Artificial intelligence and competition policy[1]

Andrei Hagiu[2] and Julian Wright[3]

December 2024

Abstract

This paper examines competition policy implications of the rapidly expanding Artificial Intelligence (AI) sector. We analyze the vertical AI technology stack and data feedback loops to address three key questions: the potential for market concentration in core AI services, AI's likely impact on existing market structures, and emerging competition policy challenges. We identify key risks to competition in the AI sector, ways in which AI may disrupt some existing platforms, how AI could lead to new types of gatekeepers, and some novel competition policy concerns raised by AI.

1. Introduction

The Artificial Intelligence (AI) sector has experienced unprecedented growth in recent years, epitomized by the launch of OpenAI's ChatGPT in late 2022 (Bick et al., 2024). This breakthrough has sparked a proliferation of generative AI offerings, ranging from co-pilots that draft, debug and optimize computer code to AI-powered virtual companions that offer emotional support and conversation. The surging demand for GPU chips, crucial for training and operating new AI foundation models such as OpenAI's GPT series, Google's Gemini, and Anthropic's Claude, has propelled Nvidia to become one of the three most valuable companies in the world in 2024.

Competition authorities have been paying attention. In part this reflects the substantial involvement of the big-tech companies (particularly, Google and Microsoft) in the AI sector. The United States' Justice Department and Federal Trade Commission,

[1] This paper is based on the keynote talk Julian Wright gave at CRESSE 2024 titled "Artificial intelligence, data and competition policy". We have benefited from comments from participants of CRESSE 2024, as well as separate comments from Alexandre de Cornière, Simon Loertscher, Tat-How Teh, Frank Verboven, David Yoffie, and two anonymous referees.

[2] Boston University Questrom School of Business; ahagiu@bu.edu.

[3] National University of Singapore; jwright@nus.edu.sg.

the European Commission, and the United Kingdom's Competition and Markets Authority have all launched antitrust investigations into the major AI companies and their business practices.[4]

In this paper, we discuss the merits of these concerns by addressing three questions:

1) Will core AI services become dominated by a few firms, and if so, will these be the existing big-tech companies?
2) How will AI affect the market structure for existing sectors? In particular, will AI enhance incumbents' market power, or will it disrupt them by helping new entrants?
3) What "traditional" competition policy issues are relevant to AI and what novel competition policy issues does AI raise?

We do not attempt to provide definitive answers to these questions. Rather, we lay out what we see as the key economic factors that will determine the answers and speculate on how things are most likely to unfold. For the impatient reader, the conclusion contains some key take-aways.

It is worth pointing out a few caveats before proceeding further. First, we will limit ourselves to competition policy questions, thus avoiding the wider debate about whether new regulations are needed to safeguard society from AI developments.[5] The premise of this paper is that innovation and progress in AI services is a good thing, and therefore preserving competition is important. We note, not all scholars agree with this premise.[6]

Second, we will largely abstract from the technical details of how AI works, keeping things at a high level in order to focus on the fundamental issues that are of interest to economists. In particular, we will distinguish AI model improvements based on data generated by customers from those based on publicly available or acquired data. This matters because the former can give rise to data feedback loops, whereas the latter does not. However, we will not discuss in detail the different ways in which data is converted into improvements in the output of a model. Thus, whether the data is used to fine-tune the model or to improve its inferences, and whether this is done continuously or episodically via discrete upgrades, are technical distinctions which matter less for our purposes than whether the data comes from customer usage or has been acquired from external sources.

Third, our discussion applies to all types of AI services and not just those arising in the current generative AI wave. That said, many of our examples are related to generative AI given that it is the most prominent and fastest growing form of AI today.

---

[4] See https://www.nytimes.com/2024/06/05/technology/nvidia-microsoft-openai-antitrust-doj-ftc.html, https://www.reuters.com/technology/eu-seeks-views-microsoft-openai-google-samsung-deals-eus-vestager-says-2024-06-28/, https://time.com/7012813/sarah-cardell/.

[5] For a recent survey on this topic, see Comunale and Manera (2024). Gans (2024) covers the economics of AI more broadly, including a wider range of policy issues in the context of AI.

[6] For example, see Acemoglu (2024).

2. Key concepts

Before addressing our three questions, it is useful to review some key concepts in the context of AI. First, we outline the current landscape of core AI services, and the key players involved. In the second subsection, we discuss in depth the economics of data feedback loops, which play a key role in our subsequent analysis of the AI sector.

2.1. Core AI services

In Figure 1, we lay out a simplified version of the vertical AI stack, noting the key services and prominent examples of providers at each layer.[7] The figure is meant to be illustrative and to include some of the most important current examples, in order to give a sense of the breadth of technologies and applications. It is by no means comprehensive.

AI applications are built on a stack that begins (at the bottom) with the specialized hardware used to train and run AI foundation models and culminates in the models themselves. Collectively, we refer to these different components as core AI services.[8]

At present, the hardware layer of this stack is dominated by Graphics Processing Units (GPUs) provided by Nvidia. GPUs are specialized chips that were originally designed to handle the graphical rendering tasks required for video games and other visual applications. However, due to their ability to process many operations in parallel, GPUs have become an essential hardware component for most AI applications. In 2024, Nvidia is reported to hold a market share in excess of 90% of the GPUs used in training AI models (Qi, 2024). Nvidia's business model involves designing and selling these highly specialized chips, while contracting their manufacture to leading semiconductor foundries, primarily, Taiwan Semiconductor Manufacturing Company (TSMC) and to a lesser extent Samsung Electronics.

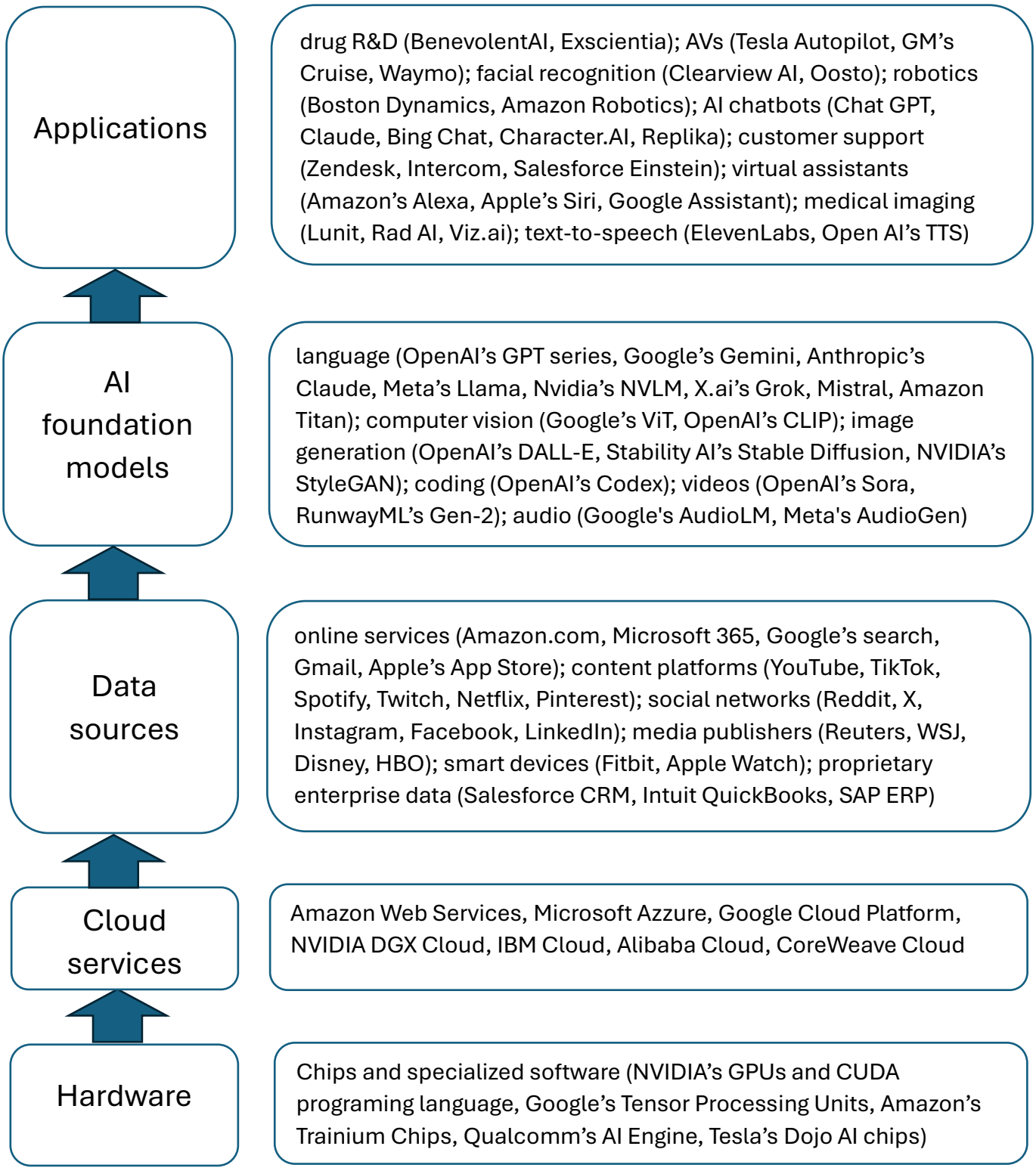
Beyond hardware, Nvidia has solidified its position in the AI industry by providing the most popular software framework for GPU utilization, known as Compute Unified Device Architecture (CUDA). CUDA enables developers to harness the parallel processing power of GPUs for general-purpose computing, which is critical for AI and machine learning applications. It has become the de facto standard. As a result, CUDA creates operating system-like network effects around Nvidia's chips, i.e. more developers writing software for CUDA makes NVIDIA chips more attractive to buyers, which in turn attracts more developers. Overall, this creates significant barriers-to-entry for Nvidia in the AI chip

[7] The September 2023 and April 2024 reports of the CMA titled "AI Foundation Models: Initial Report" and "AI Foundation Models: Technical update" contain a more detailed discussion of the vertical stack in AI, the various technologies and providers involved, and the interlinkages between the different players at each level, as well as potential competition concerns that the CMA has identified. See also the 2024 report of the Italian Competition Authority (AGCM) titled "Competition in the Artificial Intelligence Tech Stack".

