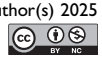
Check for updates

Article

Cloud Capitalism and the AI Transition

Politics & Society
2026, Vol. 54(2) 184–214
© The Author(s) 2025

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00323292251396395
journals.sagepub.com/home/pas


JS Tan[1] and Kathleen Thelen[1]

Abstract

This article explores the origins and implications of a new cloud business model that is powering the advance of AI. We document how this model emerged within a handful of the most dominant IT firms whose reach into all corners of the economy makes them a powerful node or "choke point" in the political economy as a whole. We then elaborate how the features of the cloud business model differ from the traditional platform model out of which it grew, as it evolved from asset-light to asset-heavy, from hierarchical organization to semivertical integration, from domination over to collaboration with partner firms, and from embracing consumer- to enterprise-facing strategies. A final section considers the technological, political, and distributional impacts of the rise of this new business model—showing how the current race to artificial general intelligence (AGI) has reinforced and accelerated its underlying dynamics (above all, intensifying the drive for scale and ever-greater asset intensity), analyzing the new techno-nationalist alliance between industry leaders and the state that the model's development has inspired, and considering the new power-distributional dynamics this model has produced.

Keywords

platform capitalism, cloud computing, artificial intelligence, techno-nationalist alliance

[1]Massachusetts Institute of Technology, Cambridge, MA, USA

Corresponding Author:
JS Tan, Department of Urban Studies and Planning, MIT, 77 Massachusetts Avenue, Cambridge, MA, USA.
Email: js_tan@mit.edu

Over the past several years, scholars in comparative political economy have increasingly turned their attention to the advent of a new "knowledge economy" and the kinds of firms that dominate it. Despite differences in emphasis, there is broad consensus on how the leading firms of the twenty-first century differ from the industrial concerns of the twentieth century. First, while the corporate behemoths of yesteryear were vast, capital-intensive operations that employed tens of thousands of (mostly unionized) workers, the firms that dominate today are famously "asset-light." Indeed, many of today's corporate giants often do not produce anything tangible at all but instead earn rents through the licensing of their designs and patents (e.g., Apple) or through their role as intermediates in the two-sided markets they create, linking workers and employers, buyers and sellers, clients and contractors, creators and viewers, or advertisers and consumers (e.g., Uber, Amazon, Meta, Google).[1]

Second, today's tech giants are seen as benefiting from economies of scale and network effects in what are essentially winner-takes-all markets.[2] First-mover firms that can conquer a market and box out competitors are the ones that prevail in a zero-sum contest for dominance. As Peter Thiel, now an active venture capitalist who himself founded a hugely successful platform (PayPal), put it, "If you're the founder, entrepreneur starting a company, you always want to aim for monopoly and you always want to avoid competition. . . . Competition is for losers."[3]

Third and finally, for traditional platform firms, data is the coin of the realm and the key to market dominance. In the consumer-facing markets in which firms like Google and Amazon compete, the goal is to lock in tens of millions of individual consumers who come to rely on them and in many cases are willing to "pay" for the services they provide with the data they share. The more data they control, the more they can refine the algorithm, attract advertisers, and increase user engagement in a self-reinforcing cycle that solidifies their dominance.

However, even as scholarly consensus on the main features of the new "knowledge economy" business model was consolidating, the strategies of key vanguard firms were quietly evolving. Over the past decade, some of the undisputed titans of the digital economy—notably, Amazon, Microsoft, and Google—have embraced a business model that is very different from the one just described. Whereas "traditional" platform firms were asset-light, these firms have become extremely asset-heavy. In 2024 alone, these three companies invested over $200 billion to upgrade their data centers and roll out new ones—and they are on track to top this in 2025.[4] And whereas traditional platform firms were mostly consumer-facing, these firms have become increasingly enterprise-facing. Rather than selling a standard service to tens of millions of individual users, they seek to cultivate long-term relationships with a smaller number of very large (ideally *Fortune* 500) corporate clients. Finally, data, while still important, is not the key to dominance. Instead, the crucial competitive resource for these firms is compute, that is, the physical infrastructure and processing power that is needed to run software or other specialized workloads for corporate clients on increasingly dedicated computing resources. The transformation within these leading companies points to a new production pattern within, and in

many ways at the forefront of, the knowledge economy, the emergence of what we call the *cloud business model*.

The core of the cloud business model is to provide IT infrastructure as a service—also known as cloud computing. Cloud computing, like any subscription service, provides customers with a way to rent servers as opposed to purchasing them. In the old days, a company would have had to run its IT systems from physical computers (servers, in IT speak) that it bought and maintained itself. It would have needed to set up its own networking cables and to purchase storage in the form of disk drives. Now that same company can outsource its IT infrastructure needs to a cloud provider and digitally rent it via the internet on an as-needed basis. The ongoing migration to the cloud by an increasing number of firms is thus driving an unprecedented concentration of computing capabilities and the consolidation of physical hardware. The fundamental building blocks of IT infrastructure—storage, network, and compute—correspond to physical components: memory chips or disk drives, fiber cables, and CPUs or GPUs, all of which are required to provide these services.[5] So, while we use the term "cloud capitalism," we do so reluctantly because the image of a cloud obscures the extremely tangible, asset-heavy underbelly of this new business model.

This profound shift in the business model of the firms at the technological frontier has also given rise to a new political dynamic, which we explore in more depth later. To preview briefly, whereas earlier platform firms leveraged their asset-light structure to engage in venue arbitrage, that is, moving operations across borders to evade (or to seek more favorable) tax or legal regimes, cloud firms are physically tethered to specific jurisdictions because of their reliance on massive data center infrastructure. The centrality of fixed assets has inspired important changes in the relationship to the state on the part of leading firms. Rather than regulatory arbitrage, we observe a strategic shift toward accommodation and, with that, the emergence of a new techno-nationalist alliance. This alliance, so clearly on display in the newly cozy relationship between Silicon Valley firms and the second Trump administration, is one in which geopolitical and commercial interests come together as the state facilitates rapid data center expansion at home (or in allied countries) while deploying export controls and sanctions to hobble competitors abroad.

The shift in the business model at the technological frontier is hard to see because the names of the firms are the same even as the center of gravity within them has been changing. Amazon is of course still operating as a marketplace for goods being sold to individual consumers, but the real investment is in the company's cloud division, Amazon Web Services (AWS), which is (for example) powering the predictive maintenance systems for auto manufacturers like Toyota. Microsoft is still offering its Office suite to individuals and businesses, but again the real action is in its cloud division (Azure), which is providing the IT infrastructure and AI models behind fraud detection systems for banks and airlines. Google continues to dominate search, but Google Cloud Platform (GCP) is providing the resources and the tools on which firms like Uber increasingly rely for powering the underlying data platform used to manage their operations.[6] In this way, today's extremely asset-heavy behemoths allow their

client firms—which also encompass much of the old platform economy—to become (or remain) asset-light by providing them with the IT infrastructure and computing power they need to compete in the increasingly digitally powered twenty-first century.

The legacy services of today's dominant cloud providers continue, for now, to generate the bulk of their revenue (except for Microsoft, whose cloud services surpassed its Office products in revenue in 2019). But having made windfall profits with their capital-light software innovations and platform services, they have reinvested their profits down the IT infrastructure stack into purchasing and building data centers and into other capital-intensive innovations like semiconductors, server-cooling technologies, and other hardware. So, while the cloud business model has emerged within the most dominant IT firms, it has done so *alongside* the existing corporate strategies in which these firms first rose to prominence. Where they were once considered equivalent to, albeit larger than, platform firms like eBay, Uber, Etsy, DoorDash, or Airbnb, today's top cloud service firms now sit in a category of their own, one that straddles old and new forms of knowledge production.

The number of companies participating in the cloud business model is very small—much smaller in fact than the number of platform firms. Amazon, Microsoft, and Google, in that order, account for two-thirds of the market; other players include Oracle, Alibaba, Huawei, Tencent, and IBM. But the cloud business model is nonetheless critical for understanding the latest developments in the knowledge economy. There are two key reasons.

First, a very large and growing number of companies across all industries in the United States have become clients of these firms because they have outsourced part or all of their IT infrastructure needs to them. Take manufacturing: John Deere increasingly relies on AWS and Azure to provide the IT infrastructure for its predictive maintenance services. Or finance: Citibank owns some of its own data centers but is increasingly migrating significant parts of its IT infrastructure to Google Cloud for the cost savings and flexibility this offers. Or platforms themselves: Netflix and Airbnb both depend on AWS to host their platforms.[7] The sheer reach of a handful of cloud providers into all corners of the economy makes them an increasingly powerful node or "choke point" in the political economy as a whole.[8] For example, when a portion of Amazon's cloud services went down in December 2021, it took a significant portion of our digital lives down with it; services at Netflix, Disney+, and the trading platform Robinhood were halted; Delta and Southwest airlines had to stop their services; even iRobot's vacuum cleaners stopped working.[9] Even as this article went into production in late 2025, we read about another AWS outage, which had an even broader impact, with disruptions reported at Venmo, Hulu, McDonald's, Coinbase, Ring, Lloyd's Bank, Bank of Scotland, Gov.uk, Signal, Slack, and WhatsApp, among others.[10]

Second, with the advance of artificial intelligence (AI), the cloud business model represents the new leading edge in the knowledge economy. Reliance on advanced IT infrastructure has long been integral to a range of digital areas like cybersecurity, data processing, 3D video rendering, and semiconductor design. It is also increasingly

essential for other fields such as genomics, financial simulations, oil and gas exploration, weather forecasting, and engineering. But it is absolutely central to the development of AI; indeed, the cloud is the foundational infrastructure upon which the entire AI ecosystem is being built. As such, the cloud business model is both a key enabler of AI development and, critically, the primary vehicle for its diffusion.

