensors, i.e., tools to scrape the necessary metrics. Finally, CAR calls for the publication of thresholds to which regulation will react – whether to make it more or less stringent – in order to ensure legal certainty. For example, policymakers may want to protect copyright holders by forcing generative AI to cite sources.[85] In such a scenario, policymakers would be asked to specify what protection they want to give to copyright holders. If they want to maintain traffic to their websites, policymakers will have to monitor traffic and say what level of traffic, after what period of time, would be considered a success. Policymakers would publish the data they collect so that third parties can analyze it. If the scheme had not achieved its objective after the specified period, policymakers would repeal or strengthen the law according to published principles.[86] The UK announced the outlines of a CAR in its white paper, "A pro-innovation approach to AI regulation," in which it expressed its willingness to monitor and evaluate the "cross-economy and sector-specific impacts of the new regime" by collecting "appropriate data (...) from relevant

[83] Jose Antonio Lanz, "AI Art Wars: Japan Says AI Model Training Doesn't Violate Copyright," Decrypt, June 5, 2023, https://perma.cc/335P-44PD.

[84] Sandy Pentland and Robert Mahari, "Legal Dynamism," Network Law Review, September 27, 2022, https://perma.cc/KW47-LE2W.

[85] Explainability often comes at the expense of AI accuracy, *see* Andrew J. Bell et al., "It's Just Not That Simple: An Empirical Study of the Accuracy-Explainability Trade-off in Machine Learning for Public Policy," in *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, (2022); Hani Hagras, "Toward Human-Understandable, Explainable AI," IEEE *Computer* 51, no. 9 (September 2018): 28-36; Usama, M., Butt, F. K., Muhammad Usama et al., "Unsupervised Machine Learning for Networking: Techniques, Applications and Research Challenges," *IEEE Access* 7 (2019): 65579-615; Anne-Marie Nussberger et al., "Public Attitudes Value Interpretability but Prioritize Accuracy in Artificial Intelligence," *Nature Communications* 13, no. 1 (October 3, 2022), 1-12.

[86] Should the collected data be sensible, policymakers could also mandate a trusted their party to analyze the data they have collected.

sources" and "improve the regime" on that basis.[87] This is a step in the right direction, but we believe that the mechanism should be made more transparent and automatic.

4.3. Enforcement Actions

Enforcement is the final piece of an effective "innovation first" policy. We suggest that enforcers focus primarily on general public foundation models.[88] As we discussed earlier (*see* 3.2), general public foundation models benefit from high increasing returns. This means that illegal behaviors are more likely to trigger a snowball effect and lead to an unfair market position than is the case with ecosystem and personal foundation models. In particular, behaviors that reduce the ability of competitors to benefit from increasing returns are likely to be the most damaging. These behaviors should be closely monitored and severely punished.

In practice, increasing returns are driven by user growth, which means that all firms in the space must have fair access to users. Anticompetitive strategies that restrict competitors' access to users, reduce compatibility between models, or degrade the performance of other (OA) models, contribute to locking the ecosystem into the dominant general public foundation model by reducing user input and thus the learning curve. We therefore recommend that (antitrust law) enforcers focus on detecting practices that restrict access to users, such as discrimination (i.e., fair access to platforms and aggregators), contractual exclusivity (i.e., from providers not offering competing foundation models), tying (i.e,. incentives not to switch between foundation models), and so on. These practices do indeed reduce uncertainty and hence competition.[89]

Conversely, we do not recommend that enforcers have a major focus of their efforts on (exploitative or exclusionary) practices that affect existing users. Nor do we recommend that enforcers spend a large part of their resources on proactively detecting price-related practices, as competition in this area is driven by innovation. And finally, we do not recommend that agencies concentrate on practices affecting access to big data as foundation models compete on the margins thanks to

[87] UK Government, "Pro-innovation Regulation of Technologies Review: Digital Technologies" (March 2023): 43, https://perma.cc/57T7-KUCP.