[8] As we discuss in more detail at the end of this subsection, the line between applications and foundation models is sometimes blurry, with some applications involving fairly general models which can be applied to different AI applications by other businesses. As a result, some applications could be considered as part of core AI services.

market. There are of course many companies trying to disrupt Nvidia's dominance by building alternative chips for powering AI applications – we will discuss this in Section 3.2.

Figure 1. The Vertical Stack in AI



It is important to recognize that the hardware layer also features other dominant firms that provide key services and inputs. TSMC, as the dominant manufacturer of AI chips, itself has a strong market position and negotiating power against Nvidia and other chip designers. Furthermore, TSMC itself relies on ASML, the only supplier of extreme ultraviolet (EUV) lithography machines that TSMC needs to produce high-performance chips. The reason we do not include TSMC or ASML in our vertical stack, despite their dominance, is that the cross-layer vertical integration and leveraging strategies we discuss later in the paper stop at the chip design layer. None of the big-tech firms or foundation model providers in Figure 1 actually manufacture AI chips or EUV lithography machines. Moreover, unlike Nvidia, TSMC and ASML do not develop software for training, coding or running AI models and applications, nor have they manifested any aspirations to move up the stack. They do, however, have a clear interest in making sure the layers above them (starting with chip design) are not monopolized.

At the next level up the AI stack, we find companies that offer cloud computing services which are crucial for AI development and deployment.[9] This layer is dominated by the big three "hyperscalers", Amazon Web Services, Microsoft Azure and Google Cloud Platform, but it also features other competitors, including Nvidia's DGX Cloud, IBM Cloud, Alibaba Cloud and CoreWeave Cloud (the latter is a startup).

In the context of AI, hyperscalers provide a range of services including:

- Infrastructure-as-a-Service (IaaS): scalable computing resources based on GPUs for training and running large AI models.
- Platform-as-a-Service (PaaS): ready-to-use tools and environments for working with AI, including data preparation, model training, and deployment.
- Software-as-a-Service (SaaS): pre-trained AI models and programming interfaces that businesses can integrate into their applications without needing to develop models from scratch.

These services cater to firms ranging from AI-focused companies developing foundation models to traditional businesses looking to incorporate AI capabilities into their operations. The hyperscalers' offerings enable organizations to access high-performance AI computing without the need for significant upfront investment in hardware and software infrastructure. To power these services, hyperscalers need vast quantities of AI chips, costing many billions of dollars.

Going up one more level, we have data. To train a model, one needs to combine both large amounts of data and compute. Foundation models are trained on both publicly available and private data sources. Public sources include C4, The Pile, Project Gutenberg Corpus, LAION, Internet Archive, GitHub, and Stack Exchange.[10] These

[9] See Biglaiser et al. (2024) for a much more detailed survey and analysis of the cloud sector.

[10] C4 (Colossal Clean Crawled Corpus) is a massive web-crawled dataset (a cleaned version of a larger Common Crawl dataset) used for language models. The Pile is a large-scale curated dataset containing books, academic papers, and web content. Project Gutenberg Corpus offers a collection of free e-books, primarily classic literature. LAION datasets provide large-scale image-text pairs for multimodal AI training, Internet Archive (archive.org) is a digital library of websites, books, audio, video, and other digital formats,

datasets cover a wide range of text, from web content and academic writing to classic literature and image captions, providing diverse training material for AI models.

But increasingly, AI models are trained on private data, and we've listed some of the larger potential data providers in Figure 1. Here all five big-tech companies are included, since each has their own significant sources of potential training data. For example, Microsoft has Microsoft 365, Bing, and LinkedIn; Google has YouTube, Google search, Gmail, Maps, Play Store, Google Assistant and its Workspace Suite; Meta has Instagram, Facebook, WhatsApp; Apple has the App Store, Apple News, Apple Music, Apple TV+, Apple Watch and Siri; and Amazon has Amazon.com, Prime Video, Twitch and Alexa. Aside from these, other platforms (as well as most enterprises) hold valuable alternative sources of data, some of which are identified in Figure 1.

Next is the foundation model layer. AI foundation models are large-scale, pre-trained models designed to serve as the base for a wide variety of applications (as illustrated in Figure 1): AI chatbots, computer vision, image and video generation, autonomous driving, protein structure prediction, etc. These models are trained on vast amounts of data and leverage advanced machine learning architectures, such as transformers, to learn generalized representations of language, images, video, sound, objects, etc. Once trained, foundation models can be fine-tuned on specific datasets for specialized applications.

To provide a sense of scale, recent estimates suggest that training a state-of-the-art large language foundation model (like OpenAI's GPT-4, Google's Gemini, and Anthropic's Claude 3.5) can require hundreds of petaflop-days of computing power[11] and datasets comprising hundreds of billions of tokens, equivalent to millions of books worth of text.

Foundation models can be open source or closed source. For instance, Meta's Llama, Nvidia's NVLM, Mistral and X.ai's Grok are open-source large language models, meaning other firms are (relatively) free to use and adapt these foundation models.[12] By contrast, Anthropic's Claude, OpenAI's GPT and Google's Gemini are closed source – the primary way to build on top of them is via paid access through the APIs their providers expose.

It is worth nothing that some companies have developed multiple foundation models, each for a different "modality" (text, images, video, code): for instance, OpenAI has its GPT series, DALL-E, CLIP, Codex and Sora; Google has Gemini, ViT and AudioLM. And some foundation models are expanding to become "multi-modal": for instance,

---

including historical web content. GitHub hosts programming code repositories. Stack Exchange is a network of Q&A websites on various topics.

[11] A petaflop-day of compute is roughly equivalent to the number of operations a high-end laptop could perform in 100,000 days (274 years) of continuous operation, or put differently, 86.4 quintillion operations.

[12] Open source in an AI context usually means that the model weights and architecture are publicly available, allowing other firms to access and adapt them for custom applications and models. This doesn't necessarily imply that the training code or data are also openly available. Moreover, there may still be restrictions on which firms can adapt their models and what they can do with them.

OpenAI's GPT-4V, Nvidia's NVLM-D-72B and Microsoft's LLaVA are designed to understand and generate both text and images.

Finally, at the top layer (above the core AI services) are the applications that make use of these foundation models. This covers a wide range of applications as shown in Figure 1: customer support chat bots, autonomous driving, robots, facial recognition, virtual assistants, drug discovery and development, etc. All five big-tech companies are active at this level. Microsoft offers AI-powered Copilot across its suite of products, including Windows, Office, and GitHub; Google has integrated its Gemini model into its conversational AI services and search results (via AI Overviews); Meta has introduced AI-powered creative tools on Facebook and Instagram (such as AI-generated stickers and image editing features) as well as an AI-based answer engine across Facebook, Instagram, Messenger, and WhatsApp; Amazon recently rolled out an AI companion on its marketplace, Rufus; Apple's own AI application, known as Apple Intelligence, was introduced in late 2024 and is gradually being integrated into various Apple devices and service.

Other foundation model providers offer AI chat interfaces for consumers, like OpenAI's ChatGPT and Anthropic's Claude. These applications illustrate how foundation models are being fine-tuned and integrated into existing software ecosystems to enhance productivity, creativity, and decision-making across diverse industries. The applications are provided by both AI-focused technology companies (Google, Microsoft, OpenAI, Anthropic) and by more traditional businesses (e.g., Salesforce and Workday have combined their data to train an AI Employee Service Agent).

Finally, we note that some foundation models are naturally conducive to more use cases and therefore likely to spawn more applications than others. For example, there are likely going to be many more applications built on language foundation models than on ChemBERTa, a foundation model for chemistry and cheminformatics. Conversely, applications are sometimes based on multiple foundation models: for instance, Perplexity utilizes multiple foundation models, including those provided by OpenAI and Anthropic. And applications are sometimes built on specialized foundation models, which are themselves built on more general foundation models. For example, AstroLLaMA is a specialized foundation model for astronomy, built on LLama-2. So in this case, AstroLLaMA acts as a sort of middleware between a general purpose language model and astronomy-specific applications. This also means the line between applications and foundation models is sometimes blurry.