AI both reinforces and accelerates the underlying dynamics of the new model. Especially since the ChatGPT moment in 2022, the dominant belief underpinning AI advancement is that progress and the attainment of an all-powerful artificial *general* intelligence (AGI) depends on scale—more data, larger models, and above all, more compute. In this context, the computing capacity made available by the vast infrastructure of the cloud has become the most critical and contested resource. Competition in AI today is not so much about developing new intellectual property (researchers and AI companies often openly publish new innovations in their model design online)[11] but having access to the capital—the physical infrastructure—to train and run the largest models with the most parameters.[12] Thus, to understand the political economy of AI, we must first understand the cloud.

This article explores the genesis, evolution, and impact of the cloud business model. We begin by tracing its origins, reviewing—in a highly telescoped and stylized way—the journey from Fordism through the platform business model to the emerging cloud business model. We explain the origins of the AI transformation, which both stems from and reinforces the cloud business model, including the latest (post-ChatGPT) chapter in AI development—the race to AGI. Second, drawing from industry reports, conference materials, and blog posts, we outline the core attributes of this model, highlighting how the corporate strategies found in leading firms defy conventional expectations. We document multiple shifts: from asset-light to asset-heavy, from horizontally to semivertically integrated, from dominance and scale economies to the cultivation of long-term relationships and more intense interfirm collaboration, and from consumer- to enterprise-facing strategies. A final section considers the technological, political, and distributional impacts of the rise of this new business model—showing how the current race to AGI has reinforced and accelerated its underlying dynamics (above all, intensifying the drive for scale and ever-greater asset intensity); analyzing the new techno-nationalist alliance between leading tech firms and the state that the model's development has inspired; and considering the new power-distributional dynamics this new model has produced.

From the Platform Economy to Cloud Capitalism

Previous works have already elaborated the ways in which platform firms transformed twenty-first-century capitalism, so a short overview will suffice.[13] The dominant mode of production in the previous era of mid-twentieth-century Fordist accumulation was both labor and capital intensive. Firms employed large numbers of workers in massive, geographically anchored production sites. Their labor intensity made them vulnerable

to worker unrest, and thus provided an incentive for them to share the rewards of productivity gains with their employees. Traditional Fordist firms encountered intense strains starting in the 1970s with rising inflation, economic stagnation, and increased competition from lower-cost producers abroad. In this context, a new corporate governance paradigm took hold. The rise of shareholder value forced firms to reorganize around "core competencies" to maximize stock prices, prioritizing shareholders' interest over workers.[14] This organizational focus on reducing expenses and streamlining operations drove the fragmentation ("fissurization") of firms that were once consolidated and vertically integrated, as lead firms retained only the highest value-added activities—primarily knowledge production and strategic control—in-house, while outsourcing or offloading everything else.[15]

The technological innovations of the IT revolution and the advent of "platform capitalism" extended and supercharged many of the strategies pioneered by the fissured firm.[16] Networked technologies now allowed firms to manage suppliers and laborers in the periphery at low cost, fragmenting and "gigifying" work in ways that allowed them to forgo labor contracts altogether while still exercising close control over their workforce through algorithmic management techniques.[17] The goal for consumer-oriented platforms such as Uber and Amazon was to generate rents by squeezing the vendors, contractors, and advertisers who rely on the platform in order to reach and retain a large customer base.[18]

Because a platform's value famously depends on the size of its network, it was imperative for these firms to seek market dominance. Such dominance not only secures a steady stream of rents; it also makes it extremely difficult for new market entrants to overthrow incumbents. This is why leading platform firms were willing to forgo short-term profits and optimize instead for growth. This strategy was underwritten by deep-pocketed investors (many of whom had acquired their vast wealth as a consequence of financialization) with the resources to fuel years of profitless growth. And of course, once a digital platform solidifies its market dominance, it can live off the "platform rents" garnered not through continuous productive activity but by leveraging the firm's market position and monetizing interactions on the platform.

Finally, for traditional platforms, data is the crucial competitive resource. The success of online services like search engines and social media hinges on their ability to "harvest" user data to deliver highly targeted and personalized ads, a process Shoshana Zuboff terms "surveillance capitalism."[19] The more data a platform gathers, the better it can refine its algorithms, attract advertisers, and enhance user engagement, creating a self-reinforcing cycle that solidifies its market dominance.

As much as existing analyses of post-Fordist corporate strategies and platform capitalism explain the "knowledge economy," they fail to capture the most recent, and most important, developments that now represent its leading edge. For that we need to understand the cloud business model that lies behind these new developments—a topic that has in recent years received some scholarly attention but has not yet been analyzed as a distinct business model.[20] The next section recounts briefly—and in a

highly simplified way—the emergence of the cloud business model that is at the forefront (and in many ways, represents the backbone) of the "AI revolution."

The Rise of the Cloud

As already noted, the cloud business model has its roots in the older platform model. Unlike previous paradigm shifts where incumbents were displaced by challengers with new corporate strategies, the incumbents of the platform firm era pioneered and now dominate the cloud business model. This shift is evident in the leadership changes at the two top players, Amazon and Microsoft. Both Andy Jassy and Satya Nadella, who now serve as CEOs of Amazon and Microsoft, respectively, previously led their companies' cloud divisions—a clear reflection of how central the cloud has become to their corporate priorities.

Among these firms, it was Amazon that pioneered the introduction of IT infrastructure as a rentable resource.[21] Amazon's online retail business initially ran on a monolithic software architecture. In line with then-standard engineering practices, all its features and operations were tightly connected within a single, massive code base, making it difficult to update or scale individual parts without affecting the entire system. Simplifying a more complex technical story,[22] in the early 2000s, Amazon transitioned its code base to a modular (or "service-oriented") architecture, where different components (e.g., the shopping cart or the payment system) were separated into independent services that communicated with each other through standard interfaces. The key breakthrough that laid the groundwork for the cloud business model was extending this modular approach to the company's IT infrastructure. By decoupling its IT infrastructure from its e-commerce operations and making it accessible and controllable through software, Amazon transformed it into a stand-alone digital service, which it began renting out to external customers in 2006 under the name Amazon Web Services (AWS).

Client firms turn to cloud providers primarily for two reasons: cost savings and flexibility. Previously, most major firms across the economy built and maintained their own data centers—known in industry speak as "on-premise" infrastructure—to manage their IT needs. By renting servers (via the internet) from cloud providers, these firms relieve themselves of the need to operate and maintain their own servers (which involves the cost of the equipment itself, the cooling systems, and the IT personnel to run them). Client firms also gain crucial flexibility; they can "dial up" or "dial down" the computing capacity on an as-needed basis, enabling a "pay-as-you-go" model that eliminates the need for heavy upfront investment in on-premise systems.

Meanwhile, cloud providers can pool demand across many customers with different usage patterns, allowing them to smooth out spikes and troughs, keeping the shared infrastructure running closer to full capacity at all times. The result is higher efficiency (and lower costs), which is passed on to client firms. This shift from purchasing and maintaining on-premise data centers to renting from cloud providers can be seen in the sharp increase in spending on cloud-owned data centers relative to on-premise data centers, as illustrated in Figure 1.

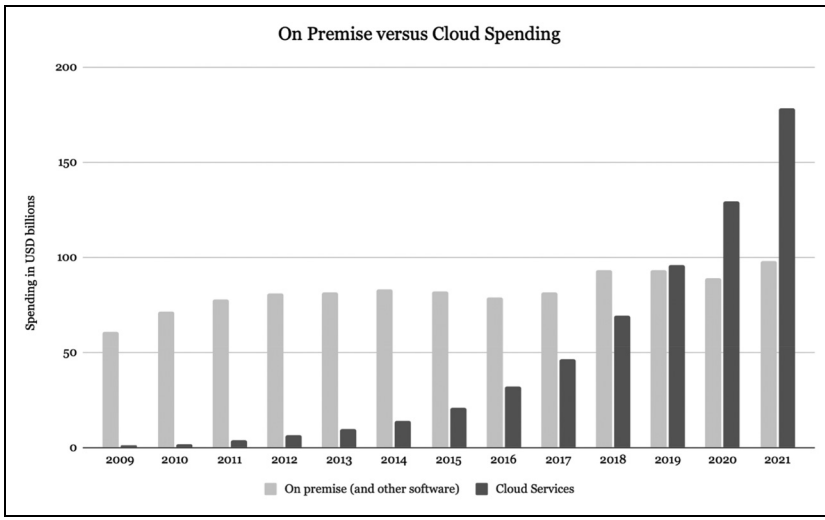


Figure 1. IT spending on cloud versus on-premise infrastructure. ***Source:*** Statistica.com, "Enterprise Spending on Cloud and Data Centers by Segment from 2009 to 2024 (in Billion U.S. Dollars)," February 6, 2025.

In the years immediately following AWS's launch, other firms—including not just Google and Microsoft but also IBM, AT&T, and Oracle, as well as smaller upstarts like Dimension Data, Joyent, and GoGrid—rushed into the market to get in on the action. Providing this type of infrastructure—essentially rented servers for computation, disks for storage, and the networking for connection—is capital intensive. But it is a relatively low-value-added service, and as such offers only relatively low margins. Since the basis of competition was cost, leading firms sought to gain an advantage by dropping prices and eating massive losses to try to box out competitors. In 2014–16, cloud providers were locked in an ongoing price battle;[23] an announcement of price cuts by one vendor would inevitably be countered days later by the other.[24] The result was an industry shake-out that eliminated some of these competitors (including AT&T, Dimension Data, Joyent, and GoGrid). This is when the cash-rich "big three" (Amazon, Microsoft, and Google) pulled ahead of the pack, with Oracle emerging as a later but increasingly strategic entrant, leveraging its political connections and enterprise foothold to stay competitive.

Value-Added Services

Oligopoly control resulted in less price competition, but the continued narrow margins on basic IT infrastructure prompted the top firms to shift their focus to higher-value-added services that they could pitch to their largest corporate clients in order to increase rents through customization. The initial offerings (known in

the industry as "platform as a service" or PaaS) sat just one layer above the infrastructural level and were neither task nor industry specific, for example, helping customers automatically scale the underlying compute and storage based on traffic; or providing new database products that automatically duplicate across multiple geographies in order to minimize data loss and downtime (i.e., if one data center fails, a backup in another location takes over); or automatically distributing a large computationally intensive task across a pool of servers (e.g., rendering an animation in parallel on dozens of servers instead of processing it sequentially on just a single one). More task-specific offerings were then built on top of these services, further abstracting away from the underlying infrastructure—for example, providing data analytics platforms that can help customers make sure that their compute resources are used efficiently or powering the primitive chatbots with which we interacted a decade ago.

