Rhodium Group

China

Silent Saboteurs: Loaded Assumptions in US AI Policy

February 26, 2025
Reva Goujon (rgoujon@rhg.com), Ben Reynolds (breynolds@rhg.com), and John Larsen (jwlarsen@rhg.com)

In the span of a single week—from the January 13 publication of the Commerce Department's AI Diffusion Framework, to the January 20 release of Chinese AI developer DeepSeek's R1 model—US AI policy has swung from extreme confidence to creeping self-doubt over the country's ability to retain global AI primacy. The irony was undeniable: Hard on the heels of an emboldened US attempt to control the global diffusion of AI came an engineering feat in China that dramatically lowered the barrier to AI diffusion globally.

This manic moment should come as no surprise. Ambitious attempts to regulate a rapidly scaling technology were bound to hit turbulence, and the fact that a Chinese firm is the face of today's technological disruption only amplifies the geopolitical stakes.

To understand what lies ahead, we break down the potential fault lines within the **five core assumptions implicit in emerging US AI policy**:

1. US chip controls will throttle China's indigenous chip production, widening the US lead in foundational AI hardware.
2. US controls can contain Chinese AI competition to its home market.
3. The US will be able to leverage dominance across the AI stack to lock in dependencies in rest of world and condition global access to compute.
4. The US will be able to protect multi-billion-dollar upfront investments in frontier model development and prevent geopolitical challengers from free-riding off US innovation.
5. Energy demand will grow exponentially in line with AI demand, requiring a massive surge in power supply to fuel rapid AI infrastructure buildouts.

In examining both the drivers and potential arrestors behind each of these assumptions, we can stress test emerging US AI policy, anticipate the areas that are most likely to undergo heavy edits under the current administration, and contemplate alternative futures in the next wave of AI competition.

Going for it

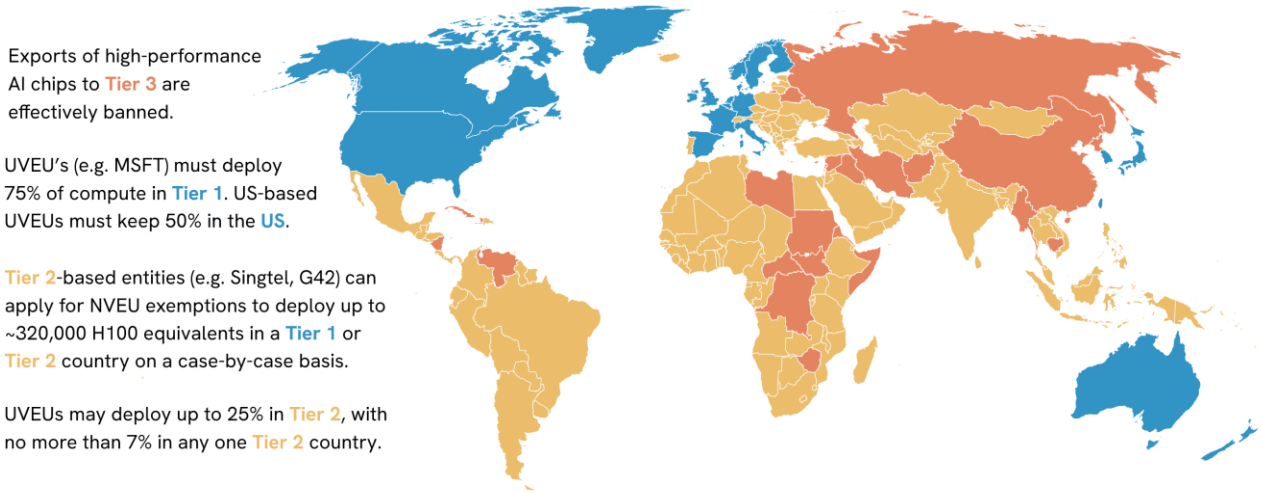
Two recent developments have massive potential consequences for the geopolitics of AI competition: the US AI Diffusion Framework and back-to-back releases of competitive, low-cost open-source AI models developed by Chinese startup DeepSeek.

A US rule to control global access to AI compute

In the final days of the Biden administration, the US released an ambitious AI Diffusion Framework as part of its burgeoning export control regime. The rule represents the most comprehensive attempt to date by the US to leverage its dominance across the AI stack—from chips to cloud services—to regulate worldwide access to computational power ("compute") for AI model development. It is set to take effect May 15 unless the Trump administration decides to overhaul it.

The US Commerce Department Bureau of Industry and Security (BIS) does so by imposing *worldwide* export restrictions on high-performance AI data center chips and closed frontier model weights.[1] BIS then introduces a three-tiered licensing framework that enables varying levels of access on a country-by-country basis (Figure 1). Eighteen Tier 1 countries qualify for license exemptions to maintain unrestricted access to controlled technologies, while Tier 3 countries, including China and other US arms embargoed countries, face an effective ban as a continuation of current policy. All other countries fall into Tier 2 and now face a series of complex rules enabling conditional access to leading-edge chips, subject to various caps and adherence to strict due diligence requirements.

FIGURE 1
US AI Diffusion Framework



[1] High-performance semiconductor restriction covers all items previously restricted for export to China and other US arms-embargoed countries under the October 17, 2023 chip controls (ECCN 3A090.a, 4A090.a, and related .z items.)

We can infer that the rule's conceptual aim is threefold:

- **Concentrate compute deployment for large frontier model training in the US and "trustworthy" partners.** By capping compute exports to Tier 2 countries and obligating US hyperscalers to maintain at least 50% of their installed compute base within US borders, the Framework aims to ensure the US preserves its dominant position as the nexus for frontier model training.
- **Lock in global dependencies on US AI infrastructure and adherence to US AI norms and standards before China can build enough chip capacity to offer an alternative.** The Framework uses new Verified End User (VEU) authorizations to compel emerging AI hubs in Tier 2 to align with US standards to pursue their AI ambitions. Cloud providers must submit a plan for limiting Chinese equipment from their data centers and supply chains to BIS to qualify for a VEU license.
- **Wholly restrict China's access to advanced compute and closed frontier model weights** as the US tries to preserve its AI lead over its chief geopolitical challenger. By imposing restrictions worldwide initially and then rolling them back with specific license exemptions, BIS asserts significant leverage to force data center providers to comply with strict requirements to track and regularly report on their global deployment of advanced chips and closed model weights in a bid to close off potential pathways for diversion of controlled technologies to China.

Verified End User licenses and country caps

The most impactful elements of the AI Diffusion Framework are the new conditions it places on exports to countries in Tier 2. Every Tier 2 country, from Switzerland to Saudi Arabia, now faces a cap on the amount of *advanced chips* it can import. Less advanced chips that fall below the BIS chip control performance threshold, including NVIDIA's H20, do not count against the cap. The cap is defined according to BIS's total processing performance (TPP) metric. Each Tier 2 country can import advanced chips with a total cumulative TPP of 790,000,000—equivalent to roughly 50,000 NVIDIA H100 GPUs—between 2025 and 2027. While this cap can be raised by as much as 100% if the country aligns with US chip controls, the rule does not establish a pathway for Tier 2 countries to graduate to Tier 1.

Tier 2 countries can circumvent the compute cap by partnering with US government-vetted cloud infrastructure providers. US hyperscalers and other Tier 1-based entities qualify for a **Universal Verified End User (UVEU)** authorization, which allows them to deploy advanced AI data center chips in any Tier 2 country independent of the country cap. Critically, UVEUs still face an alternative restriction on the amount of compute they can deploy in Tier 2. At any given time, each UVEU must maintain at least 75% of its global installed base of advanced compute within Tier 1, leaving no more than 25% to be deployed in Tier 2. Moreover, a UVEU may not deploy more than 7% of its global compute base in any single Tier 2 country. UVEUs based in the US, which account for the lion's share of current AI compute installations, face the additional requirement of keeping at least 50% of their deployed compute within US borders.

Tier 2-based providers, such as the UAE's G42 or Singapore's Singtel, qualify for a separate **National Verified End User (NVEU)** exemption. These exemptions will be granted on a case-by-case basis, meaning Tier 2 providers must apply for separate exemptions for each country where they plan to build significant AI infrastructure. The Framework creates yet another distinct quota for NVEU compute deployments. NVEUs face a gradually increasing cap on the cumulative amount of compute they can deploy in any single country (Tier 1 or Tier 2) between 2025 and 2027. This cap begins at roughly 40,000 H100 equivalents in Q1 2025 and rises every quarter before settling at a hard cap of 320,000 H100s in 2027.