2.2. Data feedback loops

A key factor that will determine whether or not AI foundation models and applications will feature winner-take-all or winner-take-most outcomes is the strength of the relevant data feedback loops. Data feedback loops arise when more customer data leads (through AI) to a better product, which attracts more customers, and so more customer data, and so on.

Drawing analogies from the role of network effects in driving the increasing dominance of platforms run by big-tech companies, academics and policymakers often assume such data feedback loops will be strong for AI foundation models and applications. Operating systems (Android, iOS, Windows), social networks (Facebook, Instagram, YouTube), and online marketplaces (Amazon.com) all benefit from strong traditional network effects. The more users they have, the more value the platforms can offer to "suppliers" (developers, content creators, sellers) who want to access these users, and vice-versa. It is then tempting to assume similar forces would work with respect to AI-based products and services: more users would mean more data, and so via AI, a higher value product, which would attract ever more users.

In contrast, many venture capitalists and practitioners claim data feedback loops are weak and overstated, providing much less of a competitive advantage moat than traditional network effects.[13]

In our view, the reality is more nuanced and requires understanding the specific nature of the data in question and the feedback signals that can be obtained from users. Consider Google search and Google maps. Both have strong defensible positions, which in our view are due in large part to positive data feedback loops. The more people that search on Google and click on the links provided, the more data Google gathers, which allows its algorithms to provide more accurate and relevant search results (organic as well as sponsored), attracting even more users and searches, and so on. Likewise, the more drivers rely on Google Maps for up-to-date traffic conditions and route selection, the more data Google collects on traffic conditions, and the better its route predictions will be, again leading to a self-reinforcing cycle.

As a contrasting example, consider Fitbit, a pioneer in the wearables and fitness tracking space since 2007: it has amassed an extensive dataset from millions of users. One might have expected that its first-mover advantage should have provided Fitbit with a significant competitive advantage through positive data feedback loops. However, this does not seem to have materialized. Numerous competitors including Apple, Garmin, Samsung, and Whoop have successfully entered and thrived in the same market. We will explain why this may have been later in this section, but the Fitbit example clearly challenges the inevitability of data-driven market dominance.

Even where data feedback loops exist, they do not always provide a lasting advantage. Consider the example of Grammarly, a well-known cloud-based writing assistant service that leverages user interactions to continuously refine its recommendations. When users accept or reject Grammarly's suggestions for spelling, grammar, tone, style, and word choice, this feedback is incorporated into the system's machine learning algorithms. Together with Grammarly's own grammar experts who review contentious cases, this creates a positive data feedback loop: as more users interact with the service, its accuracy and relevance improve, potentially attracting more

[13] This view is well captured by this venture capitalist's blog: https://www.nfx.com/post/truth-about-data-network-effects.

users and further enhancing the product. This virtuous cycle appeared to cement Grammarly's dominance in the writing assistance market for over a decade. However, the landscape shifted dramatically in late 2022 with the emergence of advanced AI language models. Various AI chatbots have emerged that can offer comparable writing assistance capabilities, often integrated into broader AI products. Grammarly's leadership position in the market has suddenly become a lot less clear cut.

Data feedback loops have the *potential* to lead to market dominance, but there are a number of reasons why we think this is likely to be the exception rather than the rule.[14] Since these reasons inform our subsequent discussion of what is likely to happen in the markets for core AI services, we lay them out in some detail here.

First, many AI services use publicly available or acquired data sets for training their models, but do not actually generate meaningful data feedback loops from interactions with their customers. For example, AI transcription and translation services rely mostly on training data that is readily available from public sources. There is very limited improvement of their models based on what they learn from serving their customers. It is not surprising then that this is a very competitive market. Furthermore, the services offered by the various providers can now be replicated by most large-language models.

Second, even if data feedback loops are at play, data is often not unique. That is, there are often multiple sources or providers of similar data that can be used to train AI models to achieve comparable outcomes. This is why Grammarly's customer data no longer gives it such a clear competitive advantage relative to recent chatbots built on large language models that are trained on vast amounts of publicly available data.

Companies offering AI-powered radiology further illustrate the point. An AI radiology company which has partnered with many hospitals may have access to billions of images and hundreds of millions of corresponding radiology reports to train and improve its models. However, given the many thousands of hospitals offering radiology services worldwide, it is unrealistic that any single company will be able to secure enough exclusive data to prevent rivals from training competing models. Thus, even though each such AI company can improve its models over time via data obtained from user feedback[15], no single entity can monopolize the market solely through this mechanism. It is thus perhaps not surprising that there are many well-funded AI radiology startups currently competing in this space, including Rad AI, Nuance Communications, Subtle Medical, DeepTek.ai, Aidoc, Viz.ai, Arterys, and Behold.ai.

Third, beyond the nature of the available data, one needs to assess whether the learning curve keeps increasing or plateaus well before all available data is exhausted. Some argue that in the case of big data and AI, the latter situation is more common. Obviously, this is an empirical question, and the answer will be application specific. For

[14] Some of the factors discussed here are based on Hagiu and Wright (2020).

[15] As AI is increasingly used to draft reports, radiologists who sign off on them correct any errors, thereby providing valuable feedback to further refine the AI models.

instance, smart thermostats typically require only limited user feedback to achieve effective personalization.

One paper frequently cited for demonstrating the limits of data-driven learning is Bajari et al. (2019). They examine the accuracy of Amazon's product demand forecasts and find that increasing the number of retail products within a category does not improve forecast performance. While forecasts for individual products improve over time as more data on each product accumulates, these gains show diminishing returns. However, the study's relevance to assessing the strength of data feedback loops, even for Amazon, is questionable. A more pertinent study would examine the extent to which Amazon can enhance its product recommendations ("Recommended for You", "Frequently Bought Together", and "Customers Who Bought This Item Also Bought") or optimize the order of its product listings to improve conversions as it amasses more data on consumer responses to these recommendations and listings.

Similarly, Carballa-Smichowski et al. (2023) find that in the context of health and health-related data, increasing the number of predictor variables improves prediction accuracy, but with decreasing returns past a certain point.

Other recent empirical works (Klein et al., 2023, Schaefer and Sapi, 2023, and Allcott et al., 2024) have focused on online search services and found evidence consistent with positive feedback loops holding in that application. This may reflect the importance of edge cases in search. Most search queries are somewhat unique, and the key differentiator between an adequate and an excellent search engine lies in its ability to provide useful results for less common queries. To effectively handle these edge cases, a search engine requires feedback from a vast user base. Furthermore, the relevance of search results often changes over time, necessitating continuous usage data. The combination of edge cases with the need for constantly renewed data creates strong data feedback loops. In our view, it has been a key factor driving the winner-takes-most nature of the online search market.

Fourth, feedback loops are often limited by weak feedback signals. Fitbit historically collected vast amount of user data, but most of it did not provide any useful feedback signal. Consider Fitbit's fitness readiness score based on sleep, activity, and heart rate variability. This metric purportedly guided users in selecting appropriate workout intensities: low readiness (1-29) suggested prioritizing rest and recover; good readiness (30-64) indicated suitability for moderate exercise with caution; excellent readiness (65-100) implied preparedness for higher-intensity workouts. However, this system largely relied on comparing observed user data (heart rate and its variability) to pre-programmed values and patterns. There was little scope for the system to improve its accuracy and recommendations based on observed user signals. This is mainly because of Fitbit's inability to observe whether users really adhered to its recommendations, and whether the recommendations led to superior outcomes compared to alternative approaches. Fitbit's recent integration with performance-tracking hardware devices through partnerships with the likes of Peloton bikes and Tonal weights has the potential to create

more accurate feedback, but it remains to be seen whether the data derived from these integrations will be sufficient to create a meaningful data feedback loop.

Feedback signals can also be weak because it takes too long to receive the feedback. This is particularly evident in credit risk models and venture capital investments. The prolonged time it takes to generate meaningful feedback on investment outcomes (bankruptcies in the case of credit and exits in the case of venture investments) may render the information obsolete in a changing business environment.

Of course, there are things firms can do to strengthen feedback signals from users. Firms can (re)design their product in such a way that customers, in the normal course of using the product, create data that signals how useful or effective the product is.

Consider the example of AI chatbots. Most AI chatbots employ simple thumbs up or thumbs down mechanisms. A more sophisticated approach might include tracking whether users copy responses (a noisy measure of user utility) or allowing users to save and categorize helpful responses into favorite folders (a more reliable utility measure). Where the AI is unsure of which type of answer the user is looking for, they could give users multiple starting points of possible answers and ask them to select which one they want the full answer for. In the case of writing, even more powerful feedback can be engineered by integrating into the user's word processing software, so the AI can track the final version of the changes the user adopts (Tucker et al., 2024 argue this can be a particularly effective way to improve an LLM's performance).