The move toward more specialized offerings also prompted these firms to develop a network of software firms that could build new services on top of the basic infrastructure and beyond these generic- and task-specific services. Through such partnerships, cloud providers then added industry-specific capabilities in data management, customer relations management software, and Internet of Things (IoT) services, among others. With their partners, they also co-developed specialized software solutions to make it easier to run complex and computationally demanding industry-specific tasks such as protein folding simulations, 3D video rendering, or gas exploration. Some of these applications, in turn, demanded new hardware, pushing cloud providers to source from more sophisticated hardware makers (and in some cases, even co-develop new hardware with them).

Meanwhile, AI (which at the time was more commonly known as machine learning) also became an increasingly important part of these higher-value-added services that cloud firms could offer their clients. As AI technology became mature enough to be commercialized (roughly in 2012), cloud providers began adding a portfolio of task-specific AI models, aimed mostly at enterprises rather than individual consumers: for instance, computer vision for face detection and industrial maintenance, speech-to-text to transcribe meetings and to power voice assistants, recommendation systems for social media and streaming platforms, and anomaly detection for fraud and medical monitoring, to name a few.

By the late 2010s, AI had in fact come to be seen as integral to the cloud business model. In 2018, for example, Microsoft renamed its cloud division from the "Cloud and Enterprise Group" to simply "Cloud and AI," reflecting the growing centrality of AI. All three of the top players now offer corporate clients a suite of AI services tailored to their specific needs: For example, Microsoft Azure developed the AI instrument for Scandinavian Airlines to detect fraud quickly in their loyalty program;[25] AWS co-developed with Toyota a predictive maintenance AI system to predict when factory equipment will fail;[26] Google provided IKEA with an AI model to serve up personalized recommendations on their website to increase company sales.[27] As a result of these developments, client firms were now using the cloud, not to cut costs but to explore new business opportunities, optimize productivity, and innovate.

The Latest Chapter: The Race to AGI

The release of ChatGPT in 2022 set off a race for AGI that has reinforced and accelerated the underlying dynamics of the cloud business model, above all intensifying the drive for scale and an ever-more asset-heavy business model. Since then, AI development at the frontier has moved away from the proliferation of task-specific models and become dominated by a single model type: large language models (LLMs). Whereas traditional models excelled at narrow tasks but required tailored training on specific, often labor-intensive, human-labeled datasets, the defining feature of LLMs is their generality—the ability to perform a wide range of tasks across domains without customization—which AI researchers see as the key toward building an all-powerful artificial *general* intelligence.

The claim in the industry is that AGI will be capable of performing any intellectual task that a human can and will surpass human capabilities across virtually all domains. The technology, according to its most fervent proponents, is thought to be so powerful that once attained, it will be able to improve upon itself, triggering a feedback loop of rapid self-improvement beyond human control or comprehension—the impacts of which will be so profound that it will lead to economic abundance and sweeping societal transformations. By this logic, the first country to attain AGI will gain such an insurmountable lead that it would cement its position as a technological hegemon.[28]

What distinguishes LLMs from previous-generation AI models is that scale is the largest contributing factor to improving its performance.[29] So whereas traditional, task-specific AI models could often be trained on a single server or a small cluster of servers, the new AI models, of which ChatGPT is one, require entire data centers for training. Indeed, this new class of AI models are compute- rather than IP-constrained; that is, the surest way to increase performance is to scale up rather than to rely on new innovations in the model architecture. The landmark study on which ChatGPT and other LLMs were based was released in 2017,[30] but it was only much later that computing became cheap and abundant enough to train the LLMs of the quality that we see today.

In this way, LLMs have made AI development even more contingent on the cloud. Because scale is key, these new frontier AI models can *only* be built on the kind of dense agglomerations of compute capacity that the cloud offers. Thus, having access to capital—the physical infrastructure—to train and run the largest models with the most parameters has become the central competitive advantage in the new AI era. The race is now basically one of computational scale—whoever has the most available compute is seen as best positioned to win in the race to AGI. Amazon's plans to build thirty data centers on 1,200 acres of farmland in Indiana—at a cost of tens of billions of dollars—or Microsoft's astounding $80 billion commitment to building data centers in 2025 alone are but the latest examples of the current dynamics.[31]

Meanwhile, however, the task-specific AI of the pre-ChatGPT era is very much alive and diffusing rapidly in client corporations in a growing range of sectors. In this way, cloud firms continue to shape and support profound digital transformations across the economy. Factories increasingly use various computer vision models

throughout their production systems; social media/streaming platforms continue to perfect AI for recommendations to increase user engagement; basic computer vision is being applied in an ever-wider range of scenarios including unlocking personal devices and autonomous driving, to name a few. The features of the new cloud business model on which these developments are founded are radically different from those of first-generation platform firms on a number of dimensions. The following section elaborates these differences, showing how the model evolved from asset-light to asset-heavy, from hierarchical to modular in structure, from dominating to collaborating with clients and partner firms, and from consumer- to enterprise-facing.

The Cloud Business Model

Asset-Heavy, Not Asset-Light

In order to provide the infrastructure capacity for a growing base of customers, cloud providers committed huge sums of capital to building massive data centers across the globe. The cost of a modern data center is extraordinarily high—far higher than the typical amount of capital that a startup can raise through venture capital. Amazon, for instance, spent $11.8 billion on just five data centers in Oregon in 2022 (roughly $2.4 billion each).[32] Data centers that can train the latest AI models cost even more as they have a far greater density of expensive NVIDIA GPU chips (the latest NVIDIA GPU for AI, the Blackwell B200, costs between $30,000 and $40,000 per chip).[33] Previous-generation data centers used to be filled with lower-cost CPUs and had lower energy and cooling requirements; now the new data centers that power AI are full of expensive GPU chips, and have far higher energy and cooling requirements.[34]

The three leading cloud firms (Amazon, Microsoft, and Google) have splurged on investments in physical infrastructure to fuel their rise. Together they are planning to spend well over $300 billion in 2025 alone, far more than the heavyweights in traditional sectors such as oil and gas. Figure 2 shows that the capital expenditure of all three took off when each first adopted the cloud business model. Similar to the platform model, the path to profitability for these infrastructure investments was uncertain and long-term, but for a different reason: not network effects but upfront capital investment costs. It took Amazon.com six years to turn a first (modest) profit; AWS took nine years. Microsoft became profitable only in the late 2010s.[35] Google Cloud reported a profit for the first time in 2023.[36]

These massive capital investments reflect a shift from the centrality of data to the preeminence of compute. Platform-era frontier firms were asset-light because storing vast quantities of data is cheap, allowing them to scale to millions of users with relatively low levels of capital expenditure. Having the best, most efficient physical hardware for data storage had nothing to do with a company's advantage because competitiveness depended more on the scale and sophistication of their data. Facebook, for instance, grew fabulously profitable because they had a vast trove of

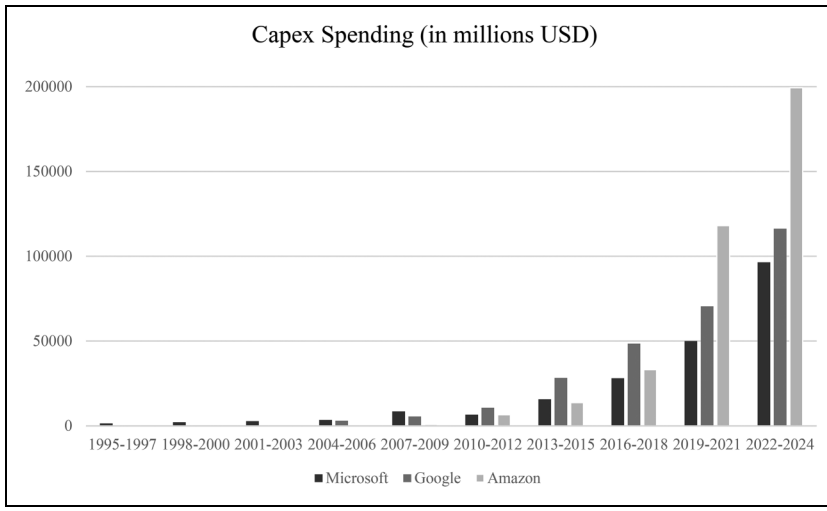


Figure 2. Microsoft, Google, and Amazon's Capex. ***Source:*** Authors.

user data that allowed them to effectively serve advertisements to some four billion users—all without the need for massive spending on IT infrastructure.

By contrast, cloud providers do not benefit from harvesting large quantities of data in the same way. Instead, they provide clients with compute infrastructure and compete on their ability to help their customers innovate, which increasingly requires running sophisticated workloads on dedicated computing resources. And as noted, they have become the primary avenue for developing and diffusing AI technology, which requires highly specialized, high-end hardware (GPUs), and can involve redesigning entire data centers (e.g., rewiring racks of servers with higher interserver bandwidth). As a result, cloud providers are constantly seeking the most advanced and capital-intensive hardware.

The extent to which the cloud business model is rooted in physical infrastructure cannot be overstated. The core of this physical infrastructure is the data center. Data centers are in effect massive, oftentimes football field–sized buildings that are filled with racks of servers, of which the most expensive components are usually the chips (this is especially true with modern data centers that are filled with expensive chips capable of training large AI models). Previously, cloud providers leased data centers from real estate investment trusts (such as Digital Realty and Equinix). But as the efficiency and capabilities of data centers have become part of these firms' competitive advantage, they increasingly shifted toward building and owning their own.