The goals of the AI Diffusion Framework overlap neatly with the Trump administration's own priorities in advancing an "America First" AI policy. The design of the rule, however, is highly interventionist both in dividing up the map between Tier 1 and Tier 2 countries (fracturing the EU's internal market in the process) and by imposing compute ratios and GPU caps that defy market logic in some cases and risk becoming obsolete as AI systems continue to evolve at a rapid clip. We discuss the implications and likely revisions to the rule further below.

DeepSeek, deep breaths

Within ten days of the AI Diffusion Framework announcement, DeepSeek upended markets and AI policy debates with the open-source release of its R1 reasoning model. But DeepSeek's breakthrough did not occur in a vacuum. Just four months prior to DeepSeek R1's release, OpenAI had unveiled a new paradigm in AI model development with the launch of its pathbreaking o1 model. OpenAI used reinforcement learning—a set of machine learning techniques that use Pavlovian reward systems to train desired behaviors into a model—to distill complex reasoning capabilities into o1. The core insight behind o1 was that training a model to autonomously think through its reasoning step-by-step and recursively scrutinize its chain of thought could enable substantial performance improvements. Critically, OpenAI's engineers showed that o1's performance improved when it was given more time, and thus compute power, to think during the inference stage.

OpenAI's breakthrough created an immediate incentive for competitors to embrace the new reasoning paradigm. Competitor models were bound to emerge quickly, but the fact that a little-known Chinese startup was the first company to do so was shocking nonetheless. DeepSeek's success in producing a comparable model to o1 at a fraction of the compute cost animated those arguing that the rapid pace of innovation in AI model efficiency invalidates a core assumption behind US chip controls: that massive deployments of cutting-edge hardware are a prerequisite to frontier AI competitiveness.

This argument centers on DeepSeek's apparent success in innovating around US export controls by focusing on building efficient models that maximize the productivity of its limited compute resources. DeepSeek-V3, a large foundation model that was released in late December 2024 and serves as the base model for R1, introduced a handful of novel algorithmic optimizations that significantly reduce the cost of both training and deploying DeepSeek's models. DeepSeek is currently offering its self-hosted application programming interface (API) for roughly 4% the cost of OpenAI's o1 API (Table 1).

TABLE 1
API pricing comparison: OpenAI o1 vs DeepSeek R1 (self-hosted and 3rd-party hosted)
USD per one million tokens

Model/Host	OpenAI o1 (self-hosted API)	DeepSeek R1 (self-hosted API)	DeepSeek R1 (Nebius)
Input	$15.00	$0.55	$0.80
Output	$60.00	$2.19	$2.40

Source: Company websites. Note: OpenAI generally charges between a 50%-75% gross margin premium for API access to their models, while DeepSeek is likely offering its models at razor-thin margins, if not at a loss. But the fact that third-party AI cloud platforms outside China like Nebius are offering R1 at only a slight premium over DeepSeek's self-hosted API indicates that R1's low inference cost is largely a function of its efficient architecture rather than low-margin cost subsidization by DeepSeek.

As impressive as DeepSeek's innovations are, they do not invalidate the case for large-scale AI compute. Algorithmic progress has always been a key vector for enhancing model performance and is best viewed as a complement to, not a replacement for, scaling compute. Frontier model developers outside China will embrace these new techniques as they have embraced similar advancements in the past, not by reducing their compute budgets, but by building bigger, more powerful models to push the boundaries of AI-driven experimentation and inference.

From the perspective of US AI policy, R1's most significant impact relates to its role as an accelerant to the open-source development and deployment of efficient reasoning models. With R1, DeepSeek became the first global frontier AI developer to publicly release a model with similar reasoning characteristics and performance to o1 and offered it to consumers and AI developers at a fraction of o1's cost. DeepSeek also released the R1's model weights and detailed information on its training process and underlying architecture free to the public. All at once, DeepSeek de-mystified and democratized the new reasoning paradigm for open-source developers worldwide.

While DeepSeek does not change the paradigm on compute demand, it does break the barrier on open-source AI diffusion, raising questions over how far Chinese AI developers will be able to invigorate the home market and expand globally while the US works to exclude Chinese players from "trusted" AI ecosystems.

Unpacking US assumptions

There are a number of loaded assumptions implicit in the US AI Diffusion Framework and emerging US AI strategy more broadly. Here we unpack the drivers, arrestors, and watch items for five assumptions that we believe will have the most bearing on the future of US-China AI competition.

Assumption 1: US chip controls will throttle Chinese indigenous chip production, widening the US lead in foundational AI hardware.

DRIVERS

The crux of the US's AI strategy lies in its ability to maintain an absolute monopoly over the mass production of advanced AI chips. This is what has informed a series of chip control measures since October 2022, which collectively aim to deny China access to the high-end chips powering today's AI revolution and cripple its ability to produce alternative chips domestically. So long as the US maintains a monopoly in high-performance chips, it theoretically has the foundational prowess to widen its technological lead with China and the leverage to globally allocate advanced compute to the rest of the world as it sees fit.

ARRESTORS

The jury is still out on whether US chip controls will bite to the point that China falls significantly behind the US in AI development. Part of this has to do with timing: The US has spent more than two years building and patching up a stack of chip controls to cover loopholes and emerging chokepoints. These controls have effectively cut China off from advanced node chip production using extreme ultraviolet (EUV) lithography—now the mainstay for the manufacture of high-performance AI data center chips at process nodes below 7nm.

But Chinese chip designers and foundries are still able to leverage older deep ultraviolet (DUV) lithography, multi-patterning techniques, and advanced packaging processes to manufacture advanced chips on a 7nm-class process node. These are proven engineering techniques—TSMC and Samsung both used multi-patterning to produce 7nm chips at scale for a brief period before migrating to EUV-based manufacturing. With EUV, chipmakers can reduce the number of steps in the manufacturing process, thereby improving their defect rate and overall yield. While Chinese chipmakers using DUV manufacturing and multi-patterning techniques will inevitably have lower yields compared to their foreign counterparts, they also have ample motivation to optimize these techniques and state-backed funding to compensate for their higher burn rates. As long as the equipment that SMIC is using to manufacture Huawei chips remains operable, China theoretically has the capability to manufacture chips on a 7nm process node in growing volumes and with declining defect rates over time. Recent reporting from the FT by China- and Hong Kong-based reporters, citing anonymous sources, suggests that SMIC has already managed to make significant process improvements in the past year, raising the yield of its Ascend AI chip production line from 20% to 40%.

This is precisely why the Biden administration painstakingly attempted to get alignment from the Dutch and Japanese governments to cover *servicing* of semiconductor manufacturing equipment (SME) in their respective export control regimes. The US intent since the launch of the October 7, 2022 controls was to strip Huawei and SMIC engineers of foreign support and throw sand in the gears of Chinese chip production. BIS invoked a new US persons rule that forced US toolmakers like Lam, KLA, and Applied Materials to immediately remove their servicing engineers from China's advanced chip lines. Meanwhile, their non-US competitors only had to remove engineers with US passports from those fabs: Non-US employees could remain at their posts without restriction.

While Japan and the Netherlands have updated their respective controls to cover spare parts and upgrades, they have not been willing so far to rewrite their export rules to also cover servicing for equipment already installed in Chinese fabs. Their firms argue that forcing them to exit the market will simply leave a gap that can be quickly filled by Chinese engineers who have been trained on the equipment and are already in high demand from Chinese SME firms. In their view, export controls that self-restrict servicing will only enable China to more quickly reduce its dependency on foreign toolmakers, while depriving non-Chinese SME firms of valuable visibility into how their equipment is being used within Chinese fabs and how China's semiconductor production capabilities are progressing more broadly.

For now, Tokyo and the Hague are holding the line on servicing. So long as this critical gap in alignment between the US and its partners persists, the US strategy to severely impair Chinese chip production and widen the US lead in AI development is flawed. If the aim of the US is to buy time by slowing China down, then that time could be evaporating with every month that SMIC is able to hone its manufacturing process with ongoing support from foreign toolmakers.