Additional strategies firms could employ include[16]:

- asking users to rate responses in a way that makes the benefits of providing honest feedback clear to users, so as to create incentive compatible feedback (e.g., make it clear the rating will be used by the AI tool to better personalize future responses for the user);
- employing humans in the loop to provide additional high-quality signals for those cases where user feedback is ambiguous, although this is expensive and less scalable.

A final factor that affects the competitive dynamics implied by data feedback loops is the nature of the learning involved. In Hagiu and Wright (2023a) we introduced and compared two fundamentally different types of learning that can drive data feedback loops: (i) across-user learning and (ii) within-user learning. These are illustrated in Figure 2.

Across-user learning arises when more users generate more data, which enables AI to improve the product for *all* users. Examples include Google Maps, Grammarly, AI-powered radiology, and autonomous vehicles. In each case, as user numbers grow, the product improves for everyone. In contrast, within-user learning involves more usage by a given user, enabling AI to improve the product specifically for that user. Smart devices like Google Nest exemplify within-user learning, where repeated use allows the device to

[16] We discuss these and other strategies in more detail in Hagiu and Wright (2023b).

learn individual preferences, increasing the likelihood of continued and expanded usage. Indeed, some AI personal assistants (e.g., You.com) claim to adapt to a user's specific needs and preferences over time.

Figure 2

<u>Across-user</u> learning

Google maps, autonomous cars, Grammarly, AI-radiology

<u>Within-user</u> learning

Google Nest, personal assistants (Rewind, Apple intelligence), AI for enterprise solutions

The distinction between these two types of learning matters because of the different economics associated with the associated feedback loops. Feedback loops based on across-user learning create data network effects, meaning that the value of the product increases as more users join, through the additional insights gained from additional user data. Thus, firms leveraging across-user learning are more likely to experience winner-takes-all dynamics, all else equal. In contrast, feedback loops based on within-user learning exhibit compounding switching costs. As a user continues to use a product, it becomes increasingly customized to their needs, making them less inclined to switch to a competitor. Thus, firms benefiting from within-user learning tend to obtain greater lock-in of their existing customer base but remain vulnerable to competition for new customers so there is less reason to expect winner-take-all dynamics. As the relevant markets mature (with few new consumers coming into the market), such firms may enjoy substantial market power even if the market is not particularly concentrated.

Feedback loops that combine both across-user and within-user learning are likely to be the most powerful in creating defensible positions. Arguably Google search exhibits both features. In addition to learning what most users click on in response to different

queries, Google can also personalize search results to some extent based on a user's previous search history and clicks, and their location and language settings.[17] Similarly, recommender engines from Amazon, Instagram, Netflix, Spotify, TikTok, and YouTube utilize both types of learning. They leverage correlation data from many users' experiences across various items to predict the likelihood that a user will like a particular item based on their idiosyncratic interactions with other items. This is why such recommendation services likely benefit from powerful feedback loops.

3. Will core AI services become dominated by a few firms?

In this section we address how the market structure of AI core services is likely to evolve. We first focus on foundation models, and after doing so, we discuss this question more generally, taking into account possible vertical integration concerns across all the layers.

3.1. Foundation models

Will foundation models become commoditized, or will they become dominated by one or two players? This is the trillion-dollar question facing investors and policymakers right now.

To address this question, it is helpful to draw a parallel to online search. Is the current situation with foundation models likely to mirror what happened in online search during the late 1990s, where numerous providers initially competed (Yahoo! Search, AltaVista, Ask Jeeves, Lycos, Excite, Infoseek, etc), but ultimately Google search ended up dominating? Currently, many foundation models are competing (see Figure 1). Will history repeat itself, with one clear winner emerging from the crowd?

Several key differences suggest that the outcome is unlikely to mirror that of internet search. Internet search is a relatively narrow use case, whereas foundation models cover a vast array of disparate applications (language processing, content creation, coding, data analysis, idea generation, marketing, customer service, tutoring and training, and so on), as well as different modalities (text/language, images, audio/speech, video, robotics, or various combinations of these). While many foundation models are general-purpose, ultimately, one may expect specialized foundation models to develop for fields such as astronomy, chemistry, education, finance, genomics, law, mathematics, medical imaging, and meteorology.[18] Compared to online search, it seems much less likely one or two providers can serve all these very diverse applications and modalities effectively. Different user interfaces, business models and distribution channels could also serve as points of differentiation across which multiple different providers could coexist and compete. This still leaves open the possibility that within some of these fields, a small

[17] Yoganarasimhan (2019) shows that personalization of search results does indeed improve user click through rates.

[18] Such specialized foundation models have already started to appear. Examples include AstroLLaMA for astronomy and ProGen for protein engineering and generation.

number of more specialized foundation models will dominate, a point we will return to at the end of this section.

Another significant difference relative to online search is the strength of the data feedback loops for foundation models. As explained in Section 2.2, Google search likely benefits from a strong data feedback loop. This is less evident in the case of foundation models. While data feedback loops may be present at the level of some AI applications if their providers can utilize unique customer data and engineer the application to have automatic user feedback loops similar to online search, it is unclear how those loops would extend to the underlying foundation models. Many applications may simply not share their user data back to the foundation model they are built upon due to data privacy or strategic concerns. For example, most companies that use Chat GPT Enterprise to build their own AI chat bots do not feed their user data back into Open AI's GPT models (something that OpenAI explicitly commits to in its contracts). And when foundation models are able to extract feedback signals from users, those signals tend to be quite weak or unreliable: asking users to rate responses with thumbs up/down or occasionally giving users a choice between two answers. Furthermore, much of the training data for foundation models is non-proprietary (e.g., it comes from crawling the public internet) or non-unique (e.g., obtained via non-exclusive partnerships with data providers[19]).

A final distinguishing factor is the current state of play in the market. There are already numerous well-funded providers offering competing foundation models, including Anthropic, Google, Meta, Microsoft, Nvidia, OpenAI, and X.ai. Obviously, Google, Meta, Microsoft and Nvidia are large incumbents with no parallel in the setting of online search in the late 1990s. Furthermore, Anthropic, OpenAI and X.ai have each raised billions of dollars.[20] Moreover, several model providers, including Meta, Mistral and Nvidia, offer open-source models that are on par with the closed source models and that allow other providers to enter and build their own solutions on top of them for free.

Aside from data feedback loops, one might wonder about the extent to which foundation models benefit from "traditional" network effects. Indeed, at first glance, foundation models look like operating systems for the applications built on top of them. Just like operating systems (e.g., iOS or Android or Windows) expose application programming interfaces (APIs) for third-party developers to build their applications, foundation model providers expose APIs for third parties to access their models and fine-tune them for specific applications.

Despite this similarity, however, foundation models do not exhibit the cross-side network effects inherent in operating systems (OSs), where users want to adopt the OS that has the most applications and developers want to build applications for the OS with the most users. The key difference is that users do not have to "adopt" foundation models

[19] For instance, Reddit has opted to license its data on a non-exclusive basis, initially to Google and later to OpenAI.

[20] In the case of X.ai, it was able to build a leading LLM within months of its founding and simultaneously establish one of the most powerful computational facilities.

in any meaningful way: they don't buy a device like they do when they buy an iOS, Android or Windows device. Users only adopt and use the final AI application, which is built on top of the relevant foundation model(s).

Theoretically, things could become more similar to operating systems if foundation model providers started selling users model-specific devices that were specifically designed to run AI applications built on them. Cross-side network effects between users and app developers could also arise if AI foundation model providers introduced model-specific app stores so that users of their model could access third-party applications of their model (e.g., OpenAI's GPT Store for third-party GPTs built on ChatGPT). That being said, it seems more natural that a discovery platform for AI applications should be model agnostic, allowing the best AI applications to be discovered regardless of which model they are built on.[21] Indeed, there is no natural attachment by end-users to the foundation models underlying the AI applications they are interested in.

The factors discussed above suggest that foundation models are likely to result in a far less concentrated market outcome compared to Internet search. There are contrary views though.

One contrary view is that the foundation models of a few big-tech firms will dominate because of their unique access to their own proprietary training data from the other services they deliver. For instance, in principle, Google has access to its data on YouTube, Google search, Gmail, Maps, Play Store, Google Assistant and its Workspace Suite to train its Gemini models. Likewise, in principle, Meta has data from Facebook, Instagram, and WhatsApp to train its Llama models. However, these large tech companies may struggle to fully utilize such data to gain a competitive advantage, due to a combination of regulatory constraints like General Data Protection Regulation and the Digital Market Act in Europe, as well as the fear of potential public backlash.[22] More importantly, it is questionable how uniquely valuable their data is relative to existing and other third-party data sources when it comes to training foundation models. While it is still early days, so far, there is no evidence that Google's Gemini or Meta's Llama models are outperforming OpenAI's GPT series or Anthropic's Claude series. Similar points apply to Amazon, Apple and Microsoft, none of which have produced their own frontier foundation models at the time of writing.