However, unlike the Fordist firm of yesteryear that was both capital *and* labor intensive, once a data center is constructed, maintaining it takes very few technicians. But it requires vast amounts of water and energy. When servers run at maximum capacity—as they do when training an AI model—they can overheat, resulting in performance

degradation or even system crashes. So to mitigate that risk, data centers circulate water through specialized cooling systems to bring the server temperatures back down.

Access to energy is also crucial, so many cloud providers have initiated collaborations with electric utility companies to ensure their energy demands are met (nine of the top ten US electric utilities say data centers are a main source of customer growth).[37] With surging demand for AI and an energy grid that is increasingly under stress, these companies are now going a step further, building their own power sources in tandem with and directly adjacent to their data centers.[38] Amazon, Google, and Microsoft have already invested in hundreds of solar and wind projects to power their data centers. They are even restarting and expanding energy sources: For example, Microsoft has invested in a shuttered unit at Three Mile Island nuclear plant;[39] Amazon has invested in X-energy, a company developing nuclear energy, and partnered with Energy Northwest to build a data center campus next to the Susquehanna nuclear facility in Pennsylvania.[40]

Semivertically, Not Horizontally, Integrated

A second major difference to the previous platform model relates to corporate structure. In the platform era, the decoupling of software and hardware allowed platform firm applications (such as the Airbnb app, WhatsApp, or Netflix) to operate across a wide array of devices, whether on Windows, Linux, or Mac, Android or iOS. By decoupling the two, firms prioritized user reach and interoperability over the deep integration and optimized performance that comes from tightly coupled systems. This approach reflected the goals of platform firms—to quickly saturate markets and achieve network effects—and their willingness to trade profits for horizontal expansion.

By contrast, in the cloud business model, competitive advantage is increasingly defined by the recoupling of software and hardware, with new hardware optimized to run highly sophisticated software workloads. Initially, cloud providers operated on the logic of the platform era, filling their data centers with cheap and generic hardware—and it did not matter because any software was designed to run on any hardware. They sourced most of their components—the chips, the servers, the racks that servers sit on, the cooling system, and the plethora of networking equipment—in the market, while still retaining control over the infrastructure itself, to maximize their ability to coordinate complex systems.

But when cloud firms moved beyond low-margin infrastructure to offer differentiated, specialized software capabilities (such as AI), many of these higher-value-added services required tighter integration of hardware and software up and down the "cloud stack," that is, the layers of technology that make up the firm's IT infrastructure, from high-level software applications down to hardware components like semiconductors. Cloud providers have thus increased their role in the entire data center value chain, using their own designs and hardware rather than buying components off the shelf. For instance, they have developed their own cooling technology, implemented energy

efficiency improvements, and partnered with server rack makers to manufacture custom rack designs.[41]

When it comes to complex workloads such as AI, the ability to optimize hardware requirements for specific software demands is a hallmark of the cloud business model. Google was the first to align its software and hardware around AI. In 2015, the company introduced a software programming language called TensorFlow, dedicated to developing AI models. To enhance TensorFlow's performance, Google also released Tensor Processing Units (TPUs), specialized data center chips that maximize the efficiency of TensorFlow code. This tight integration between software and hardware allowed Google to boost the efficiency and speed of its AI services and research efforts. Other firms have similarly moved to developing their own chips, among other reasons, to reduce their reliance on external suppliers like Intel, AMD (Advanced Micro Devices), and, most of all, NVIDIA. For example, Microsoft has recently created its own AI-specialized chips, which, according to a Microsoft-authored blog post, "nestle onto custom server boards, placed within tailor-made racks that fit easily inside existing Microsoft data centers [allowing] hardware [to] work hand in hand with software."[42]

In short, the emerging model differs from the arm's-length, market-based contractual arrangements that have become the de facto interfirm relation in the knowledge economy. Although they are capital intensive, cloud providers are not vertically integrated in the traditional Fordist sense either. In theory, ownership of every layer of the cloud stack would allow firms to deliver the highest-performing solutions because coordination occurs within a single company rather than through the market. However, innovation in all layers of the cloud stack is advancing rapidly. For example, the best hardware technology in one period can quickly lose its dominance in the next. Consequently, the challenge for the cloud providers is to integrate with the best technology in any given layer in the stack, while maintaining their ability to coordinate integration across the stack.

To solve this problem, cloud providers are not only integrating vertically but are doing so *modularly* through the relationships that they cultivate with multiple players at each layer of the stack, that is, with sophisticated software vendors up the stack and with the makers of the most powerful processing chips down the stack. The nature of these relations can vary significantly in form and intensity. In some cases, for example, where software applications are easy to install, the relationship is a simple technical matter, and does not require extensive coordination between cloud providers and their partners. But in other cases, cloud providers assign staff to sit in the offices of their partner firms to ensure quality integration or conduct joint research. In still other cases, partnerships are deemed so valuable that cloud providers invest in their partners or acquire them outright to keep them from integrating with competing cloud providers. But even in the case of acquisition, the approach that cloud firms take represents a stark departure from the platform era. Acquisition decisions by previous platform firms were driven by the desire to expand the user base and achieve network effects or neutralize competitors (think of Facebook's acquisitions of Instagram and WhatsApp). By contrast, cloud providers make acquisition decisions to gain control over critical

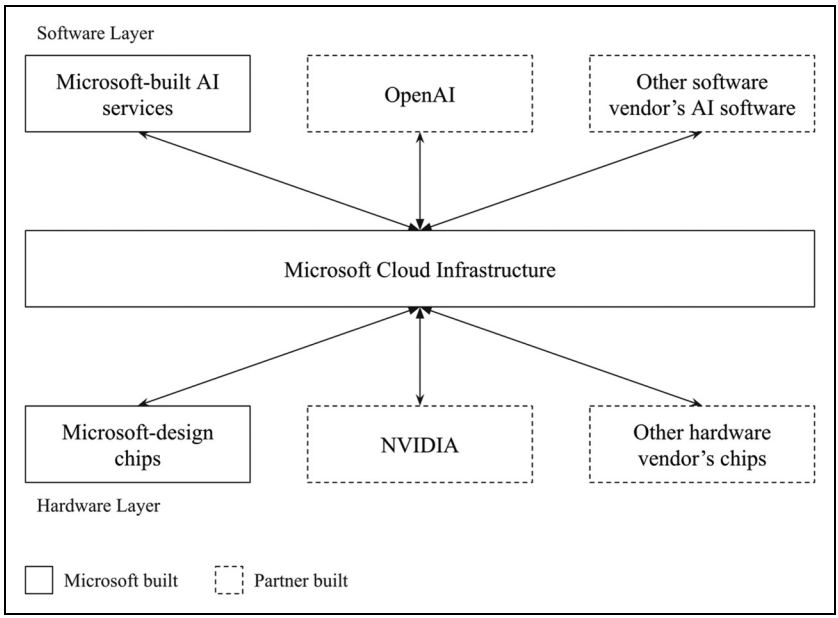


Figure 3. Example of Microsoft's "modular" integration of AI services. ***Source:*** Authors.

technologies and components that enhance their ability to deliver differentiated services. Figure 3 is a stylized illustration of the hardware and software components in Microsoft's AI stack.

Collaboration, Not Domination

In the fissurized, franchised version of the knowledge economy, interfirm relationships are characterized by lead firms dominating those lower in the supply chain.[43] For instance, Apple sets the terms of its partnership with hardware manufacturer Foxconn, which in turn sets its own terms with its lower-tier manufacturing subcontractors. The result is a cascading pattern of domination, with the highest-value producers exploiting "subordinate" firms with lower productivity. Platform firms also tend to dominate the relationships they enter, as for example in the relationship between Uber and the drivers who depend on it to connect them to riders.[44]

The cloud business model breaks from this pattern, as the relationship with their various partners spans a continuum from domination to collaboration. Cloud firms do dominate in the relationship with some partners. For instance, their relationships with suppliers who make server racks or with the utility companies that supply them with electricity resemble the hierarchical relationships that lead firms have with their suppliers as described in the global value chains literature. However, their relationships to other partners are often much more equal, characterized more

by mutual benefit and sometimes even mutual dependence. The relationship between Microsoft and OpenAI is one example, because OpenAI needs Microsoft Azure to train and run its models and because Microsoft's own models are not (yet) as powerful as OpenAI's. The same is true for the relationship between Microsoft and NVIDIA—indeed here, if anything, Microsoft needs NVIDIA (whose chips are far superior to its own) more than the other way around (since many firms are clamoring for NVIDIA's chips). Such patterns of collaboration apply more broadly, both to downstream partnerships with hardware manufacturers and to upstream partnerships with software providers, and sometimes even extend to (more limited forms of) horizontal collaboration with competing players.

Down the stack, cloud providers partner with some of the most technologically advanced companies in the world. Chips are the data center's most complex, technically sophisticated, and capital-intensive component. So the most significant partnerships at the hardware layer are with semiconductor manufacturers such as Intel, AMD, and NVIDIA. These relationships have enabled cloud providers to access the latest hardware innovations and collaborate closely on optimizing the performance of new hardware in their data centers. For example, NVIDIA's GPUs have become the cornerstone of AI development and the new race toward AGI through collaborations with companies like Microsoft and Amazon. Microsoft has also teamed up with AMD to develop specific AI processors;[45] they have also partnered with Cray, a subsidiary of Hewlett Packard Enterprise, to bring the company's supercomputers used for climate modeling, precision medicine, and other scientific research to Microsoft Azure's clientele.[46]

Such partnerships demand deep engineering capabilities on both sides, and can often also involve joint research and development efforts. To stay with the previous example, Microsoft (even while developing their own AI chips) has worked closely with NVIDIA to build a massive cloud-based AI supercomputer.[47] *Business Insider* estimates that Microsoft alone accounted for 19 percent of NVIDIA's revenue in 2023.[48] Collaborative research and development can even lead to significant advances in computational capabilities. For instance, Quantinuum and Atom Computing, both manufacturers of quantum hardware, joined forces with Microsoft to achieve new breakthroughs in the field of quantum computing.[49]

Up the stack, cloud firms rely on partnerships with software makers such as SAP, Salesforce, Databricks, Anthropic, and OpenAI, known in industry-speak as Independent Software Vendors (ISVs), to expand their portfolio of services. ISVs build on top of the basic cloud infrastructure that cloud firms provide to create a wider range of applications and services from which client firms can select. Sometimes the services offered by ISVs are as simple as helping client firms move to the cloud by re-creating legacy applications that were originally designed to run on on-premise infrastructure. For instance, the German enterprise software maker SAP offers versions of its software that allow customers to continue to run SAP software as they migrate to cloud-based data centers. In other cases, ISVs provide entirely new services on the cloud. For

instance, Databricks—which runs a cloud-native big data processing service—partners with all three cloud providers to offer a scalable big data analytics platform.