WHAT TO WATCH AHEAD

Extraterritorial enforcement of chip controls: The Biden administration invested considerable diplomatic effort in convincing partners to harmonize their own export control policies with the US. However, it also built a formidable set of new extraterritorial tools that could be used to compel compliance. While the Biden administration continued to court partner alignment to the bitter end, we suspect the Trump administration will be far less accommodating. Demands on export control alignment and enforcement may be coupled with steep tariff threats and an expansion of long-arm controls. BIS already laid the groundwork for extraterritorial enforcement in the December 2, 2024 chip controls, which included a "single chip" de minimis provision designed to assert US writ over tools made in any factory anywhere in the world that contains a single US chip (see December 9, "Slaying Self-Reliance: US Chip Controls in Biden's Final Stretch"). It also created license exemptions for "Supplement 4" partner countries, including Germany, and imposed US restrictions on countries like South Korea and Singapore unless they align with US export controls.[2] This now becomes a question of enforcement and prioritization for the Trump administration, which has already shown a penchant for openly threatening long-standing US allies with tariffs and withdrawal of security cooperation when pushing its demands.

The high-bandwidth memory (HBM) chokepoint: On December 2, 2024, BIS imposed broad restrictions on the export to China of all generations of HBM currently in production. In theory, these restrictions should pose a severe challenge to China's ability to continue producing homegrown AI chips, as Huawei's Ascend AI processors are wholly dependent on HBM imports from Korea. In reality, the Biden administration's delay in implementing these restrictions has already undermined their potential impact. Western media reported on the potential controls six months before they were implemented, giving Huawei ample time to build up a stockpile before the controls came down. Those stocks may provide Huawei with a short-term buffer to meet its Ascend production targets in 2025, but they are not a long-term solution. China will have to produce a viable domestic HBM supply chain to achieve its advanced AI chip ambitions. The US has also taken

[2] Supplement No. 4 to part 742 country list includes EU members (minus Cyprus and Malta) plus UK, Norway, Switzerland, Iceland, Canada. Australia, New Zealand, and Japan (Singapore, South Korea excluded.)

significant steps to close off this path with recent entity listings of Huawei-connected DRAM suppliers and updated technical thresholds to further restrict SME sales to advanced DRAM production facilities in China.[3] While the Biden administration made a controversial decision to spare China's HBM champion, CXMT, from the entity list in deference to requests from partners, the Department of Defense added CXMT to its Section 1260H list on January 7 and further restrictions are likely coming.

Expanding scope of chip restrictions on China: DeepSeek admits that constrained access to GPUs due to US export controls is a significant impediment to its progress but evidently cobbled together a large enough compute cluster to develop its V3 and R1 models. A wide range of estimates have been circulating over how many NVIDIA GPUs, including H100s (banned in the first round of chip controls in October 2022) and H800s (banned in the second round in October 2023), the company was able to acquire as the US chip noose was tightening over the past couple years. The question ahead is what DeepSeek and others will be able to optimize without access to banned chips, forcing them to rely on lower compute chips like the NVIDIA H20s that were designed to be compliant with US export controls and, increasingly, on homegrown Huawei Ascend chips. It appears likely at this point that the US chip ban will expand to cover below-threshold chips as the US tries to strip China of access to foreign technology for AI development.

Beefing up compute governance: Beyond restrictions on the actual GPUs, however, we expect to see a revival of proposals over compute governance that would attempt to restrict Chinese developers from leveraging US technology to build leading-edge AI models. As the AI diffusion rule postulates (albeit without concrete solutions), US regulations are moving toward requiring chipmakers to granularly track exports of their products on a chip-by-chip basis to prevent diversion. We would also expect to see a more targeted approach in which chipmakers and cloud service providers develop ways to monitor the networking capabilities of high-performance chips to prevent them from linking together to form large, powerful clusters without authorization.

Widening the gap with next-gen chip hardware: China's AI champions may have managed to keep within striking distance of their US rivals to date, but BIS officials believe their competitiveness will inevitably erode as advancements in cutting-edge AI hardware expand the compute gap between China and the US-led AI ecosystem. The first major test of this theory is now underway with the introduction of NVIDIA's next-generation Blackwell GPU platform, which introduces substantial improvements in training and inference performance and energy efficiency over its predecessor, Hopper (of the aforementioned H100 chip). Blackwell servers started to make their way into US hyperscale data centers in late 2024 and will become the dominant platform powering AI development and cloud-based deployment outside China by 2026. BIS anticipates that the impact of its export control strategy will become more apparent as deployments of these, and other, advanced chips move forward, while tightening restrictions on Chinese access to foreign chips, SME, and AI cloud services relegate China's AI developers to increasingly outdated compute infrastructure.

BIS is also betting that US-aligned chip manufacturers will extend their process lead over China's emerging domestic champions over the next two years, as SME advancements enable a shift to new architectural paradigms. In leading-edge logic, the shift to gate-all-

[3] HBM is produced by stacking and connecting multiple DRAM silicon dies vertically to achieve high memory density and accelerated data transfer rates.

around transistors and new backside power delivery network architectures will enable efficient scaling beyond 3nm. Memory chipmakers like South Korea's SK Hynix are also integrating next generation packaging techniques like hybrid bonding to increase the number of DRAM layers they can stack up within a single HBM module. All of these advances will be facilitated by innovations at the bleeding edge of various segments of the SME supply chain, from atomic layer deposition and high aspect-ratio etch to metrology and inspection. The gradual integration of ASML's next generation high-NA EUV lithography machines into advanced logic and memory processes also presents an opportunity for the US and partners to extend their lead.

Assumption 2: Chinese AI competition can largely be contained to its home market.

DRIVERS

US policy appears to be moving past the argument that tech controls will accelerate Chinese tech self-reliance. That is now a given, and an implicitly accepted cost. The focus has shifted instead to trying to ensure that China is left to its own devices in indigenizing technology supply chains, and that its tech champions can be largely contained to China's domestic market amid an already slowing economy. The first part of that argument is premised on the idea that China will be stripped of foreign support to advance its chipmaking abilities. But as discussed above, there are still critical gaps in that strategy. A new US Foundry Due Diligence rule is designed in a similar vein: Chinese fabless chip designers can no longer rely on a foreign foundry like TSMC to manufacture advanced chips unless a series of stringent conditions are met to verify that the chip does not exceed BIS's high-performance computing threshold.[4] These new restrictions have had an immediate chilling effect on foreign foundries' willingness to contract with Chinese chip designers for advanced node production and may presage a total divorce in the near future. In other words, even as Chinese firms are making notable advancements in chip design, they will need to increasingly rely on Chinese foundries like SMIC to manufacture their product.

The assumption, then, is that SMIC will be so overwhelmed with demand from Chinese chipmakers that it will inevitably run into challenges in allocating capacity, maintaining sustainable yields, and controlling its burn rate. As a result, the theory goes, Huawei or any other Chinese firm capable of designing chips for AI platforms will hit supply constraints as SMIC struggles to keep up with demand in the home market. SMIC's capacity constraints would in theory prevent Chinese competitors from rivaling AI chipmakers like NVIDIA and hyperscalers like Amazon Web Services, Microsoft Azure, and Google Cloud in building and operating data centers in third markets. The US would then have free rein to impose conditions on Tier 1 and Tier 2 countries to contract with US firms

[4] The Foundry Due Dilligence rule requires "front-end fabricators" like TSMC to presume any advanced logic chip meets the performance thresholds stipulated in ECCN 3A090.a and are thus subject to US high-performance chip controls unless one of three conditions is met: 1) the chip designer is on a list of "approved" IC designers included in the appendix to the rule, and that designer attests that the chip falls below the ECCN 3A090 threshold; 2) The chip is packaged by a front-end fabricator located outside Country Group D:5 (China and other US arms embargoed countries) or Macau, which verifies that the chip does not contain HBM and has fewer than 30 billion transistors (gradually raised to 40 billion by 2029); or 3) the chip is packaged by an "approved" outsourced semiconductor assembly and test provider (OSAT), which verifies that the chip does not contain HBM or 30 billion+ transistors.

to build out data centers powered by US-made AI chips without having to worry about Chinese tech rivals offering competing offers with fewer conditions.