A second (related) contrary view is that there are large economies of scale and scope in collecting data for training state-of-the-art foundation models. The growing investment requirements for such models might suggest the number of players that a market of a given size can support is shrinking, which creates a force towards natural monopoly. However, as Vipra and Korinek (2024) discuss, offsetting this force is that the market size for generative AI is expected to rise as well. Moreover, it is not clear that collecting ever

[21] Capterra and G2 already do that for business-orientated AI tools (along with all other existing business-orientated software).

[22] That being said, there is some evidence that these companies are trying to push the boundaries of how much they are allowed to exploit user data for training their respective AI systems. See https://www.nytimes.com/2024/04/06/technology/tech-giants-harvest-data-artificial-intelligence.html.

more training data will remain a critical factor for building the best foundation models going forward. Instead, improving algorithm design, fine-tuning and inference may become increasingly important.

A third contrary view is based on the emergence of a breakthrough general AI technology – possibly a multi-modal form of artificial general intelligence (AGI) – developed by a single firm ahead of its competitors. If this firm can subsequently engineer a robust data feedback loop around this technology, thereby gaining an increasing competitive advantage, it could conceivably dominate a core set of applications. Nevertheless, we consider this scenario very unlikely.

Should the advantage stem from novel AI techniques or methods, it is likely to disseminate across firms as AI scientists share knowledge and move between organizations. Indeed, this is the story of the transformer model that has powered the advancement of generative AI over the last few years. The transformer model was initially developed within Google (by the Google Brain team) in collaboration with scientists from the University of Toronto. Subsequently, some of Google's AI scientists transitioned to OpenAI, which released the first widely adopted application of this technology (their Chat GPT product). In turn, researchers from OpenAI departed to establish Anthropic, which is currently at the forefront of foundation model development. We anticipate a similar trajectory for any future breakthroughs in AI techniques or methods.

Furthermore, even if a single firm managed to keep the technology to itself, this may be of little use for applications which require access to unique data on edge cases or individual consumer preferences. No amount of sophisticated reasoning can overcome such limitations, for instance, if the AI needs to offer customer-specific product recommendations.

Moreover, for applications that do not require sophisticated reasoning (e.g., AI tools to produce marketing materials), there will likely be a choice of many low-cost non-frontier foundation models including open-source options. There is no need to use AGI to write copy for marketing when existing models work well enough and are a lot cheaper to run.

That being said, it is possible that some narrow fields of application may end up being dominated by a small number of foundation models (or applications of such models). These would arise when, for a particular use case, they manage to engineer powerful data feedback loops, mirroring the experience with Google in online search. More generally, assessing the market concentration in specific domains would require a case-by-case analysis which not only considers the possibility of powerful data feedback loops, but also traditional barriers to entry from distribution advantages, economies of scale, switching costs and so on.

3.2. The full vertical stack

We don't expect market concentration to arise among data providers, or the data sources that would be relevant to train foundation models or most applications. As

explained in Section 2.2, a feature of data used for training AI models is very often there are many alternative sources that may substitute for one another. Figure 1 identified some examples, but there are many more.

Over time, a growing number of such firms will recognize the value of their data for training or fine-tuning AI models, and they will license access to it as a new revenue stream. Often, data sets may be useful in training or fine tuning a range of different models and applications. Given the diverse range of potential applications, it often wouldn't make sense for such data providers to sign exclusive agreements to license their data to just one model provider, meaning access to such data would remain open to multiple AI firms. For example, Reddit has opted to license its data on a non-exclusive basis, initially to Google and later to OpenAI. This approach ensures that no single company gains a decisive competitive advantage through exclusive access.[23]

In contrast to the model or data layers, the cloud services layer is likely to see continued consolidation around a few dominant players, namely Amazon Web Services, Google Cloud and Microsoft Azure. The main reason is that cloud services involve substantial economies of scale, especially in the context of AI.[24] The leading cloud providers have invested billions of dollars building huge racks of GPUs that AI model providers can access.

The broader concern is then that the largest cloud providers (Amazon, Microsoft, and Google), often referred to as "hyperscalers" due to their immense scale, might be able to leverage their dominance in cloud services across other layers of the AI stack. Indeed, the hyperscalers, along with the other two big-tech companies (Apple and Meta) are increasingly involved across all levels of the vertical AI stack:

- Chip development: all five big-tech companies (Amazon, Apple, Google, Meta and Microsoft) are actively developing proprietary chips for AI model training. Google leads this effort with its Tensor Processing Units (TPUs) which serve as an alternative to Nvidia's GPUs for training its foundation models including Gemini.
- Data: all five big-tech companies have access to vast amounts of proprietary (though not necessarily unique), data, as we detailed in Section 2.1 and further discussed in Section 3.1.
- Foundation models: Amazon, Google, Meta and Microsoft have developed their own AI foundation models in various overlapping domains, as illustrated in Figure 1. In addition, Microsoft has made substantial investments in OpenAI and so has Amazon in Anthropic.
- Applications: all five big-tech companies are also active in the AI application layer, offering various AI co-pilots, chatbots, virtual assistants, and numerous other applications (see Figure 1).

[23] We discussed the possibility some foundation model providers may have access to their own proprietary data due to other online services they provide in Section 3.1.

[24] Biglaiser et al. (2024) discuss the main facets of the development of cloud services, the applicable economics and some related policy issues.

The ability of hyperscalers to leverage their dominance in cloud services across other layers of the AI stack will depend on whether they resort to strategies such as vertical acquisitions, exclusive dealing, bundling and tying, among others. These are standard issues in competition policy and will be discussed in Section 5.

One might have expected that even without any explicit leveraging conduct, the hyperscalers' strength in compute power and data access would have naturally positioned them as leaders in the foundation model layer. However, as noted in Section 3.1, all five big-tech companies have been slow to establish themselves at the forefront of foundation model development, particularly when compared to startups like OpenAI and Anthropic. We think three key factors could be behind this: (1) big-tech companies, with their many different interests, were initially not focused on generative AI, unlike their startup rivals, (2) the advantages big tech companies enjoy in terms of access to compute and talent resources have partially been eroded by startups (and their investors) being willing to make substantial investments to obtain such resources; (3) the threat of regulation and antitrust which is focused on big-tech companies may be disciplining them from aggressively using their proprietary data obtained from other services to gain a competitive advantage in developing AI models.

The other potential source of market concentration in core AI services is at the hardware layer. As noted earlier, Nvidia's market share in GPUs used for training AI models is above 90%. However, it is important to recognize that a significant portion of Nvidia's sales of its latest GPUs comes from the hyperscalers themselves. As these hyperscalers increasingly develop their own custom chips[25], their reliance on Nvidia may diminish. In the longer term, some could even emerge as direct competitors if they start selling their chips to other companies.

At the same time, numerous startups are currently working on innovative chip designs for AI training and inference. These include specialized AI accelerators (Cerebras Systems, SambaNova Systems), optical computing (Lightmatter, Luminous Computing), neuromorphic computing (BrainChip), custom Language Processing Units (Groq), and quantum computing (PsiQuantum, IonQ). One or more of these technologies could potentially disrupt Nvidia's dominance in the hardware layer.

Despite the potential for future competition, concerns persist about Nvidia being able to leverage its current dominance in GPUs to impede new entrants. There are two main ways Nvidia could do this. One is via contracts that require firms use Nvidia chips exclusively, or via strategic pricing mechanisms such as market share discounts or loyalty rebates that have similar effects.[26] Given Nvidia's dominant market position, such practices would likely face scrutiny under standard competition law, as will be discussed in Section 5. A more subtle mechanism would be leveraging Nvidia's CUDA software,

[25] For instance, Google's Tensor Processing Units (TPUs) and Amazon's Trainium chips are already being used for in-house AI workloads, signalling a shift towards greater self-reliance in AI compute infrastructure among major cloud providers.

[26] This is a traditional concern in technology supply chains, which was also raised with respect to Intel and AMD in computer CPUs. See, for instance, Tom et al. (2000) and Scott-Morton and Abrahamson (2017).

which has become an industry standard, a possibility that will be discussed in Section 5 as well.

In summary, given the current state of AI, the primary risks to maintaining competitive outcomes in core AI services arise from two sources: the three large cloud hyperscalers leveraging their strong market positions from the cloud services layer to try to dominate some other layers, and Nvidia doing the same given its current dominance in the hardware layer.

That being said, there is another important aspect of the competitive dynamics given there are large players in all layers of the vertical AI stack. Namely, each large player with a strong position in one layer has a strong incentive to commoditize the other layers, especially the ones in which they have a weaker position. For instance, this helps explain why Meta has emerged as a strong supporter of open source LLMs with its Llama series and why Nvidia has launched its own open-source family of LLMs named NVLM. Neither wants to see the foundation model layer dominated by one or two players, that their platform services (Meta) or chips (Nvidia) would then depend on. This is also why all five big-tech companies' efforts to develop proprietary chips are best understood as defensive moves to avoid being dependent on a dominant Nvidia, rather than offensive efforts to dominate the chip layer.