In general, ISVs allow cloud firms to meet a broad range of customer needs without having to develop all these services in-house. In the new AGI era, OpenAI's partnership with Microsoft is perhaps the most prominent example.[50] Microsoft initially invested in OpenAI in 2019, giving the latter access to the computational capacity required for training AI models. In the early years of their agreement, Microsoft's cloud computing platform (Azure) became OpenAI's exclusive cloud provider. Relationships like this can also push cloud providers to deepen their infrastructure capabilities. For instance, to meet OpenAI's computational needs, the two firms worked closely to construct AI supercomputing infrastructure that was used to train OpenAI's cutting-edge models.[51] Training massive AI models requires advanced supercomputing infrastructure or clusters of state-of-the-art hardware connected by high-bandwidth networks. The supercomputer built for OpenAI provided the infrastructure for the breakthrough models (e.g., DALL-E and ChatGPT) that the AI startup developed.[52] And integration of OpenAI's advanced AI models with Azure now allows Microsoft customers to seamlessly incorporate these models into existing solutions already built on Microsoft's cloud.

The most technically advanced ISVs, such as OpenAI, may have outsized power in their relationship with cloud providers. However, for other ISVs, the cloud providers play an important role in setting agendas and optimizing the contributions of their partner firms. Microsoft, for example, hosts an annual conference called Microsoft Ignite, where Microsoft showcases its latest innovations, infrastructure developments, and future roadmap, providing ISVs with cues on where to focus future work. For example, Microsoft may announce new hardware offerings at such conferences (or invite their downstream partners to do the announcement) to encourage upstream ISVs to integrate them into their product development; Microsoft may also promote new features within their AI platform, sending a clear signal to ISVs to augment their services with AI. Google and Amazon also host a similar annual conference for their ISV partners. Having these insights allows ISVs to align their development efforts and ensure compatibility with new developments in the cloud infrastructure. In this way, cloud providers empower their partner ecosystem while reinforcing their own position in the core.[53] Beyond this, cloud providers also coordinate vertically across their upstream and downstream partnerships. For example, Microsoft responded to OpenAI's demand for specialized chips by mobilizing hardware partners such as NVIDIA to ramp up the supply of chips designed for AI model training.[54]

While competition between leading cloud providers like AWS, Azure, and GCP is fierce, there is also an undercurrent of collaboration in some areas where mutual benefit can be realized. Unlike the platform era, first-mover advantage is not necessarily decisive. Enterprise customers almost always adopt multiple cloud providers (as high as 81 percent of firms do so, according to a recent Gartner survey)[55] to diversify their risk and capitalize on the best-of-class solutions that different cloud providers offer. Thus, certainly at this juncture, the true opportunity for all these firms lies in expanding the

market generally by broadening the adoption of cloud services and AI across industries, to the benefit of all of them. For instance, when Capital One Bank transitioned its IT infrastructure to AWS, it was not just a win for Amazon. It also served as a validation for the entire public cloud model, indirectly benefiting Azure and GCP by signaling to other financial institutions the viability of cloud adoption. In a similar vein, certifications and trainings provided by AWS benefit other cloud providers because many of the skills are industry- and not firm-specific.

Collaboration on setting industry standards can also have positive-sum impact. Establishing common standards is essential for ensuring interoperability, security, and reliability across different cloud environments. For instance, the Cloud Native Computing Foundation (CNCF) and other open-source initiatives have brought together engineers and executives from various cloud providers to work on cloud-agnostic projects like Kubernetes, which has become the modern standard for deploying software applications across cloud platforms because (simplifying greatly) it allows software workloads to be packaged into self-contained units that can be consistently and reliably executed on across different computing environments. These dynamics underscore the importance of expanding the "total pie" of cloud computing rather than just competing for a larger slice of what already exists (cloud spending surpassed $660 billion in 2023, yet it remains only a fraction of the $4.7 trillion in IT spending from the same year).[56]

Finally, horizontal coordination can be defensive, serving as a strategic counterbalance against partners in the cloud stack who hold outsized bargaining power. An example of this is a consortium recently formed by Google, Microsoft, Meta, AMD, Intel, Broadcom, and others to establish a new industry standard for the components that link together AI accelerator chips (i.e., GPUs) within data centers, enabling the scaling of AI systems.[57] NVIDIA, which currently dominates this standard, was deliberately excluded from the consortium. This move is part of a broader strategy to wrest control away from NVIDIA, whose dominance gives it significant leverage over the AI hardware ecosystem.

Enterprise-Facing, Not Consumer-Facing

A last key difference to the platform business model is the customer base. Whereas "traditional" platform firms are predominantly consumer-facing, cloud firms are almost exclusively enterprise-facing. While an individual person can rent a server with any of the cloud providers, there is hardly a use case for nonbusinesses to do so. Instead, cloud providers rely not on millions of individual users but on developing steady revenue streams from a much smaller set of enterprise customers. With large clients (e.g., a *Fortune* 500 firm or a government agency), a single contract could be worth hundreds of millions of dollars in cloud spending. Because they can generate revenue regardless of the extent of their market dominance, cloud providers focus on product differentiation and technological superiority rather than direct competition for market share.

This orientation results in corporate strategies that differ sharply from the previous platform model of cultivating superficial connection to a wide consumer base. For cloud providers, developing deep and enduring relationships with customers is a priority. It is also a critical strategic resource because their ability to innovate and stay competitive depends on the circulation of knowledge from corporate clients back to the R&D staff. The ability to tap into outside sources of knowledge—known in the innovation literature as a firm's absorptive capacity—is critical to the success of the cloud business model. Intimate familiarity with their clients' operations allows cloud providers to develop industry-specific services and product innovations, ensuring that they can tailor their offerings to meet the unique needs of their clients. According to Cohen and Levinthal, a firm's absorptive capacities depend not only on basic skills or shared technical language but also awareness of rapidly evolving developments in the field.[58] Such awareness is essential for recognizing the value of new information, assimilating it, and applying it to new domains to drive innovation. It allows firms to make informed decisions about what partnerships to forge and what innovations to pursue.

To build and maintain strong relationships with their customers, cloud firms rely on entire subgroups within the organization—such as sales teams or "customer experience engineering" departments—to manage these relationships. They spend a significant portion of their hiring budgets on technically adept sales staff to foster close relationships with current and potential clients. In 2023, AWS had over 60,000 employees focused on customer relations, accounting for roughly half the total number of AWS employees.[59] This army of sales staff operates as middlemen between clients and the cloud provider's R&D staff, helping not only to manage their relationships but also to ensure that the increasingly sophisticated services on the cloud are adopted by clients and that feedback makes it into product roadmaps. Relationships with large customers can be so important that senior executive staff sometimes get directly involved in managing them.

Besides hiring an extensive and technically competent sales staff, cloud firms also rely on deep partnerships with professional services organizations. Such firms provide consulting services to help corporate clients adopt and integrate cloud technologies into their business operations. Like sales staff, these professional services firms also facilitate the flow of information between customers and cloud providers, which informs the cloud provider's product development and R&D decisions and helps tailor cloud offerings to industry needs.

These firms bridge a critical gap for cloud providers because of their dual expertise in both cloud computing and the specific industries they serve. For instance, Accenture, a global professional services company, has played an essential role in guiding its client base—many of which are legacy companies—through the complex process of migrating to the cloud. Accenture can also help develop bespoke cloud solutions for their clients. For example, Accenture worked with the Schaeffler Group, a German automotive and industrial supplier, to build a new cloud-based industrial automation solution using robotics.[60] Accenture, in this project, was responsible for implementing

this solution, which it could do because of its dual expertise in the cloud and digital engineering.

Professional services firms often have deep relationships in specific industries, so partnering with such firms allows cloud providers to expand their reach to new sectors or deepen their penetration of existing ones. To stay with the Accenture example, the company has extensive networks and expertise across multiple industries, including banking, automotive, chemicals, retail, health care, and manufacturing, among others, as well as a dedicated business group called Accenture Cloud First, whose 70,000 employees produced $11 billion in cloud-related revenue for the company in 2019.[61] Because these partnerships are critical for adopting cloud services, cloud providers build close relationships with these firms; in 2019, Microsoft and Accenture teamed up to launch the Accenture Microsoft Business Group, a 45,000-person initiative focused on jointly delivering cloud-powered software solutions to their global client base.[62]

The cloud business model marks a radical departure from its platform predecessor that has conventionally been associated with fissurization and deskilling (e.g., outsourcing, offshoring, and utilizing low-skilled gig workers). This new model creates upward pressure on the demand for highly skilled technical workers in client firms to allow them to take full advantage of the cloud. For instance, in order for customer firms in sectors like health care, finance, and manufacturing to take full advantage of what the cloud has to offer, they must have employees who are equipped with advanced technical skills, for example, the ability to migrate legacy, on-premise IT infrastructure onto the cloud. This trend is reflected in the proliferation of cloud-related certifications. Programs like AWS Certified Solutions Architect, Microsoft Certified Azure Administrator, and Google Professional Cloud Architect are becoming increasingly important and widespread across industries. These trainings are used by the internal sales staff of cloud providers, professional services firms like Accenture, and customers themselves. Such certifications are not merely professional credentials but essential tools that equip workers with the technical skills required to take advantage of the cloud's capabilities.