ARRESTORS

If gaps in US-partner alignment persist over the servicing of China's installed base, then Huawei and SMIC theoretically still have the means to manufacture advanced node chips in growing volumes and improve on yields over time. The question then is whether SMIC will run into hard constraints allocating capacity to the production of Huawei Ascend 900-series processors for AI applications versus smartphone processors, especially as AI competition intensifies and the state may be compelled to steer resources toward industrial AI development instead of consumer devices. Meanwhile, the rapid adoption of DeepSeek's open-source model already compromises the assumption that US models will be the default platform for AI development globally.

Meanwhile, the rapid adoption of DeepSeek's open source model already compromises the assumption that US models will be the default platform for AI development globally.

WHAT TO WATCH AHEAD

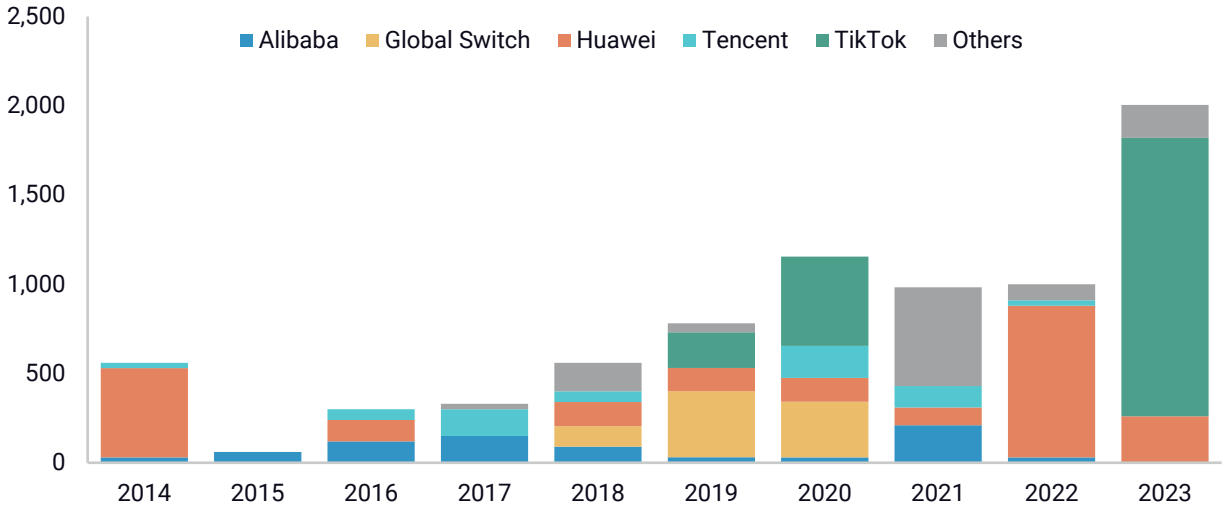
SMIC's technology constraints: While it is exceedingly difficult to get an accurate picture of SMIC's production challenges, signs of capacity constraints should become more visible with time. For example, we could see persistent delays to new product launches by Chinese device makers and data center buildouts by Chinese cloud providers attributed to production challenges and chip shortfalls. Mobile device teardowns can also provide clues on how much progress SMIC is making in refining and upgrading its advanced node processes. In September 2023, semiconductor industry research firm TechInsights published a teardown analysis of the Huawei Mate 60 that highlighted SMIC's progress in refining its 7nm multi-patterning process to enable mass production of smartphone chips. The event triggered a panic in DC that raised the Biden administration's urgency to tighten export controls. A more recent TechInsights teardown of Huawei's new Mate 70 pro revealed only marginal design upgrades to the phones core processor, upsetting rumors that Huawei's next-generation premium smartphones would be powered by SMIC's developing 5nm multi-patterning process.

We also need to watch the impact of the US Foundry Due Diligence rule in deterring foundries like TSMC from contracting with Chinese chipmakers. There are already early indications that TSMC has been forced to terminate contracts with Chinese chip design firms due to its inability to meet BIS compliance standards. If companies are unable to pass the due diligence requirements of the rule, then more orders will presumably be diverted to SMIC to fulfill. The question then becomes whether SMIC will be able to expand production capacity fast enough to meet growing demand for Chinese customers losing access to foreign foundries.

Chinese overseas investments: Chinese outbound FDI in data centers will be another leading indicator of whether Chinese hyperscalers (Alibaba, Tencent, Huawei, Baidu) are able to compete with US cloud service providers overseas. If US export controls are designed to deny Chinese firms access to foreign-made high performance chips specially designed for use in data centers, then it will fall to Huawei and SMIC to supply chips for the home market and expansion abroad. Preliminary Rhodium Group data on Chinese outbound FDI shows Huawei has been active in overseas data center buildouts (Figure 2),

particularly in the Middle East and North Africa. But none of those data centers are equipped with the hardware required to power AI applications, at least not yet. If China manages to develop enough homegrown AI chip capacity to enter the global AI data center market, we would expect to see a sharp uptick in China's data center OFDI, especially in Tier 2 markets.

FIGURE 2
Chinese outbound greenfield FDI in data centers
USD million



Source: Rhodium Group China Cross-Border Monitor. Note: Data is preliminary and subject to revision.

Blocking sanctions ahead? There is strong potential for Huawei and affiliates to come under more severe sanctions. Huawei is already under the highest degree of US export controls (foreign direct product rule restrictions), which is what also drove the tenacious firm's efforts to rapidly indigenize its supply chain. But if the US were to impose full blocking sanctions on Huawei, it would force companies anywhere in the world to choose between continued business with Huawei or access to the US financial system. As a result, Huawei's ambitions to build AI infrastructure in third markets and expand global adoption of Huawei devices and its HarmonyOS software platform for IoT applications would effectively be short-circuited. License exemptions could be granted to avoid immediate disruptions to Huawei's extensive telecom contracts all over the world, but those too would likely come under stress if the US were to impose blocking financial sanctions on the firm. Given that Huawei was already a chief target of Trump 1.0 and is now the face of China's self-reliance campaign, we suspect news of Huawei breakthroughs will be a trigger for more aggressive measures by the Trump administration, especially if the president views Huawei as unfinished business from his first term.

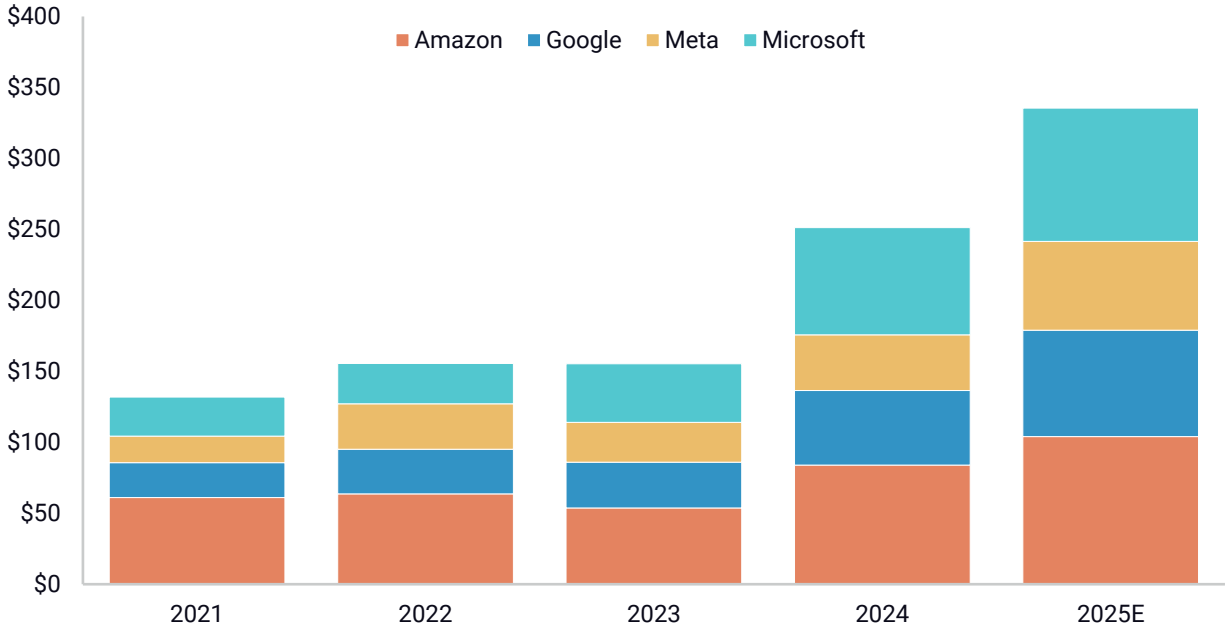
Assumption 3: The US will be able to protect multi-billion-dollar upfront investments in frontier model development and prevent geopolitical challengers from free-riding off US innovation

DRIVERS

US companies have invested enormous sums to build out the foundational infrastructure for large-scale AI training and deployment over the past few years. While initial market reactions to the DeepSeek breakout reflected concern that the return on these investments may be undermined by the proliferation of more compute-efficient models, US frontier model developers have downplayed this threat. At this early stage in AI development, their goal is to build increasingly capable models. They view architectural innovations, therefore, not as a means to cut costs on compute, but as an opportunity to build more powerful models while continuing to scale up compute.