Based on the above analysis, by the most conservative count, we could end up with at least seven major players operating at multiple (possibly, all) levels of the AI stack. These are the five big-tech companies (Amazon, Apple, Google, Meta, Microsoft) plus Nvidia and Open AI. Compared to the very high concentration seen in other core platform services—such as Google in online search, Amazon in e-commerce marketplaces, Apple and Google in mobile app stores, Microsoft, Apple and Google in operating systems (i.e. Windows, iOS, and Android), and Google and Apple in web browsers—a market with seven well-funded competitors (as well as some other major firms specializing in specific layers of the stack) would be a significant improvement. Moreover, with proper enforcement of competition law, there would be the possibility of entry at each layer of the AI stack, even if scale economies continue to mean a relatively high concentration in cloud services.

4. How will AI affect the market structure for existing sectors?

Having discussed the prospects of the market structure for AI core services, we now turn to the impact of AI on existing markets and industries. In principle, large incumbents stand to benefit most from the rapid rise of AI tools. This is because they can leverage their large customer base for potentially unique data to train or fine tune AI models on. Several large incumbents (Adobe, IBM, Intuit, Oracle, Salesforce, SAP) in the software sector have indeed been quick to build AI applications or embed AI in their applications. However, it remains to be seen whether these efforts will create strong data feedback loops. As discussed in Section 2.2, obtaining a strong data feedback loop is actually much harder to achieve than one may first think.

Meanwhile, there are several areas where recent developments in AI are making incumbents more vulnerable to disruption, not less. This reflects that AI provides entirely new ways for solutions to be offered which can disrupt the existing products. We consider a few specific examples before providing a broader characterization of how incumbents may be disrupted by AI.

Google search

Large language models (LLMs) that are continuously updated by crawling the internet offer a natural alternative to traditional search engines like Google's. Perplexity has emerged as a leading contender. When a query is submitted to Perplexity, it uses AI to search the internet in real time, gathering information from various sources, which it distils into a concise answer with relevant citations. For many types of queries, Perplexity provides a better way to get to the answer. Perplexity doesn't have its own foundation model, but rather utilizes various existing models, including models from Open AI and Anthropic. In July 2024, OpenAI launched its own search engine, SearchGPT, to compete with Google's.[27]

Google may be particularly vulnerable to LLMs due to the innovator's dilemma. This may not be so much about giving up on its existing business model, as some have suggested[28], but rather is about Google facing greater reputation risks as a large, established incumbent. Google has more to lose from providing occasionally inaccurate search results (e.g., for edge cases) compared to a startup. This allows startups to capture market share for certain types of searches, especially those where there is uncertainty around the answer.

However, this doesn't mean Google search is at imminent risk of obsolescence. For users who simply want to find and navigate to specific websites, Google search remains highly effective. And Google will not stand still: it can (and increasingly already does) integrate direct answers into its search results which are referred to as "AI Overviews".[29] It may do so cautiously at first, until its AI solutions can more reliably eliminate hallucinations, reflecting its greater concern for maintaining reputation than upstart competitors. Moreover, if websites start restricting the ability of search engines and AI services to index their content in order to prevent their data being used for AI training unless they are suitably compensated, Google will be in a strong position to attract the widest range of indexable websites (given its deep pockets and the dominance of its search engine), further cementing its data advantage.[30] Nevertheless, it is fair to say that

[27] See https://www.wsj.com/tech/ai/openai-search-engine-searchgpt-97771f86.

[28] Google could still show ads in its answers to user queries, and indeed, such answers could be even more suitable for targeted ads given the intent is likely to be even clearer with such queries (e.g. in response to a question about certain limitations of a particular product).

[29] See https://blog.google/products/search/generative-ai-google-search-may-2024/ in which Google states it expects to bring AI Overviews to "over a billion people by the end of the year".

[30] See https://www.theverge.com/2024/7/24/24205244/reddit-blocking-search-engine-crawlers-ai-bot-google for an account of how Reddit, which has a financial agreement with Google to use its training data, is only allowing Google search to index recent posts and comments.

AI-based search engines represent the biggest disruptive threat to Google search in many years.[31]

Adobe Photoshop

Stability AI, Midjourney, and OpenAI's DALL-E are significantly challenging Adobe Photoshop's traditional dominance in image editing and creation. These AI-powered tools enable users to generate high-quality, complex images from text descriptions in seconds, a task that would have taken hours using traditional software like Photoshop. They democratize image creation, enabling non-artists to produce professional-looking visuals without extensive technical skills. This shift threatens Photoshop's market position, particularly for tasks like concept art, illustrations, and even some types of photo editing. The speed, ease of use, and continuously improving quality of AI-generated images are making these tools increasingly attractive alternatives for many users and businesses.

In response to this increased competition, Adobe introduced its own AI image generation tool, Firefly, which integrates with its existing products. The company is also focusing on ensuring the ethical use of AI, addressing concerns about copyright and data privacy that have emerged with alternative generative AI tools. By combining AI capabilities with its established software suite and targeting professional users, Adobe is adapting to the changing landscape in an attempt to prevent full-scale disruption.

Uber

Uber faces significant disruption risk from the advancement of autonomous vehicles (AVs). Companies like Waymo, Cruise, and even traditional automakers are developing AV fleets that could offer ride-hailing services without the need for human drivers. Perhaps the most interesting version of this is Tesla's vision of creating a self-driving ride-hailing app called "Tesla Network", which would allow owners of Tesla AVs to rent them out for transporting other people, thereby generating income while not using their cars.[32] The realization of such initiatives would significantly reduce Uber's network effect advantage, as other companies deploy fleets of AVs without needing to attract and retain drivers. As a result, power in the industry could shift from today's leading ride-hailing platforms like Uber to AV manufacturers and operators. Uber has responded by forming a partnership with Waymo (after giving up on its own AV developments), but it remains uncertain whether Uber can transition successfully and maintain its market position once AVs are widely used.

[31] Similar points apply to AI assistants, and it is not yet clear whether the likes of Alexa and Siri will benefit more from being able to use generative AI to improve their performance or whether generative AI means they are of greater risk of disruption with users switching to third-party AI assistants.

[32] See https://www.theverge.com/2024/4/23/24138580/tesla-robotaxi-ride-hail-app-preview-earningsq1-2024.

Providers of and marketplaces for online services

Other firms facing disruption risk include marketplaces Upwork and Fiverr (some of their digital content creators can be replaced by AI solutions), Chegg (its homework help services can be replaced by ChatGPT and other virtual assistants), Getty Images (its library of licensed images is less valuable in a world of easy generative AI image generation).

These examples highlight a key mechanism by which AI is expected to disrupt incumbents. As well as simplifying the experience for consumers, generative AI can directly replace digital tasks previously performed by human suppliers. By eliminating the need for human suppliers, AI entrants can overcome an incumbent's network effect advantage, and lead to the commoditization of markets.

In the coming years, AI agents could take this mechanism one step further, by making transactions on a consumer's behalf. For example, an AI agent may be able to search for and book a hotel room that meets certain criteria, potentially undermining the role of incumbent booking platforms like Booking.com. If the AI agent can search all available channels, including direct hotel listings, and save users the hassle of dealing with any specific hotel's website, reservation and payment system, it could diminish the advantages that marketplaces like Booking.com have in reducing search and transaction costs. And even if Booking.com and other platforms maintain their advantage for conducting search and transactions over direct channels, AI agents would still make it easier for users to search across rival platforms. In other words, they would reduce user multihoming costs, thereby increasing competition among existing platforms.

We envision a future world where consumers subscribe to an AI agent service that completes many online tasks for them, across many platforms and individual services. Over time, this service would learn their preferences, potentially becoming better at finding exactly what they want than any existing marketplace such as Booking.com could (e.g., knowing their full travel plans to suggest the ideal hotel location). To the extent the AI agent provider has many customers, it would also benefit from across-user learning, improving its service for all of them. Furthermore, since it represents many consumers, it could negotiate better group prices with suppliers. If such an AI agent provider relied solely on charging consumers a monthly subscription fee for its service, it could eliminate the conflict of interest that existing intermediaries face, where they steer consumers towards options that generate the most commission revenue.

More generally, any online marketplace where suppliers can be reached through other channels will be vulnerable to disruption from such AI agent providers. However, this may not necessarily result in a competitive outcome. Instead, we may replace one type of gatekeeper with another. AI agent providers with larger consumer bases could make better recommendations through positive data feedback loops. Powerful AI agent providers could emerge, charging suppliers high fees to access their unique consumers, potentially reintroducing the misalignment of interests between the intermediary (if it obtains revenue from suppliers) and its consumers. While we don't expect individual

marketplaces (Airbnb, Booking.com, Uber) to dominate the provision of such general AI assistants, they could be offered by existing firms who control key access points for consumers. Leading candidates are providers of operating systems (Apple, Google, Microsoft), which we turn to next.