In sum, the cloud business model departs, on multiple dimensions, from the platform model out of which it grew. Its emergence at the frontier of the knowledge economy has also unleashed a host of new dynamics, both in the market and in politics. The next section considers some of the most important of these.

The Cloud Business Model: Impact and Effects

Technological Dynamics: The Race to AGI

According to its most zealous advocates, the promise of AGI is that it will solve the greatest challenges of our time—from mitigating climate change to unlocking entirely new scientific breakthroughs to dramatically increasing economic abundance.[63] At this juncture, however, the concept of AGI is less a technical milestone than it is an

imagined socioeconomic future or—more practically—a marketing concept. The narrative is that achieving AGI—the inflection point at which AI will forever transform civilization—is ever imminent but never here. This always-on-the-horizon narrative is what sustains market momentum and justifies the hundreds of billions of dollars that the top cloud firms are pouring into compute infrastructure.

Yet even as the ChatGPT moment and the drive toward AGI has produced a new (almost pyramid-scheme-like) logic of investment, a killer application, significant commercialization opportunities, or meaningful productivity gains have yet to materialize. What commercialization has occurred remains far from justifying the capital being poured into scaling the technology. Last year, for instance, OpenAI collected $3 billion in revenue but spent $7 billion (a large portion of which went to training and deploying their models).[64] This raises the possibility that the hundreds of billions of dollars invested in compute to power the current AI paradigm may look more like a speculative bubble than the foundation of the next technological revolution.

This is why DeepSeek was such a threat and caused markets to dip in early 2025. The cost optimizations achieved by DeepSeek's model undermined the scaling laws underpinning the AGI race, making the massive investments in building compute infrastructure look unnecessary. Many thought the breakthrough would pop the bubble. But the scare was only temporary. Just as nineteenth-century economist William Stanley Jevons observed that technological improvements in coal efficiency led to increased —rather than decreased—coal consumption by unlocking new demand (a phenomenon later known as the Jevons Paradox), the efficiency gains achieved by DeepSeek's AI models are thought to similarly boost rather than diminish demand for cloud infrastructure. This was in fact Microsoft CEO Satya Nadella's argument—that far from threatening the cloud business model, DeepSeek's breakthrough would drive even greater demand for compute, as cheaper and more efficient models unlock new demand for AI.[65]

It is too early to predict the precise trajectory of the AI "revolution." Whether the current "bigger is better" paradigm will remain dominant or whether we will see a reversal to more task-specific AI models is an active debate within the AI research community.[66] What is already certain is that the handful of dominant cloud providers will continue investing aggressively in support of the scale hypothesis, driven by the fear of missing out should the promised gains ever materialize. And at this stage, cloud providers have no choice but to perpetuate the AGI narrative, since taking their foot off the gas would undermine the legitimacy of the extraordinary sums of capital already sunk in.

Political Dynamics: The Rise of Techno-Nationalism

The rise of the cloud business model has also unleashed new political dynamics. One debate revolves around the capacity of the state to steer this new model and to mitigate its negative impacts. The asset-light nature of previous-generation platform firms allowed them to engage in venue arbitrage in order to circumvent regulations, which are typically enforced within state or national borders. Their mobility enabled them

to shift operations to jurisdictions with more favorable legal, tax, or regulatory environments, making them difficult to regulate effectively. By contrast, this kind of regulatory flexibility is far more challenging for cloud firms to achieve because their physical infrastructure ties them to specific locations. As we have seen, these companies have made heavy investments in the physical assets required to operate their data centers, up to and including the land on which the data centers sit and the energy sources needed to run them.

It seems clear that American tech leaders are quite aware of the ways in which the resulting vulnerability could give the state new leverage in shaping this next era of the digital economy. Pundits have offered a range of plausible explanations for the abrupt turn on the part of American tech titans toward embracing Donald Trump. Gone are the days when CEOs of the top companies openly challenged Trump policies and even banned the then-former president from participating on their platforms.[67] Most accounts of this turnaround stress political pragmatism (fear of reprisal), growing discontent with the policies of the Biden administration toward the tech industry, and a breakdown of the Silicon Valley consensus view of innovation and tech entrepreneurship as an unalloyed good.[68] To these we would add another that has received virtually no attention: namely, a wholly new situation in which previously footloose firms are now more invested in immovable assets—creating enormous material incentives for them to curry favor with national policymakers both domestically and abroad.

But the dependence runs both ways, as the government itself relies increasingly on the cloud. The United States has several multibillion-dollar contracts with cloud providers to modernize various aspects of the government, including the military. It also has enormous contracts with Palantir, which provides the advanced data analytics and integration for a wide range of federal agencies, in particular in defense, law enforcement, and intelligence. Palantir does not have its own data centers, but instead relies on partnerships with the dominant cloud providers analyzed here. And although government reliance on Palantir pre-dates the current administration, the scope of that reliance has expanded massively since Trump took office.[69]

Perhaps more importantly, geopolitical dynamics have if anything deepened the mutual dependence of the government and the top cloud firms. Winning the so-called AI arms race with China had become a top priority for both major US political parties even before the more recent race toward AGI, which has only intensified the geopolitical competition. In 2018, the Congressional House Armed Services Committee turned to former Google CEO Eric Schmidt to chair a National Security Commission on Artificial Intelligence. The commission, which also included top executives from Microsoft and Amazon, was tasked with making recommendations to "advance the development of artificial intelligence, machine learning, and associated technologies to comprehensively address the national security and defense needs of the United States."[70]

Under the Biden administration, one of the signature pieces of legislation—the CHIPS and Science Act—included generous funding to promote the manufacturing of the type of advanced semiconductors on which cloud providers rely. Even energy

investments, like those in the Inflation Reduction Act aimed at expanding renewable energy capacity, would ultimately benefit the power-intensive operations of major cloud firms. A late-term executive order called for measures to "advanc[e] United States leadership in artificial intelligence infrastructure," including making federal land available for building data centers, while however also calling for guardrails to ensure AI safety.

The current Trump administration, for its part, has exhibited more unbridled enthusiasm, rolling back the Biden-imposed conditionalities for state support that, for instance, enforced compliance with security measures and required cloud firms to use renewable energy sources. Trump has also endorsed a massive new wave of data center construction in the proposed "Stargate" plan, which earmarks over $500 billion for data center infrastructure, ensuring that there will be sufficient electricity to power the new data centers. At the center of the initiative is Oracle, the plan's key cloud partner—positioned to benefit directly from state-backed expansion as it competes with the other cloud giants. The administration clearly sees AI as a key strategic asset, as evidenced in the 2025 executive order "Removing Barriers to American Leadership in AI."

This dependence on the state has created a new and still evolving alliance between an increasingly nationalist political elite and AI firms including, above all, the cloud providers. Among the most surprising turns is the one that now has Donald Trump and OpenAI's Sam Altman on the same side (in 2017, Altman was among the first to vocally oppose Trump's nationalism, even writing a blog post to rally other industry executives to "take a stand" against the administration's anti-immigrant policies).[71] Cloud and AI firms are now much more amenable to the federal government's nationalist goals. Microsoft has, for instance, pulled its AI research operations from China and cut services to Chinese universities amid geopolitical concerns, throwing away decades of the company's efforts to globalize its operations.[72] This marks a sharp departure from the earlier era when US-based platform firms eagerly sought access to the Chinese market. So whereas the footlooseness of the previous era enabled platform firms to be much more globalist in their approach, the state-tech nexus that characterizes the politics of the new cloud business model atop which the "AI race" is waged results in a much more nationalist approach—as the fate of the cloud is now seen as more closely tied to the strength of their home country, and vice versa.

This new alliance between the state and the top cloud providers is problematic for other countries. Across the Atlantic, high levels of dependence on a handful of American cloud firms is increasingly viewed as a source of geopolitical vulnerability. After Trump's recent flare-up with the European Union, policymakers were reminded just how easily a hostile US administration could cut off access and shut down parts of the European economy, prompting renewed urgency and investment in cloud sovereignty across the continent.[73] Meanwhile, export limits on AI chips and other trade measures designed to hinder China's advancement in AI technology have only fueled the Chinese government's commitment to achieve technological independence and thus to intensify support to Chinese firms—including, notably, domestic cloud providers.[74]

Distributional Dynamics: New Centers of Power

Finally, the cloud business model has also given rise to new distributional dynamics that are quite different from the earlier platform era. Platform firms, as their founders never tired of emphasizing, "disrupted" legacy industries. Airbnb, for instance, upended the traditional hotel sector by turning any homeowner into a potential bread-and-breakfast host. Amazon's e-commerce business supplanted much of the brick-and-mortar retail sector by moving shopping online. Fintech startup Robinhood redefined retail investing by offering users commission-free access to financial markets through its easy-to-use mobile app. In each of these cases, legacy industry incumbents were disrupted by new platform firms emerging from outside of each industry's established core.

In contrast, and because of its enterprise-facing nature, cloud providers have the potential to help industry incumbents stave off competition, especially from technologically savvy disruptors. Rather than posing a threat, the services that the cloud offers provide the most dominant legacy firms with the means to undergo digital transformation themselves and increase their own competitiveness. For instance, Walmart has become a seasoned user of Microsoft Azure and its value-added services, allowing it to compete with Amazon in online retail. Banks such as JPMorganChase are also fast adopters of the cloud, helping them stave off competition from fintech upstarts. So, whereas the platform business model wrests power away from legacy industry incumbents and toward Silicon Valley, thereby widening the digital gap across industries (between the tech sector and the rest), the cloud business model may reverse this trend.

As should be clear by now, the client firms that benefit the most from the cloud business model are those with robust partner networks and the technical know-how to take full advantage of high-value-added cloud services. The most adept clients might even employ a fleet of technical staff with cloud-based certifications. This means that, within industries, we should expect the rise of the cloud business model to deepen inequality since only the top players (e.g., JPMorganChase or Walmart) will have the capabilities to fully leverage the cloud. Less sophisticated players (e.g., local banks without the tech savvy to take advantage of the cloud; mom-and-pop retailers) will fall further behind.