US chipmakers and AI data center providers share the view that DeepSeek's actual impact is more likely to support *growing* compute demand in the long run. As a result, they continue to plan massive capital expenditures to build AI infrastructure in 2025 (Figure 3). While the majority of their data center investments in recent years have focused on scaling compute for AI model training, their expectation is that relative compute demand will gradually shift toward deploying these models for real-world AI inference applications over time.

FIGURE 3
US big four data center hyperscaler capex (historical and 2025 projected)
USD billion



Source: Company financial reports, earnings transcripts. Note: Microsoft capex estimate covers FY 2025 (July 2024-June 2025). Estimate for Microsoft's calendar year 2025 derived by assuming flat growth from H1 2025 to H2 2025. While capex figures include non-data center oriented spend, majority of capex for each company has been devoted to AI cloud infrastructure buildout during this time frame. Within this group, Amazon spends the largest share of capex on non-data center items to support physical infrastructure for its e-commerce business.

US cloud service providers (CSPs) view DeepSeek as a catalyst for increasing demand for AI inference deployment for two primary reasons. First, as the cost of training and deploying models with frontier capabilities falls, they become more accessible to the public. Third party developers will exploit access to more efficient, high-performing models to build new AI-powered applications that will attract more demand for cloud-hosted inference compute. This dynamic is reflected in an obscure economic concept that Silicon Valley's AI boosters have latched onto in the wake of R1's release, Jevons Paradox, which holds that increasing the efficiency of resource use often leads to higher overall consumption of that resource.

DeepSeek's commoditization of reasoning-based inference models is also likely to drive growing demand for compute. As explained above, the performance of reasoning models scales up as more compute is applied during the inference phase. While these models are particularly well-equipped to handle math and coding applications currently, they are expected to drive more general progress across a wide range of domains moving forward. By lowering the barrier to widespread adoption of reasoning models, DeepSeek R1 contributes to the acceleration of this transition to a more compute-heavy inference paradigm. By this logic, US policies that aim to concentrate compute in US hands while tightening restrictions on Chinese access to compute are likely to endure as the US tries to appreciably widen its technological lead over China.

ARRESTORS

While DeepSeek's emergence does not undermine the technological logic for large-scale investments in compute infrastructure, it does raise legitimate questions about the return on investment for massive closed frontier model training runs. OpenAI achieved a major leap forward in reasoning and problem-solving capabilities with o1, but the competitive advantage it gained from this breakthrough was short-lived. DeepSeek's introduction of a comparably performant model with significantly lower inference costs already threatens to erode OpenAI's pricing power. While this is great news for consumers and AI developers, it poses a serious challenge to OpenAI's business model. OpenAI is estimated to earn a profit margin of between 50% and 75% on its API offerings but still reported a $5 billion loss on $3.7 billion in total revenue in 2024 due to the massive scale of investments the company is devoting to model development.

DeepSeek's emergence and the advancement of open-source frontier models more generally have heightened existing doubts surrounding the ability of closed-source frontier model developers like OpenAI and Anthropic to preserve a competitive moat that can justify their massive upfront investments in model training. If open-source developers in China, or elsewhere, continue to keep pace, the case for pouring massive investments into closed-source model development could be compromised.

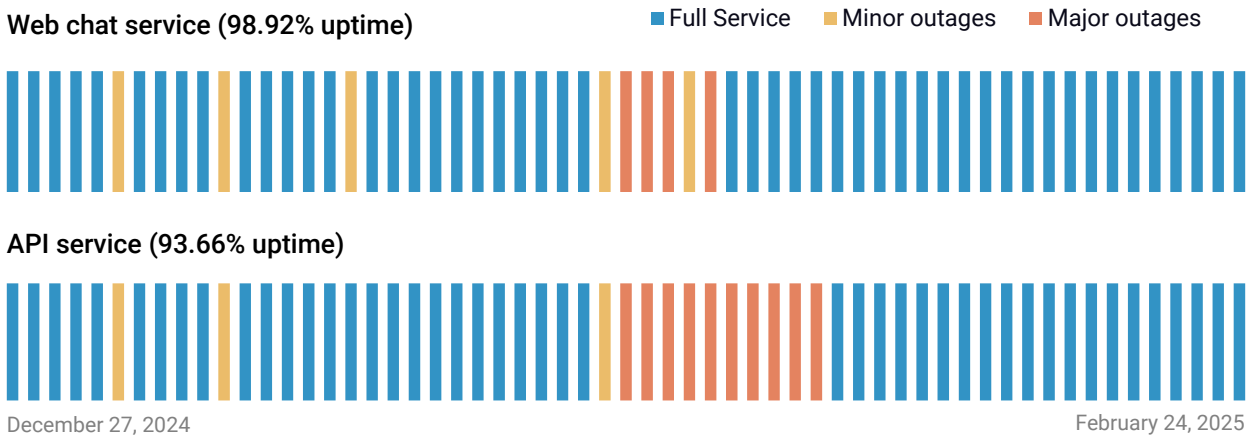
WHAT TO WATCH AHEAD

Rapid technological diffusion is inherent to AI development. There is only so much that can be controlled, especially in the open-source domain. Those benefiting from accelerated AI diffusion will argue for a "do no harm" approach to technology restrictions to enable rapid innovation. Nonetheless, US policy anxiety will inevitably grow over the prospect of competitive open-weight models developed by Chinese firms becoming default platforms for AI development. As a result, **US tech controls will naturally gravitate toward the access points for compute: end user controls for cloud service providers and economic**

security or "trustworthiness" standards designed to prevent integration of Chinese models into critical infrastructure and industry.

Scope of data security and cybersecurity restrictions. As DeepSeek popularity soared across the globe, several governments (including US federal and state governments, Italy, South Korea, Australia, Taiwan, and India) imposed varying restrictions on usage of the DeepSeek app for government agencies. This dynamic has parallels to the TikTok debate in the US, with arguments being made that the DeepSeek app can be a vector for malware, is subject to Chinese censorship, collects sensitive data (for example, user and proprietary data inputs through usage of the app) and that data can theoretically be transmitted to Chinese state-owned entities. While the DeepSeek app shot to the top of app store charts outside China in late January, its momentum quickly stalled as the firm's limited compute capacity proved ill-equipped to serve a mass influx of new users (Figure 4). This issue has dissipated in recent weeks as a combination of foreign government bans and, more importantly, the integration of DeepSeek's open-source models by US cloud providers and app developers like Perplexity has reduced foreign direct demand for DeepSeek's internally hosted app and APIs. Nevertheless, foreign government responses to the potential data security concerns raised by the DeepSeek app suggest that AI apps hosted by Chinese companies may face broader restrictions on national security grounds, multiplying the TikTok effect.

FIGURE 4
DeepSeek API and web chat service status (December 27, 2024 - February 24, 2025)
Service uptime (% of time service is operating at full performance)



Source: DeepSeek

The data and cyber security arguments surrounding the DeepSeek app are distinct from the use case of companies adopting DeepSeek's open-source model weights for fine tuning internal models. In the latter scenario, a US cloud provider hosting DeepSeek on its platform becomes the main conduit for data flows with end users, neutralizing the risk of a China-based entity accessing sensitive data from the end user unless the cloud provider itself suffers a major cyber breach. This does not preclude the potential for further restrictions, however. Research by Anthropic on the potential for embedding so-called "sleeper agents" in source code is drawing attention to the risk of more subtle backdoors that can be difficult to detect in standard safety evaluations. In essence, a large model injected with exploitable code can use chain-of-thought reasoning to deceive the safety

training process itself, thereby masking the risks. This security argument could be used as a foundation for policymakers and tech influencers arguing for broader restrictions to prevent US cloud providers from hosting LLMs developed by countries of concern like China. A narrower application would limit Chinese LLM integration into critical US infrastructure. The US may also consider imposing long-arm measures to further restrict the ability to host Chinese models on AI clouds in other global markets as a follow on to the AI Diffusion Framework.