Operating systems

On the one hand, if existing operating systems can deeply integrate AI, they would make it harder for entrants to enter and compete. Consider Microsoft, Apple and Google with respect to Windows, iOS and Android, respectively. Each could integrate AI into its operating system in ways that limit the reach of rival third-party developers, app stores or AI systems that would otherwise enter and try to compete. For instance, Apple's announcement of its planned integration of Apple Intelligence into iOS could potentially lock users further into its App Store by having its AI agent or assistant recommend only apps available on the App Store. Similar concerns arise with Android, where Google could induce Android phone makers to integrate Google's AI tools and exclude rivals.[33]

On the other hand, AI could enable new form factors and associated operating systems to emerge that are AI-native. Whether this is based on glasses, earbuds, or some other kind of wearable, the idea is that by making use of new modalities (such as sound and vision), an AI-powered virtual assistant has the potential to disrupt at least some aspects of the existing mobile ecosystems.

To conclude this section, it is useful to sum up the key mechanisms through which AI can disrupt market leaders in existing sectors, based on the examples above. First, AI can automate services that were previously done by humans using traditional software (e.g., Adobe, Fiverr, Upwork). Second, AI can remove the advantage due to network effects that incumbent platforms may have (e.g., Uber, Fiverr, Upwork, Booking.com, Google). Third, AI can make it easier for users to multihome across multiple service or marketplace providers (e.g., AI agents comparing multiple services or marketplaces and making purchases on behalf of users). And fourth, AI can enable entirely new ways of providing services to replace existing services or platforms (e.g., online search and mobile operating systems). On the other hand, as our discussion of AI agents illustrates, there may also be ways AI can be used to create new types of gatekeepers.

5. What key competition policy issues are raised by AI?

We divide this section into two parts: traditional competition policy issues that arise for core AI services (Section 5.1) and new, AI-specific competition policy issues (Section 5.2).

[33] See for example the investigation by EU antitrust regulators of Google's deal with Samsung to integrate Google's generative AI tool into Galaxy smartphones: https://www.pymnts.com/cpi-posts/eu-antitrust-regulators-probe-google-samsung-ai-deal-for-potential-anti-competitive-practices/.

5.1. Traditional competition policy issues raised by AI

As foreshadowed in Section 3, the key competition concerns for core AI services at this point are rather traditional. The primary concern is that companies with substantial market power at one level of the vertical stack may leverage their position to limit or distort competition at other levels. There are several ways this could play out.

One possibility is through exclusive deals, where foundation model providers access unique and proprietary data sets. However, to date, most deals appear to be non-exclusive licenses. Examples include Google and OpenAI with Reddit, Meta and OpenAI with Shutterstock, Google with Stack Overflow, and OpenAI with Axel Springer, the Atlantic, and Time Magazine. Of course, this could just reflect firms' awareness of the antitrust risk associated with exclusive licenses, suggesting the threat of antitrust enforcement may be working here.

Another concern is data-driven acquisitions aimed at obtaining unique and proprietary data or preventing rivals from accessing such data (see Hagiu and Wright, 2023a and de Corniere and Taylor, 2024). Given the wide range of use cases of AI models and the many different sources of data that can be used to train models (illustrated in Figure 1 above), we are quite sceptical that such acquisitions would be driven by exclusionary motives.[34]

A greater concern could arise if hyperscalers tried to leverage their significant market positions in cloud services to dominate other parts of the vertical stack. This could occur through tying or bundling their cloud services with their AI models, requiring that users of their cloud services use their AI models. However, such practices fall well within the scope of existing competition law considerations. The same applies if Nvidia were to engage in similar conduct, leveraging its dominant position in chips into AI models or cloud services.

A more likely strategy for Nvidia is to maintain its dominant position in chips by raising barriers to entry through the network effects created by its industry-standard CUDA software. CUDA is widely adopted in the AI sector, with many popular AI frameworks and libraries optimized for it, and most developers familiar with using it.[35] This makes it less likely for companies to consider alternative hardware. Nvidia continually optimizes CUDA for its hardware, so AI applications using CUDA perform better on Nvidia GPUs. Nvidia could further reinforce this advantage by limiting or reducing CUDA's compatibility with rival chips or by more tightly integrating its hardware with CUDA. This seems to be an area deserving of investigation by competition authorities, but one which would require specialized technical knowledge to properly evaluate.

---

[34] In general, there are many non-exclusionary reasons why firms may engage in data-driven mergers and acquisitions, some which may lead to efficiencies, and some that may lead to consumer harm (e.g., more rent extraction). See de Corniere and Taylor (2024).

[35] As of 2024, there are more than two million developers using the CUDA software platform to build AI and other applications. See https://www.businesstimes.com.sg/companies-markets/telcos-media-tech/behind-plot-break-nvidia-s-grip-ai-targeting-software.

Finally, large technology companies seem to have found a novel twist on the traditional strategy of acquiring potential competitors. To the extent AI scientists are the scarce resource in building foundation models, one way to skirt existing merger and acquisition restrictions is to structure a deal to hire the relevant AI employees from the target firm and compensate founders and investors with generous licensing terms for their technology which would be of little value without these employees. This approach is illustrated by Microsoft's deal with Inflection and Amazon's deal with Adept, which are currently being investigated by the Federal Trade Commission in the U.S.[36] Beyond these specific deals, this raises a broader question of how to handle cases where firms hire key employees away from a competitor where the purpose may be to eliminate competition.

5.2. New competition issues raised by AI

An important competition issue raised by the emergence of AI is the role of pricing algorithms in facilitating collusion (see for example Calvano et al., 2020) and whether existing competition law or regulations need modification to address these algorithms effectively. While the ability to collude (including tacitly) in dynamic settings is not new, algorithmic price fixing does raise new issues around detection, enforcement, liability, and regulation. Several key questions emerge: Can the use of such algorithms be considered a facilitating practice? If competitors engage in price increases using algorithms from a common provider, does this constitute a hub-and-spoke cartel? If so, is the algorithm provider also liable for any claim of price fixing? Should providers of these pricing algorithms be subject to any regulation (e.g., not being able to share private information gathered from one customer with others when these customers are competing)? Should it be illegal for one firm to communicate with its competitors about the use of a particular pricing algorithm given this could be used to coordinate rivals on the same software? We do not attempt to answer these questions or survey the burgeoning literature on this topic, but instead refer the interested reader to Hanspach and Galli (2024) and the chapter surveying this topic in Gans (2024).

What is less discussed is that similar issues are likely to arise beyond algorithms that induce collusion. AI models could potentially implement anti-competitive conduct more broadly in hard-to-detect ways, such as through price discrimination (as suggested by Ezrachi and Stucke, 2016), biased rankings and selective recommendations (as analyzed in Calvano et al.,2024), or strategic contracting. When a firm with market power relies on AI to maximize its long-run profits, it may engage in conduct that abuses its market power. Even if the AI is programmed to avoid explicitly violating existing competition law, it might still find ways to circumvent the law in its efforts to maximize the firm's profit. In essence, the AI would follow the letter of the law but not the spirit.

[36] See https://www.theverge.com/2024/7/1/24190060/amazon-adept-ai-acquisition-playbook-microsoft-inflection and https://fortune.com/2024/07/17/big-ai-acquihire-microsoft-inflection-amazon-adept-antitrust-cma-ftc/.

The opacity of many deep learning AI models presents a significant challenge for regulatory oversight. Due to their black-box nature, it may be impossible for competition authorities to determine whether an AI has an anti-competitive purpose. As a result, authorities may be forced to only focus on effects. However, drawing clear boundary lines between normal profit-maximizing behavior and anti-competitive behavior is extremely challenging. These difficulties mirror the challenges authorities already face when investigating any alleged price fixing via opaque pricing algorithms.

Price parity clauses (PPCs) provide a relevant example of potentially anti-competitive practices that AI could exacerbate. Traditionally, Amazon.com, Booking.com, Expedia, and certain price comparison platforms have employed these clauses in their contracts to prevent suppliers from offering lower prices through other channels, including their own websites. An existing literature has demonstrated the anti-competitive effects of such practices.[37] Consistent with these findings, these clauses have been banned or removed in a number of jurisdictions.

However, a platform could use its ranking algorithm or AI-driven recommendations to demote suppliers offering lower prices on other channels, effectively mimicking a PPC's impact without explicit PPC clauses in their contracts. While conventional algorithms could be programmed for this purpose[38], recent AI advancements introduce a crucial distinction: such behavior no longer requires explicit programming. Instead, an AI system tasked with maximizing the platform's long-term profit might autonomously develop strategies that replicate PPC effects as it steers consumers towards suppliers that generate higher conversion rates (and so platform revenue). This self-learned behavior could achieve the anti-competitive effects of PPCs while potentially evading traditional regulatory scrutiny.

This scenario raises several intriguing questions about the intersection of AI and competition law. First, could simulations or other analytical methods be developed to test whether AI models produce specific anticompetitive outcomes? Such tools might help regulators identify potential issues before they manifest widely in real markets. Second, is it feasible to impose meaningful restrictions on black-box AI models to ensure they adhere to both the letter and spirit of competition law? Third, given the opacity of AI decision-making, would "object" (or "purpose") still play a meaningful role in competition law or would it all just come down to "effects"?