The most notable new center of power of all lies with the cloud providers themselves relative to the rest of the tech sector. In the post-ChatGPT era, cloud and AI technologies have become firmly entrenched among the top players, such that innovation and its profits remain locked within a handful of existing cloud giants. Indeed, cloud firms are buttressing their competitive advantage by increasing the sophistication of the AI services they provide to their clients. And as they do so, they bring ever-greater portions of the AI supply chain under their control, everything from semiconductor hardware and specialized data centers to high-level AI services.[75] They are also dramatically increasing their asset intensity to provide the infrastructure to train and deploy the most advanced AI models.

This means that even fewer companies can participate in developing these AI models, let alone compete in the race to AGI. So whereas firms like Apple and Netflix could, in the past, develop their own task-specific AI models in-house (e.g., Apple's face recognition to unlock the iPhone; Netflix's personalized recommendations), even these highly tech savvy companies must now rely on the cloud to access cutting-edge models like ChatGPT. For instance, Apple plans to use these models to power its Siri upgrade;[76] Netflix is integrating ChatGPT into its search functionality to allow users to find content using conversational language.[77] As a result, the top cloud players have pulled far ahead of even the most advanced platform firms including Uber, Airbnb, and Netflix—all of which rely on cloud providers not only for their infrastructure needs but for integrating the latest advancements in AI. In sum, the cloud business model represents a reorganization—and intense centralization—of power and profit at the leading edge of both the tech sector and the knowledge economy—one defined by asset intensity and underpinned by state support and geopolitical competition.

Conclusion

This article has outlined a major evolution at the technological frontier of the knowledge economy: the emergence of the cloud business model. Departing from the previously dominant, consumer-facing, and asset-light platform model, cloud firms have become intensely asset-heavy, enterprise-oriented, collaborative, and vertically integrated operations. The top cloud players, which once thrived on data-driven, consumer-centric platforms, have increasingly shifted their focus toward providing sophisticated IT infrastructure services and specialized computational capabilities to a much smaller number of large enterprise clients. In doing so, these cloud giants have accumulated immense physical infrastructures—including data centers, specialized hardware, and even energy facilities—that constitute a much more capital-intensive digital economy. The rise of the cloud business model thus represents a new development that forces us to reconsider many of the features we have come to expect from the knowledge economy.

As we have seen, the cloud is not a passive backdrop for AI's rapid advancement; it is the foundational infrastructure upon which the entire AI ecosystem is being built and the key player in the race to AGI. Indeed, the prevailing belief that scaling compute infrastructure is the key to reaching AGI further reinforces and accelerates the underlying dynamics of the cloud business model described above, intensifying the drive for scale and ever-greater asset intensity. This evolution has led to the deep embedding of cloud providers at the center of the entire AI ecosystem—not only as the key actor in the race for AGI but, more importantly, as a critical "choke point" with the diffusion of AI across the political economy as a whole.

Unlike platform firms that thrived under a noninterventionist free-market regime where state inaction is the most beneficial policy, the cloud model signals the emergence of a new techno-nationalist alliance—one in which the state clears the way

for unbridled data center expansion at home while deploying export controls and sanctions to hobble competitors abroad. This diagnosis of the nature and origins of the cloud business model draws our attention to the emergence of a new and potentially enduring state-tech nexus—one that risks reinforcing an ever-more top-heavy economy, in which the transformative gains of the cloud and AI accrue overwhelmingly only to the most powerful corporate and political actors.

Acknowledgments

We are grateful to Chase Foster, Peter Hall, Jason Jackson, Ben Schneider, and Christine Trampusch for their valuable feedback and encouragement throughout the development of this article. Special thanks to Henry Farrell, whose insights and suggestions were especially instrumental in shaping the core arguments. We also thank the editorial board of *Politics & Society* for their thoughtful guidance during the review process. This work benefited from conversations at the Max Planck Institute for the Study of Societies in Cologne, particularly with Lucio Baccaro and Timur Ergen.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Notes

1. Herman Mark Schwartz, "From Fordism to Franchise: Intellectual Property and Growth Models in the Knowledge Economy," in *Diminishing Returns*, ed. Lucio Baccaro et al. (Oxford University Press, 2022).
2. See, e.g., Martin Kenney and John Zysman, "The Rise of the Platform Economy," *Issues in Science and Technology* 32, no. 3 (2016): 61–69; Nick Srnicek, *Platform Capitalism*, Theory Redux (Polity, 2017).
3. See, e.g., *Competition Is for Losers with Peter Thiel (How to Start a Startup 2014: 5)*, 2017, https://www.youtube.com/watch?v=3Fx5Q8xGU8k.
4. See Rafe Uddin and Stephen Morris, "Big Tech Lines Up over $300bn in AI Spending for 2025," *Financial Times*, February 7, 2025, sec. Artificial intelligence. (This document, as are many of the other documents cited in the notes, is online and easily found using a standard search engine. For that reason, the URLs in most cases have been omitted.)
5. Storage, compute, and networking are the core pillars of cloud-based IT infrastructure. Storage handles data storage and retrieval, built on physical devices like solid-state drives or hard disks. Compute provides processing power for applications and workloads, relying on underlying chips (CPUs or central processing units, and GPUs, graphics processing units). Networking enables communication between systems, ensuring seamless data

exchange. While each has a hardware foundation, they can be deployed via software in the cloud.

6. Bob Evans, "Why Uber Picked Google Cloud: The Inside Story," *Cloud Wars*, February 17, 2023; Abhi Khune et al., "Modernizing Uber's Batch Data Infrastructure with Google Cloud Platform," *Uber Blog*, May 30, 2024; Daniel Newman, "Uber Goes Big with Google and Oracle as Cloud Architecture Debate Continues," *Forbes*, February 21, 2023.
7. Ashley Stewart, "Airbnb's IPO Filing Suggests It Could Spend at Least $1.2 Billion with Amazon Web Services by 2027," *Business Insider*, November 16, 2020.
8. On "choke points," see esp. Henry Farrell and Abraham L. Newman, "Weaponized Interdependence: How Global Economic Networks Shape State Coercion," *International Security* 44, no. 1 (2019): 42–79.
9. See Annie Palmer, "Dead Roombas, Stranded Packages and Delayed Exams: How the AWS Outage Wreaked Havoc Across the U.S.," CNBC, December 9, 2021.
10. Joel Patterson, "A Partial List of Sites and Apps Affected by Outages," *New York Times*, October 20, 2025.
11. Both DeepSeek and Meta have published papers on arXiv detailing key architectural innovations in their respective models (DeepSeek-R1 and LLaMA). Each organization has also released open-weight versions of these models, allowing others to deploy and fine-tune them at no cost.
12. Training a single AI model can cost hundreds of millions in computational capacity (OpenAI's GPT4 supposedly took $100 million to train). DeepSeek trained its language model at far lower cost, but running the model at scale still requires capital-intensive cloud infrastructure.
13. For more detail see, e.g., K. Sabeel Rahman and Kathleen Thelen, "The Rise of the Platform Business Model and the Transformation of Twenty-First-Century Capitalism," *Politics & Society* 47, no. 2 (2019): 177–204.
14. Neil Fligstein and Taekjin Shin, "Shareholder Value and the Transformation of the U.S. Economy, 1984–2001," *Sociological Forum* 22, no. 4 (2007): 399–424.
15. Schwartz, "From Fordism to Franchise." The term "fissurization" is from David Weil, *The Fissured Workplace: Why Work Became So Bad for So Many and What Can Be Done to Improve It* (Harvard University Press, 2014).
16. Rahman and Thelen, "Rise of the Platform Business Model."
17. See, e.g., Min Kyung Lee et al., "Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (CHI '15: CHI Conference on Human Factors in Computing Systems, Seoul Republic of Korea: ACM, 2015), 1603–12, https://doi.org/10.1145/2702123.2702548; Brishen Rogers, *Data and Democracy at Work: Advanced Information Technologies, Labor Law, and the New Working Class* (MIT Press, 2023); Alex Rosenblat, *Uberland: How Algorithms Are Rewriting the Rules of Work*, A Naomi Schneider Book (University of California Press, 2018).
18. Pepper D. Culpepper and Kathleen Thelen, "Are We All Amazon Primed? Consumers and the Politics of Platform Power," *Comparative Political Studies* 53, no. 2 (2020): 288–318.

19. Shoshana Zuboff, *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power* (PublicAffairs, 2019).
20. See, e.g., Gerald Berk and Anna Lee Saxenian, "Rethinking Antitrust for the Cloud Era," *Politics & Society* 51, no. 3 (2023): 409–35; Vili Lehdonvirta et al., "Weaponized Interdependence in a Bipolar World: How Economic Forces and Security Interests Shape the Global Reach of U.S. and Chinese Cloud Data Centres," working paper, January 24, 2025, http://dx.doi.org/10.2139/ssrn.4670764.
21. It is often claimed that the idea for AWS came from renting out excess IT infrastructure during Amazon's retail off-season—but this is more myth than fact. For more details on the founding of AWS, see the Acquired Podcast episode "Amazon Web Services: The Complete History and Strategy of Amazon Web Services," https://www.acquired.fm/episodes/amazon-web-services.
22. For a more in-depth (also slightly more technical) account of the evolution of cloud computing, see JS Tan, "The Evolution of the Cloud Computing Business Model," *Value Added*, June 5, 2025.
23. See David Byrne et al., "The Rise of Cloud Computing: Minding Your P's, Q's and K's," Working Paper 25188 (National Bureau of Economic Research, October 2018), https://www.nber.org/papers/w25188.
24. See Gladys Rama, "Guthrie: Azure and AWS Competing on Features, Not Price," *Redmond Channel Partner*, September 14, 2016.
25. See "Scandinavian Airlines Reduces Loyalty Program Fraud with Microsoft Azure Machine Learning," Microsoft.com, May 13, 2020.
26. See https://www.youtube.com/watch?v=bi6Li7WaupQ (11:53).
27. Bryan Wassel, "IKEA Partners with Google Cloud to Enhance Omnichannel and Customer Service Capabilities," *Retail Touchpoints*, October 27, 2020.
28. Karen Hao traces the concept of AGI in her book, *Empire of AI*, showing how the idea slowly took Silicon Valley by storm over the 2010s, and how OpenAI set off the arms race for AGI with the launch of ChatGPT 3.5. Karen Hao, *Empire of AI* (Penguin Press, 2025).
29. The industry is pushing ahead with the assumption that scale is key to model performance, but this hypothesis has not gone unchallenged, most of all by DeepSeek, which we return to in the conclusion.
30. The landmark 2017 paper referenced here was authored by Google scientists and called "Attention Is All You Need."
31. Karen Weise and Cade Metz, "At Amazon's Biggest Data Center, Everything Is Supersized for A.I.," *New York Times*, June 24, 2025.
32. See Columbia River Enterprise Zone board meeting agenda at https://tinyurl.com/5n7f54k9.
33. See Alex Koller, "Nvidia Shares Close Up After Company Unveils Latest AI Chips," CNBC, March 19, 2024.
34. "Inside the Relentless Race for AI Capacity," *Financial Times*, July 31, 2025.
35. The profitability of Microsoft's cloud business is not clear based on official financial statements because the company bundles other products (e.g., Office365, Windows Server, and SQL Server) with their cloud platform in their reporting.