The (still open) open source debate: Some policymakers may argue that the only way to protect US investment in frontier models is to double down on the OpenAI closed model paradigm and restrict open-source model developers like Llama from releasing their most sensitive IP (model weights, training data, source code, model architecture, critical research on the model training.) This would be a big move that would violate the core principles of the open-source philosophy: that releasing IP to the world is the most rapid and secure pathway to innovation as both technological capabilities and safety oversight get diffused to a large number of players instead of concentrated among a small cadre of well-capitalized Big Tech firms.

To date, the US has so far refrained from trying to regulate open source, going only so far as to restrict the export of *closed* model weights in the AI Diffusion Framework. A July 2024 National Telecommunications and Information Administration (NTIA) study on the pros and cons of regulating open-weight models informing the AI Diffusion rule concluded that it was still too early to warrant restrictions on open-source models.

Moreover, DeepSeek has already broken the barrier on commoditizing AI model development and, along with other major players like Alibaba, has ambitions to become the dominant open-source platform globally. The US will be reticent to undermine its own leading open-source AI model champion, Meta, when China is already rapidly gaining ground. Nonetheless, we need to watch regulatory debates around licensing restrictions for open-source models. For example, there is still a spectrum of open-source licensing, with Meta requiring licenses for organizations with more than 700 million active users while DeepSeek's simpler MIT license is highly permissive. Some policymakers may argue for country-based licensing restrictions to prevent Chinese companies from integrating US-built models, but that may have a limited effect if Chinese LLMs like DeepSeek prove to be a competitive substitute to US models. Regulations focused on tightening licensing restrictions for open-source models would also be inherently difficult to enforce.

Evolution of copyright restrictions: LLM developers are already in uncharted legal territory given that the models they develop are trained on vast amounts of copyrighted material. This has spurred an active debate over IP protections for both model inputs and output. OpenAI and Google have attempted to mitigate these concerns by signing licensing agreements with major media outlets to secure access to their content. This dynamic can give model developers access to more exclusive content but is also bound to substantially raise the costs of development. Cross-jurisdictional issues will be paramount in this trend. It can take years to negotiate IP protections in a multilateral framework, and the current geopolitical climate is not conducive to such coordination. For example, if AI distillation—a training technique which uses output from a larger "teacher" model to distill knowledge into a smaller "student" model—enables a Chinese model developer to train off a US model that is paying licensing fees for content, it could lead regulators to impose country-based restrictions for API model access. If Chinese developers continue to double down

on open-source releases in trying to become a default global AI standard, however, restrictions on US model developers could also undermine their own competitive edge.

China's policy response. China's domestic AI industry has benefitted immensely from the open-source paradigm. Open access to research and model weights from leading foreign developers like Meta and Mistral has been a key enabler of the rapid progress of DeepSeek, Alibaba, and other emerging AI leaders in China. However, that dynamic is becoming increasingly bidirectional as China emerges as a leading innovator at the frontier of open-source AI development. DeepSeek is now doubling down on its open-source ethos, releasing code on a daily basis as part of its "#OpenSourceWeek" in the spirit of "full transparency" and "pure garage energy and community-driven innovation."

While Beijing may chafe at the idea that Western firms can feed off China's open-source innovations even as US restrictions take aim at China's AI developers, it is unlikely to act to constrain open-source practices anytime soon. The Chinese government recognizes that open source offers China's AI community a valuable lifeline in the context of tightening US chip controls. US AI developers can leverage their access to large-scale compute to run more experiments and test more complex, compute-intensive architectures to improve their models and discover new paradigms. In a purely closed-source environment, this dynamic would place compute-constrained Chinese developers at a massive disadvantage. Open source mitigates that disadvantage to an extent by enabling Chinese developers to benefit from knowledge transfers across a broad global community.

Beijing's policy response is more likely to focus on restricting domestic market access for foreign-produced models, while promoting use of indigenous LLMs over US-based models. We will be watching to what extent Beijing is able to tame its statist instincts in promoting rapid AI development at home without stifling private sector innovation. Xi Jinping chaired a rare symposium on February 17 with top tech leaders, including DeepSeek founder Liang Wenfeng. The event indicated that China's leadership is exercising restraint on private sector regulation for now, choosing instead to endorse open-source development and encourage widespread adoption of the DeepSeek model, along with a Huawei-led effort to build out AI infrastructure. We still need to watch targets for state-backed funding for AI development and efforts to centralize compute resources, as such moves will be watched closely by US policymakers for sanctions targets. Proximity to the Huawei-SMIC nexus runs a real risk of US restrictions ensnaring Chinese AI developers.

Divergent incentives drive conflicting stances on US controls among US tech champions. DeepSeek's breakout revealed growing fault lines emerging within Silicon Valley regarding the wisdom and possible future trajectories of the US government's strategy to outcompete China in AI. While all the key players are aligned on the need to support a massive expansion in US data center capacity, they are increasingly in conflict on the issue of whether, and how, to disrupt China's AI development via tightening technology controls. This dynamic will be critical to understanding which tech influencers will ultimately have Trump's ear in shaping AI policy.

The clearest line of demarcation exists between companies that benefit from the diffusion of low-cost open-source models and those that view this trend as a direct threat to their business models. Closed frontier model developers like Open AI and Anthropic have taken on billions of dollars in losses to invest in frontier model R&D but are vulnerable to the impact of price erosion by fast-following open-source competitors. These companies have

expressed optimism that their access to large-scale compute will allow them to widen the gap with smaller competitors as they continue to push the frontier of the new inference scaling paradigm. However, in recent months, they have also leaned into lobbying efforts to convince the US government to expand its controls on China and the global diffusion of AI. While they present these stances as reflective of US national security interests, these companies have an apparent business interest in constraining opportunities for challengers to emerge outside the US, particularly in China, which is both highly competitive in AI and committed to forwarding the open-source paradigm.

China's increasing competitiveness in open source raises a complex set of threats and opportunities for Meta. On the one hand, the rise of open-source competitors like DeepSeek and Alibaba challenges Meta's strategy to entrench its Llama family of models as the foundational platform for global open-source development, potentially undermining Meta's ability to extract business license fees from large-scale Llama deployments. However, open-source innovation also supports Meta's more pressing goal of commoditizing frontier AI to undercut its closed model competitors and lower the cost of deploying inference. Meta has generally avoided taking a stance on US tech control policy toward China specifically, but has lobbied aggressively against potential US restrictions on open-source model weight sharing, pointing to the risk of ceding the market entirely to China.

TABLE 2
Overview of US AI companies' positions on US tech controls and DeepSeek's impact

Company	Position
Closed model champions	▪ **Advocating for tighter US tech controls** to restrict China's AI development, while positioning themselves as standard-bearers for US-led democratic AI. ▪ **Anthropic CEO Dario Amodei:** DeepSeek's releases "make export control policies even more existentially important…[to keep] democratic nations at the forefront of AI development." ▪ **OpenAI CEO Sam Altman:** US should "set out rules of the road for what sorts of chips, AI training data, and other code—some of which is so sensitive that it may need to remain in the US—can be housed in the data centers that countries around the world are racing to build to localize AI information."
Meta	▪ **Opposition to US open-source restrictions** framed as necessary for preserving US influence, preventing Chinese developers from establishing dominance over global open-source standard-setting. ▪ **CEO Mark Zuckerberg:** DeepSeek reaffirms that "there's going to be an open-source standard globally. And I think for our kind of national advantage, it's important that it's an American standard."
Hyperscale CSPs	▪ **Embracing DeepSeek's low-cost models** to accelerate inference deployment while undermining CSP's strategic dependence on expensive closed model partnerships. ▪ **Broad opposition to AI Diffusion Framework:** The rule imposes heavy compliance burdens on CSPs and potentially caps ability to expand in attractive Tier 2 AI data center markets.