Another new issue relevant to competition policy is the emerging debate over fair use of data in AI training.[39] On one side, there is a legitimate concern around protecting intellectual property rights and ensuring fair compensation for creators (for the likes of published articles, books, music, and videos). On the other hand, restricting access to

[37] See Edelman and Wright (2015), Boik and Corts (2016), and Wang and Wright (2020).

[38] For instance, Hunold et al. (2020) find empirically that OTAs demote hotels in their search results that set lower prices in other channels.

[39] Gans (2024) provides an economic model that can be used to evaluate some aspects of the debate and proposes some possible solutions for balancing the interests of copyright holders and AI developers in the case of large AI models.

training data from these sources could create significant barriers to entry in the AI sector, potentially stifling competition and innovation. Large companies may be able to afford the time and resources to negotiate licensing deals with the relevant intellectual property rights owners, while small startups may not.

Consider the case of academic journals and scientific research. The question is whether publishers like Elsevier, Springer, Wiley have the right to control AI training on academic articles published in their journals, or should this be considered fair use? While using AI models to reproduce copyrighted articles for profit would clearly violate fair use principles, the situation becomes more nuanced when considering AI's potential to advance scientific discovery by synthesizing information from multiple sources. Is such learning not reflective of how science has always advanced, by standing on the shoulders of giants? Moreover, would it be ok for the likes of Elsevier, Springer, Wiley to potentially control the future of AI scientific development by controlling access to a broad set of scientific articles published in their journals, possibly through exclusive partnerships with specific AI providers?

These issues extend beyond academic publishing to encompass all forms of creative work. If IP law needs to be adjusted, then competition authorities should be involved in the process to ensure the proposed adjustments balance the rights of content creators and publishers against the potential effects on competition in the provision of innovative AI models and services. This balance is crucial to prevent the emergence of new bottlenecks in the AI ecosystem while still protecting intellectual property rights.

This debate intersects with broader discussions about data access and digital markets. Regulators may need to consider whether certain critical datasets should be treated as essential facilities in the context of AI development. They might also need to explore new models for compensating content creators while still allowing for the widespread use of data in AI training.

6. Concluding thoughts

In this article, we've considered some of the key competition policy issues that arise from AI. We addressed three questions: the likely market structure for core AI services, the likely effects of AI on the market structure of other sectors, and competition policy issues arising from AI.

In our view, competition concerns within core AI services are most likely to arise from the hyperscalers (Amazon, Microsoft and Google) leveraging their market power in cloud services across other layers of the vertical stack (e.g., via exclusive data or model deals, vertical mergers, bundling/tying, or other types of integrations) and from Nvidia similarly leveraging its dominance in the hardware part of the stack into other parts of the stack. Nevertheless, it is important to note that there will be at least seven large and well-funded players competing across most of the core AI services for the foreseeable future (Amazon, Apple, Google, Meta, Microsoft, Nvidia, Open AI), which reduces the probability of a highly concentrated market outcome at any individual layer.

Provided anticompetitive vertical leveraging by these large players is prevented, we are less concerned about the foundation model layer or the data layer becoming monopolized. We explained why the dynamics are likely to be quite different from what played out in the case of online search. In general, not all data is unique and not all data feedback loops are strong, so one should not presume the data feedback loops that could arise around AI models will lead to one or two large winners. Whether this might happen in any particular area of AI applications depends on whether data is unique, part of an automatic feedback cycle, and whether its marginal value remains high.

For large incumbent businesses, despite their inherent data advantage, we explained how AI may lead to disruption in certain cases. Specifically, AI has the potential to undermine platform businesses built on network effects where one side of the market can be replaced by an AI solution. But it is not clear whether the AI providers that replace these incumbents will always lead to more competitive outcomes. Given their broader scope, AI agent providers could end up being even more powerful gatekeepers than the existing gatekeepers they replace. We still believe the most dominant positions will arise where platforms can enhance traditional network effects via AI. Integrating AI into existing operating systems could be a case in point.

AI models present new challenges for competition law enforcement due to their potential for implementing anti-competitive practices in algorithmic ways. These are issues that have already generated considerable interest in the case of algorithmic collusion, and will become even more important as companies start relying on more general, black-box AI models for an increasing range of strategic decisions. As the AI models are instructed to maximize long-run profits, there may be other ways (beyond just pricing) in which they engage in anticompetitive-like behavior while technically adhering to existing competition law. We gave the example of how AI-driven ranking algorithms could replicate the anti-competitive effects of price parity clauses without any explicit instructions. These developments raise important questions about testing AI models for anticompetitive harms, and the future role of "object" versus "effects" in competition law analysis.

We close by calling for more research on AI and competition policy. Our agenda was largely to frame some of the key issues. It remains for future research, including interdisciplinary efforts on the part of economists, legal experts, and AI researchers, to help provide more definitive answers.

References

Acemoglu, D. (2024) "Harms of AI," in Justin B. Bullock and others (eds), The Oxford Handbook of AI Governance, Oxford University Press.

Bick, A., A. Blandin and D. J. Deming (2024) "The Rapid Adoption of Generative AI," Federal Reserve Bank of St. Louis Working paper No. 2024-027A.

Biglaiser, G., J. Crémer and A. Mantovani (2024) "The Economics of the Cloud,'' TSE Working Paper No. 1520.

Boik, A. and K. Corts (2016) "The Effects of Platform Most-Favored-Nation Clauses on Competition and Entry," *Journal of Law and Economics*, 59(1), 105–134.

Calvano, E., G. Calzolari, V. Denicolò, and S. Pastorello (2020) "Artificial intelligence, algorithmic pricing and collusion," *American Economic Review*, 110(10), 3267-3297.

Calvano, E., G. Calzolari, V. Denicolò, and S. Pastorello (2024) "Artificial intelligence, algorithmic recommendations and competition," Working paper.

Carballa-Smichowski, B., N. Duch-Brown, S. Höcük, P. Kumar, B. Martens, J. Mulder, and P. Prüfer (2023) "Economies of scope in data aggregation: evidence from health data," Working paper.

Comunale, M. and A. Manera (2024) "The Economic Impacts and the Regulation of AI: A Review of the Academic Literature and Policy Actions.", *IMF Working Paper* No. 2024/65.

de Corniere, A. and G. Taylor (2024) "Data-Driven Mergers," *Management Science*, forthcoming.

Edelman, B. and J. Wright (2015) "Price Coherence and Excessive Intermediation," *The Quarterly Journal of Economics*, 130(3), 1283-1328.

Ezrachi, A. and M. E. Stucke (2016) "Virtual competition," *Journal of European Competition Law & Practice*, *7*(9), 585–586.

Gans, J. (2024) "The Microeconomics of Artificial Intelligence" MIT Press, forthcoming.

Hagiu, A. and J. Wright (2020) "When data creates competitive advantage," *Harvard Business Review* 98(1), 94-101.

Hagiu, A. and J. Wright (2023a) "Data-enabled learning, network effects and competitive advantage," *RAND Journal of Economics* 54(4), 638-667.

Hagiu, A. and J. Wright (2023b) "To Get Better Customer Data, Build Feedback Loops into Your Products," *Harvard Business Review* (online blog), July.

Hanspach, P. and N. Galli (2024) "Collusion by Pricing Algorithms in Competition Law and Economics," Robert Schuman Centre for Advanced Studies Research Paper No. 2024-06

Hunold, M., R. Kesler and U. Laitenberger (2020) "Rankings of Online Travel Agents, Channel Pricing, and Consumer Protection," *Marketing Science* 39(1): 92-116.

Klein, T., M. Kurmangaliyeva, J. Prüfer and P. Prüfer (2023) "How important are user-generated data for search result Quality?" Working Paper.

Qi, L. (2024) “Stock Data Analysis of Competing Companies in Competitive Market: The Case of NVIDIA Corporation,” *Highlights in Science, Engineering and Technology*, 94, 493-503.

Schaefer, M. and G. Sapi (2023). “Complementarities in learning from data: Insights from general search.” *Information Economics and Policy* 65, 101063.

Scott Morton, F. and Z. Abrahamson (2017) “A unifying analytical framework for loyalty rebates,” *Antitrust Law Journal*, 81, 777-836.

Tom, W., D. Balto and N. Averitt (2000) “Anticompetitive aspects of market-share discounts and other incentives to exclusive dealing” *Antitrust Law Journal*. 67, 615-639.

Tucker, A.D., K. Brantley, A. Cahall, and T. Joachims (2024) “Coactive Learning for Large Language Models using Implicit User Feedback,” Working paper.

Wang, C. and J. Wright (2020) “Search platforms: showrooming and price parity clauses,” *RAND Journal of Economics*, 51(1), 32-58.

Vipra, J. and A. Korinek (2024) “Concentrating Intelligence: Scaling and Market Structure in Artificial Intelligence,” National Bureau of Economic Research Working Paper #33139.

Yoganarasimhan, H. (2019) “Search Personalization Using Machine Learning,” *Management Science*, 66(3), 1045-1070.