36. Mark Haranas, "Google Cloud Begins Profitability Era: 5 Huge Q2 Earnings Takeaways," *CRN*, July 26, 2023.
37. See Laila Kearney et al., "US Electric Utilities Brace for Surge in Power Demand from Data Centers," Reuters, April 10, 2024, sec. Energy.
38. See Heather Clancy, "Amazon, Google and Microsoft Signal Growing Interest in Nuclear, Geothermal Power," *Trellis*, March 25, 2024.
39. See C. Mandler, "Three Mile Island Nuclear Plant Will Reopen to Power Microsoft Data Centers," NPR, September 20, 2024.
40. See "7 Ways Amazon Is Thinking Big About Nuclear Energy," Amazon.com, https://tinyurl.com/546tbz9w.
41. Microsoft, for instance, has its own custom chips and server rack designs. See Sebastian Moss, "Microsoft Announces In-House Arm CPU and AI Accelerator Chips, Custom Racks," *Data Centre Dynamics*, November 15, 2023.
42. Jake Siegel, "With a Systems Approach to Chips, Microsoft Aims to Tailor Everything 'from Silicon to Service' to Meet AI Demand," *Microsoft Source*, November 15, 2023.
43. Schwartz, "From Fordism to Franchise." See also William Milberg and Deborah Winkler, *Outsourcing Economics: Global Value Chains in Capitalist Development* (Cambridge University Press, 2013).
44. For details on how Uber exercises control over its drivers, see Alex Rosenblat and Luke Stark, "Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers," *International Journal of Communication*, no. 10 (2016), https://ijoc.org/index.php/ijoc/article/view/4892.
45. MarcCharest, "Introducing the New Azure AI Infrastructure VM Series ND MI300X v5," *Azure High Performance Computing (HPC) Blog*, May 21, 2024, https://tinyurl.com/mtzh62bh.
46. Jason Zander, "Cray Supercomputers Are Coming to Azure," *Microsoft Azure Blog*, October 23, 2017.
47. "NVIDIA Teams with Microsoft to Build Massive Cloud AI Computer," *HPC Wire*, November 16, 2022.
48. Matthew Fox, "A Single Customer Made Up 19% of Nvidia's Revenue Last Year: UBS Thinks It's Microsoft," *Business Insider*, May 31, 2024.
49. Dennis Tom and Krysta Svore, "How Microsoft and Quantinuum Achieved Reliable Quantum Computing," *Microsoft Azure Quantum Blog*, April 3, 2024.
50. This example also shows that these ISVs are themselves high-margin firms that can influence, and in some cases even dictate, the terms of their relationship.
51. Microsoft Corporate, "Microsoft and OpenAI Extend Partnership," *Official Microsoft Blog*, January 23, 2023.
52. Microsoft Corporate, "Microsoft and OpenAI Extend Partnership."
53. See Cecilia Rikap, "Intellectual Monopolies as a New Pattern of Innovation and Technological Regime," *Industrial and Corporate Change* 33, no. 5 (2024): 1037–62.
54. See Anissa Gardizy et al., "OpenAI Leaders Say Microsoft Isn't Moving Fast Enough to Supply Servers," *The Information*, October 8, 2024.
55. See Laurence Goasduff, "Why Organizations Choose a Multicloud Strategy," *Gartner*, May 7, 2019.

56. Statista estimates that cloud spending exceeded $660 billion in 2023, while Gartner places total IT spending for the year at $4.7 trillion. See "Public Cloud—Worldwide," accessed March 2, 2025, https://tinyurl.com/ydfe9frs; "Gartner Forecasts Worldwide IT Spending to Grow 6.8% in 2024," *Gartner*, January 17, 2024.
57. Doug Eadline, "Everyone Except Nvidia Forms Ultra Accelerator Link (UALink) Consortium," *HPC Wire*, May 30, 2024.
58. Wesley M. Cohen and Daniel A. Levinthal, "Absorptive Capacity: A New Perspective on Learning and Innovation," *Administrative Science Quarterly* 35, no. 1 (1990): 128–52.
59. See Anissa Gardizy, "Amazon Prepares to Shake Up AWS Sales Group," *The Information*, December 20, 2023.
60. See "Accenture and Schaeffler Pave the Way for Industrial Humanoid Robots with NVIDIA and Microsoft Technologies," *Accenture Newsroom*, April 1, 2025.
61. See Christian Harper and Mylissa Tsai, "Accenture Cloud First Launches with $3 Billion Investment to Accelerate Clients' Move to Cloud and Digital Transformation," *Accenture Newsroom*, September 17, 2020.
62. See "New Accenture Microsoft Business Group Will Empower Enterprises to Thrive in the Era of Digital Disruption," Microsoft.com, February 4, 2019.
63. See https://x.com/tsarnick/status/1842401670225125539; Sam Altman, "The Intelligence Age," September 23, 2024, https://ia.samaltman.com.
64. Cade Metz and Tripp Mickle, "Behind OpenAI's Audacious Plan to Make A.I. Flow Like Electricity," *New York Times*, September 25, 2024.
65. https://x.com/satyanadella/status/1883753899255046301.
66. Yann LeCun, one of the pioneers of modern AI and Meta's chief AI scientist, has argued that we won't reach AGI under the current paradigm—that such a breakthrough requires an entirely different approach. Fei-Fei Li, widely considered the godmother of AI, has also cast skepticism on the whole notion of AGI. See also Helen Toner, "Unresolved Debates About the Future of AI," *Rising Tide*, June 30, 2025, https://helentoner.substack.com/p/unresolved-debates-about-the-future; Henry Farrell et al., "Large AI Models Are Cultural and Social Technologies," *Science* 387, no. 6739 (2025): 1153–56.
67. See, e.g., Theodore Schleifer and David Yaffe-Bellany, "In Display of Fealty, Tech Industry Curries Favor with Trump," *New York Times*, December 14, 2024, sec. Technology; Nicole Narea, "Why Tech Titans Are Turning Toward Trump," *Vox*, July 17, 2024; Kevin Rector and Laura J. Nelson, "Smart Business? Currying Favor? Why Big Tech Leaders Are Friending and Funding Trump," *Los Angeles Times*, January 11, 2025.
68. Henry Farrell, "Why Did Silicon Valley Turn Right?," https://www.programmablemutter.com/p/why-did-silicon-valley-turn-right.
69. Sheera Frenkel and Aaron Krolik, "Trump Taps Palantir to Compile Data on Americans," *New York Times*, May 30, 2025.
70. *National Security Commission on Artificial Intelligence Final Report*, https://reports.nscai.gov/final-report/preface.
71. Sam Altman, "Time to Take a Stand," January 28, 2017, https://blog.samaltman.com/time-to-take-a-stand.

72. See Wency Chen, "Microsoft Abruptly Cuts Services to Chinese University, Genomics Firm," *South China Morning Post*, April 10, 2025; Wency Chen, "Microsoft Shutters AI Lab in Shanghai, Signalling a Broader Pullback from China," *South China Morning Post*, March 31, 2025. Amazon has also pulled its AI operations in China; see Ryan McMorrow and Zijing Wu, "Amazon Shuts Down Shanghai AI Research Lab," *Financial Times*, July 22, 2025.
73. See "Commission Welcomes Member States' Declaration on EU Cloud Federation," Shaping Europe's Digital Future, European Commission (blog), October 15, 2020, https://tinyurl.com/yu2cxtme.
74. Dan Wang, "2020 Letter," January 1, 2021, https://danwang.co/2020-letter/.
75. For more details on the vertical integration of the AI supply chain, see Leonardo Gambacorta and Vatsala Shreeti, "The AI Supply Chain," BIS Papers 154 (March 2025), https://www.bis.org/publ/bppdf/bispap154.pdf.
76. Mark Gurman, "Apple Weighs Using Anthropic or OpenAI to Power Siri in Major Reversal," *Bloomberg*, June 30, 2025.
77. Lauren Forristal, "Netflix Debuts Its Generative AI-Powered Search Tool," *TechCrunch*, May 7, 2025.

Author Biographies

JS Tan (js_tan@mit.edu) is an MIT PhD candidate in the International Development Group in the Department of Urban Studies and Planning. His work addresses the political economy of innovation in the United States and China, with a focus on cloud and AI.

Kathleen Thelen (kthelen@mit.edu) is Ford Professor of Political Science at MIT. Her work focuses on the political economy of the rich democracies, with a current emphasis on the study of American capitalism in comparative perspective. Her latest book is *Attention, Shoppers! American Retail Capitalism and the Origins of the Amazon Economy* (Princeton University Press, 2025).