[88] Competitive dynamics in ecosystem and personal foundation models are better addressed by policy actions such as described in 4.2 such as allowing small and medium companies to share costs, pool resources, etc.

[89] Nicolas Petit and Thibault Schrepel, "Complexity-Minded Antitrust," *Journal of Evolutionary Economics*, February 24, 2023, https://perma.cc/TE6S-W5SP.

innovative design and fine-tuning. Instead, we recommend that {regulators focus on "real time" regular auditing of outcomes in order to have early detection of harms, and allow them learn what sorts of harms require the most regulation and enforcement.

Detecting practices can be challenging in a world where reactive methods – i.e., enforcement agencies waiting for complaints – are ineffective because users cannot easily detect violations. Investigations can also be complicated by the fact that some major foundation model players do not document or annotate their training data.[90] However, automated audit trails that continuously monitor system outputs and alert when they exceed pre-defined tolerance ranges can help agencies to be proactive.[91] The automation of these audits helps to connect enforcers and technology. They enable near real-time enforcement. We recommend that such audit trails be mandated by law for major players.

Where potential infringements are identified, we recommend that regulators and courts severely punish harm to innovation as a stand-alone practice. We also recommend that these agencies and courts consider excusing potentially anticompetitive practices when they benefit innovation in the sector.[92] These last two recommendations follow from the fact that innovation drives competition between foundation models. As far as remedies are concerned, we do not recommend the imposition of data sharing and data portability (*see* 3.2.i) but, rather, remedies to ensure access to users.

5. Conclusive Words and Steps Ahead

Generative AI is experiencing strong and dynamic competition, which seems to favor an open ecosystem rather than a proprietary one. At first glance, existing foundation models players are not and will not be able to live the "quiet life" of

[90] Melissa Heikkilä, "OpenAI's Hunger for Data Is Coming back to Bite It," MIT Technology Review, April 19, 2023, https://perma.cc/FE2N-2WD4; Nithya Sambasivan et al., "'Everyone Wants to Do the Model Work, Not the Data Work': Data Cascades in High-Stakes AI," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, 1–15.

[91] Agencies can implement automatic audit trails themselves if they have the expertise. They can also use second or third-parties to implement them. On the subject of audits, *see* AI Now Institute, "Algorithmic Accountability: Moving beyond Audits," AI Now Institute, April 11, 2023, https://perma.cc/LYS3-DH98.

[92] Thibault Schrepel, "A Systematic Content Analysis of Innovation in European Competition Law," Social Science Research Network (April 9, 2023), https://perma.cc/6H8G-MK7C.

monopolists.[93] However, a closer look reveals that not all foundation models experience similar competitive dynamics. Ecosystem foundation models, and even more so general public foundation models, benefit from strong increasing returns. As history teaches us, this type of increasing returns can lead to robust market dominance and opportunities for abuse.

Enforcers have an important role to play in protecting the competitive dynamics where increasing returns are highest. They can do this by protecting innovation and mitigating the effects of anti-competitive practices designed to freeze the ecosystem. This will require AI expertise.[94] We recommend that they build capacity in this area without delay.

Policymakers also need technical expertise. This expertise will enable them to ensure that AI regulation does not stifle dynamism. We are already seeing calls for regulation from big players and governments that want to lead the regulatory space. Policymakers must resist the calls of AI doomsayers and dominant market players, and instead address the issues at hand while prioritizing innovation in the space. We recommend the use of publicly documented regulatory experiments. Such experiments from different regions of the world will allow for comparisons and counterfactuals. If researchers systematically track these experiments and review their impact, and if policymakers learn from their findings, AI regulation will be on its way to further improving the public good.

*

*　　*

[93] J. R. Hicks, "Annual Survey of Economic Theory: The Theory of Monopoly," *Econometrica* 3, no. 1 (January 1935): 1.

[94] *See* Stanford Law School, "Computational Antitrust," Stanford Law School, https://perma.cc/65JA-ZPZV.