	▪ **Oracle Executive VP Ken Glueck:** "The government proponents of this rule claim that they are protecting US hyperscalers from global competition. Respectfully…we don't need a ride, we need government to get out of the way."
NVIDIA	▪ **Strong promoter of AI sovereignty, with an eye toward expanding global market for NVIDIA products and services; Opposed to caps imposed by AI Diffusion Rule,** which would restrict market access in Tier 2 and risk creating an opening for Huawei competition ▪ **Ned Finkle, VP of Government Affairs:** AI Diffusion rule "would only weaken America's global competitiveness, undermining the innovation that has kept the US ahead."

Assumption 4: The US will be able to leverage dominance across the AI stack to lock in dependencies in the rest of the world.

DRIVERS

This assumption is a central tenet of the AI diffusion rule. The US has extraordinary leverage across the AI stack—in chips, software, and cloud services—and is readily exercising that leverage to condition global access to AI compute before Chinese competitors pose a credible threat in third markets. The rapid pace of technological development, the promise of AI-driven productivity breakthroughs, and the multi-dimensional national security challenges posed by AI are together injecting a sense of urgency among governments to invest in AI infrastructure buildouts. This means that most countries, while aspiring toward a "sovereign AI" future free of dependencies on either the US or China also cannot afford to wait to build up an indigenous tech ecosystem. AI chipmakers like NVIDIA and US hyperscalers will still pervade even the boldest of sovereign AI strategies, including French President Emmanuel Macron's recent announcement of €109 billion ($112.6 billion) in private AI investment in France.

This puts the US in pole position for now to define AI security standards through a geopolitical lens. The Trump administration may be boisterously pulling away from the Biden administration's emphasis on AI safety but it is still likely to double down on AI security and trustworthiness standards. This is the key to ensuring AI buildouts don't commit the mistake of 5G buildouts when Chinese telecom firms were already sweeping global markets. This time, the US has the lead in AI buildouts, and the US theoretically has the arguments and tools to demand that countries divorce themselves from Chinese tech partnerships on cybersecurity, data security, and overall national security grounds.

ARRESTORS

The most obvious constraint to this strategy emerges from the design of the AI Diffusion Framework itself, which limits the freedom of US AI cloud providers to expand in foreign markets by requiring them to maintain at least half of their deployed compute base in the US and prohibiting them from building more than 7% in a single Tier 2 country or 25% in Tier 2 as a whole. Proponents of the rule assert that these ratios will have little, if any, immediate impact, since they simply reflect the state of global AI deployment as it is today—well over 50% of the global installed base of AI compute currently resides in the

US, and while a handful of Tier 2 countries have formulated ambitious AI plans, they are still in the early stages of their AI infrastructure buildouts. The intent, at least for the Biden administration, which wrote the rule, is not to reshape the landscape of global AI compute as it exists today, but rather to steer its future development in a direction that comports with a US vision of a secure, pro-democracy AI world order.

While most countries fall into the Tier 2 bloc, these conditions are apparently targeted at constraining the growing AI ambitions of two countries in particular, the UAE and Saudi Arabia. Within the past year, both countries have announced ambitious plans to invest in domestic infrastructure for frontier model training and deployment. The Biden administration also curiously divided the EU single market in its tier classifications, relegating countries like Poland, Austria, and the Baltic and Balkan states to Tier 2.

Meanwhile, DeepSeek has captured policy and engineering minds alike on how to enable AI model development more broadly and in line with a particular country's economic strengths, language, culture, and values. Companies like Meta want to be the global standard and platform for such development, but open-source models like DeepSeek are gaining traction fast in third markets. The imposition of trustworthiness standards could be applied to limit usage and integration of Chinese LLMs in the US and partner markets: This would preserve an arena for competition among "trusted" developers, but would also require convergence around national security arguments. At present, there is a lack of goodwill among the US and partners to advance such an agenda. If the right to compute is viewed increasingly by sovereigns as a universal right, then US attempts to ration compute is bound to hit diplomatic turbulence.

WHAT TO WATCH AHEAD

Streamlining the AI Diffusion Framework. The Trump administration may simplify the top-down tiers and ratios designed into the AI Diffusion Framework, leaving ample opportunity for Trump-style transactionalism with countries to access compute depending on whether they can reach an economic and security understanding with the US. This bodes well for Tier 2 countries like the UAE that are investing in the US's $500 billion Stargate AI infrastructure initiative. However, it poses challenges for EU countries already divided between Tier 1 and Tier 2 status in the current rule and facing a litany of trade and security frictions with the Trump administration. This includes the thorny issue of digital services taxes that disproportionately target US tech firms, which the Trump administration wants to neutralize as part of his pressure campaign on Europe.

Economic security standards: The evolution of economic security standards across the US and G7 countries may be one of the most important variables defining the next four years. The US is already bluntly defining such standards in asserting that Chinese providers of information and communications technologies—from chips and routers to software—are not trustworthy to justify new restrictions on Chinese firms' access to the US market. The AI Diffusion Framework follows a similar logic in forcing VEU applicants to establish a "supply chain risk management plan to limit PRC-origin equipment from entering their data centers. These could become de-facto standards for US and partner countries that will endure well beyond the fractious years of the Trump administration.

The Draghi effect: High friction with the Trump administration will drive more ambitious strategies for building out sovereign AI capabilities, but those strategies will require both significant funding and a conducive regulatory climate for innovation. This is a major litmus

test for Europe, already prodded by the recommendations of former European Central Bank President Mario Draghi's to *de-regulate* if Europe wants to be competitive in an AI economy. It will be especially important to watch to what degree emboldened member states like France internalize the Draghi effect and whether that in turn invigorates a bigger shift within the EU bureaucracy to fight an impulse to regulate a burgeoning AI economy. If Trump demands on scaling back digital services taxes yield a mini trade deal with the EU that includes digital provisions for cross-border trade, this could be another driver for change and innovation.

Anti-competition cases: US qualifications for a "universal verified end user" to access US-made GPUs and deploy compute outside the US could give way to anti-competition cases by other countries if the policy is viewed as disproportionately favoring US incumbents and preventing the emergence of competitive firms as countries try to build out their own AI capabilities.

Assumption 5: Energy demand will grow exponentially in line with AI demand, requiring a massive surge in power supply to fuel rapid AI infrastructure buildouts.

DRIVERS

The relative distribution of AI compute demand between model training and inference deployment will be a critical factor driving power supply buildout decisions moving forward. While Meta and others are developing new techniques to enable large models to be trained across geographically distributed networks of data centers, training frontier models currently requires extremely low latency. This requires large amounts of *geographically concentrated* compute. As a result, large, localized power supplies are needed to develop and train ever more capable AI. This paradigm has dictated US policy and data center buildouts to date—go big and then go bigger. As long as large, localized frontier model training remains a critical enabler of AI model development, countries that can build large installations of excess generating capacity fast will be best positioned. That is what is driving hyperscalers to scour the US and the world for opportunities to build multiple gigawatts (GWs) of electric power capacity in one place at one time.

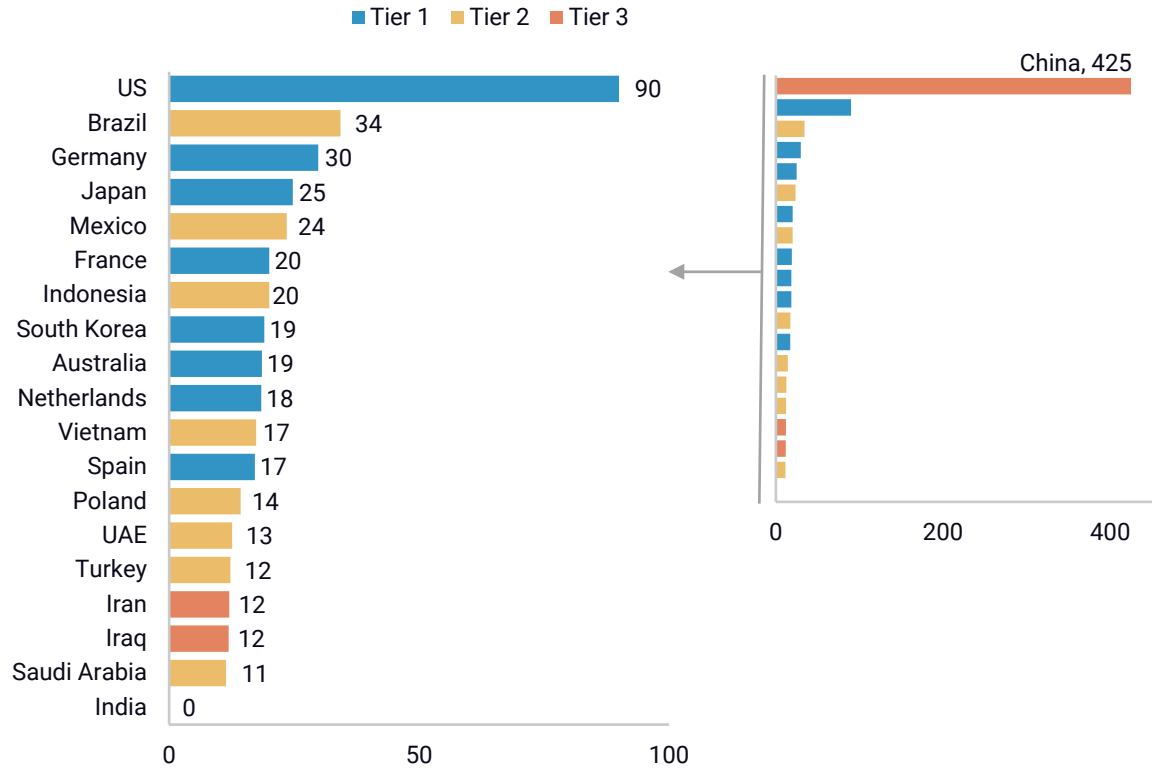
We find[5] that over the past five years, the US on average has added enough net capacity to provide 90 TWh per year of generation above and beyond what was needed to meet load growth (Figure 5). That is the most of any Tier 1 country and three times more than Germany, the next Tier 1 country. China, with its industrial overcapacity and buildout of all types of generation, added over 400 TWh per year: Access to electricity is not a significant constraint on China's ability to sale AI. Among the ten countries that have built the most excess energy capacity over the past five years, six are located in the Tier 1 country bloc (with China, Brazil, Mexico, and Indonesia serving as the exceptions). These countries have either built large amounts of capacity fast or maintained a decent build

[5] To evaluate which countries are most capable of providing a large amount of available energy capacity to scale AI compute quickly, we first focus on the countries with the fastest recent capacity build rates around the world. We adjust recent capacity additions by country based on average capacity factors and estimate how much new generation a country has added over the past five years that is surplus to what is required to meet demand growth over the same period. This allows for a consistent cross-country assessment of each country's potential to meet the expected near-term surge in AI demand.

rate and experienced slow or declining electric demand across their economies, leaving slack capacity that could be used to supply AI. It is worth noting that we are looking at electricity capacity writ large in this analysis, including coal and other fossil capacity. AI buildouts in each of these countries would have very different conventional and carbon pollution implications based on energy composition.

Assuming the training paradigm holds, Tier 1 countries in Europe, Asia, and Australia appear to be in the best position to meet AI demand in the near-term. However, past performance is not a guarantee of future results. Tier 2 countries outside the top ten in this analysis, such as the UAE, have large ambitions to expand their data center and associated energy capacity and probably have the means and capabilities to do so. As long as large, centralized electricity generation is considered essential to developing frontier models, countries with resources and attractive investment and regulatory environments could expand the map for AI development under the right circumstances.

FIGURE 5
Recent five-year average surplus net generation
TWH per year



Source: Rhodium Group analysis of Ember data. Notes: Generation values are estimated from capacity data normalized across generation types based on average capacity factors then subtracting average electric demand over the same period to calculate surplus generation additions in each jurisdiction. India and Indonesia did not add more generation that is required to keep pace with demand growth over the period hence values of zero.

ARRESTORS

As compute demand for inference becomes more dominant, scale and centralization of energy buildouts will matter less. It is less clear how the pace of electricity buildout will be impacted. Energy demand will still grow with AI diffusion more broadly as efficiency gains drive growth in demand, but the timeline will be dictated by the uptake of AI services. Recent analyses have shown that training is currently by far the biggest source of AI-related electricity demand. However, as the availability of low-cost, high-performing models enables the development of more advanced AI applications, it may not take long for inference driven demand to dwarf training.

Inference-driven load growth will look fundamentally different from training-driven growth primarily by where it is located, rather than its speed and scale. Inference typically does not require large, concentrated compute. It can be distributed across multiple data centers all over the US or the globe. This reduces the need for building giant installations of concentrated electric power capacity and instead leaves a lot of flexibility in how much, where, and what kind of generation is needed. Indeed, if data center load can be managed in a flexible way, up to 100 GWs of new load could be integrated on the grid with minimal impact on reliability and little new additional capacity. With decentralized, flexible AI demand driven by inference, the opportunities for countries to scale up AI expand far beyond the countries that can make large amounts of power available quickly, raising questions about how the diffusion rule does or doesn't constrain build out.

Regardless of whether inference ends up driving energy demand, if DeepSeek or other model developers continue to act as fast followers to frontier model developers, the return on investment from ever larger data centers and centralized power may not be compelling, leading to a slow down or even a stall along the training paradigm.

WHAT TO WATCH AHEAD

The next two to three years of capital expenditures on large, concentrated data centers and associated electric power supply in the US are largely baked in. We will be watching what people are planning for after that. If the training paradigm continues, speed and scale will be paramount. However, the practical supply chain, siting, and permitting constraints associated with this kind of buildout may prove obstacles despite the Trump administration's efforts to streamline things. The US is currently on track for roughly 2% annual electric demand growth over the next decade (see Rhodium Group's Taking Stock projections)—a pace the grid has not seen since the 1990s. Since World War II, the US has only managed to maintain linear growth, making it hard to envision how it can manage a rapid acceleration to maintain the training paradigm indefinitely. Over the past 25 years, electric demand globally has also been linear. If the inference paradigm takes center stage, we will be watching for more distributed data center and electric power buildout that is more opportunistic, flexible, and grows at the pace of AI adoption.

Time is tyranny

The most potent variable in the current era of AI competition is time. As we have seen from innovations from OpenAI to DeepSeek, three mere months of model development can lead to path-breaking innovation, turning the market on its head. That innovation can yield rapid diffusion, enabling countries to turbocharge long-stagnant economies.

But time is tyranny for the world of regulators. There is a fundamental asymmetry between the pace of innovation and the speed at which regulators can, or even should, react. Months of critical delays to patching and enforcing tech controls threaten to eviscerate a nation's technological lead. That in turn feeds an anxiety that threatens to pollute a climate of AI innovation with ever-blunter controls.

The five loaded assumptions we've unpacked in this note offer a sobering lesson in humility. Big moves with monumental implications are being made in compressed time. Striking the delicate balance between fostering cutting-edge innovation, safeguarding those investments, and containing geopolitical challengers demands unprecedented technical acumen, strategic foresight, and deft diplomacy. In the absence of any one of these critical elements, the consequences for global stability and technological progress could be extreme.

ABOUT RHODIUM GROUP

Rhodium Group is an independent research provider with deep expertise in policy and economic analysis. We help decision-makers in both the public and private sectors navigate global challenges through objective, original, and data-driven research and insights. Our key areas of expertise are China's economy and policy dynamics, and global climate change and energy systems. More information is available at www.rhg.com.

DISCLOSURES

This material was produced by Rhodium Group LLC solely for the recipient. No part of the content may be copied, photocopied or duplicated in any form by any means without the prior written consent of Rhodium Group. Redistribution, forwarding, translation, or republication of this material in any form by you to anyone else is prohibited. Rhodium Group LLC is not an investment advisor. Any information contained herein is not intended to be relied on as investment advice and this information is not purported to be tailored advice to the individual needs, objectives or financial situation of a recipient of this information. This report is intended for informational purposes only and does not constitute a recommendation, or an offer, to buy or sell any securities or related financial instruments. The information contained herein accurately reflects the opinion of Rhodium Group at the time the report was released. The opinions of Rhodium Group are subject to change at any time without notice and without obligation of notification. Rhodium Group does not receive any compensation from companies that may be mentioned in this report. No warranty is made as to the accuracy of the information contained herein.

New York | California | Washington, DC | Paris

Website: **www.rhg.com**

Rh g Rhodium Group