Statement before the
Committee on Science, Space, and Technology of
the U.S. House of Representatives

"DeepSeek: A Deep Dive"

A Testimony by:

Gregory C. Allen
Director, Wadhwani AI Center, CSIS

Tuesday, April 8, 2025
2318 Rayburn House Office Building

WWW.CSIS.ORG
1616 RHODE ISLAND AVENUE NW
WASHINGTON, DC 20036
TEL: (202) 887.0200
FAX: (202) 775.3199

Allen: Written Testimony, House SST

Chair Obernolte, Ranking Member Stevens, and distinguished Members of the Committee, thank you for inviting me to testify today. The Center for Strategic and International Studies (CSIS) does not take policy positions, so the views represented in this testimony are my own and should not be taken as representing those of my current or former employers.

I currently serve as the director of the Wadhwani AI Center at CSIS, where I lead a team conducting policy research at the intersection of technology, economics, and national security. Prior to CSIS, I spent three years working at the U.S. Department of Defense Joint Artificial Intelligence Center, where I most recently served as the director for strategy and policy. My primary professional background is corporate strategy roles in technology-driven industries. On March 7, 2025, I published a report through CSIS titled "DeepSeek, Huawei, Export Controls, and the Future of the U.S.-China AI Race," and many of my remarks today reflect my conclusions from that research effort.

For my testimony today, I hope to offer a useful perspective on the origins of DeepSeek's AI models and what DeepSeek's progress in AI model development, as well as Huawei's progress in AI chip design and manufacturing, mean for U.S. competition with China.

DeepSeek did not come out of nowhere. Its parent company, High-Flyer Capital Management, has roots in AI-enabled high frequency trading that provided a strong technical foundation in terms of both computing infrastructure and workforce talent. Algorithmic trading firms frequently own and operate their own data center infrastructure, which they use to develop proprietary investment strategies powered by machine learning artificial intelligence. These efforts routinely include intense research on developing proprietary and unorthodox data center engineering methods for improved speed and efficiency. DeepSeek's heritage from High-Flyer's machine learning research and intensive computing optimization techniques are evident in some of the strategies that it has pursued to develop its AI models.

Many media and policy commentators in the United States and China have argued that DeepSeek's efficiency gains are evidence that U.S. export controls have failed to restrain China's AI sector and that DeepSeek proves that huge numbers of cutting-edge U.S. AI chips are not required for developing and serving high-performance AI capabilities. There is some merit to aspects of the first claim, but the second is wrong.

To begin, DeepSeek's V3 research paper states that their models were trained on 2,788,000 GPU-hours using Nvidia H800 chips, which, at an estimated cost of 2 dollars per GPU-hour, equates to $5.576 million. This was only the final, successful pretraining run. It did not include the compute cost of the hundreds of experiments that were run beforehand to generate the insights necessary for that final run, nor did it include the compute costs associated with post-training fine-tuning or inference compute workloads. Thus, it is not an apples-to-apples fair comparison to say that DeepSeek's AI model development cost only $5 million while American firms require hundreds of millions of dollars.

The H800 is a degraded version of the H100 chip that Nvidia specifically developed for the Chinese market to comply with the U.S. export controls imposed on October 7, 2022. In particular, those controls restricted the sale of chips that exceeded performance thresholds across two metrics: total processing power and interconnect speed.

Government sources told me that the U.S. government knew that these restrictions would prevent Nvidia from shipping its (at the time) market-leading A100 chip and upcoming H100 chips. They inferred that Chinese customers would be restricted to using the V100, which was first introduced in 2017 and has significantly lower performance than its successors. Additionally, the U.S. government assumed that, if Nvidia were to develop a new chip specifically for the Chinese market that exceeded export control performance thresholds in one metric but not the other, this would require the typical multiyear AI chip development timeline.

However, Nvidia had a mechanism for post-manufacturing modification of its existing chip products. Specifically, each Nvidia chip is designed for redundancy, recoverability, and defect tolerance to minimize the impact of manufacturing defects. Nvidia blew fuses on A100 chips to reduce their interconnect speed (but not their processing power) below the export control performance thresholds, thus creating the A800 product lines that were legal to export to China.

The Biden administration ultimately realized that continued sales of the A800 and H800 chips meant that their policy would not have the intended impact on China's AI ecosystem. Former government officials told me that this was clear internally by December 2022. However, it took the administration 12 months after their initial package of export controls to act. The U.S. government modified the export control performance thresholds in October 2023 to block exports of A800s, H800s, and any non-Nvidia chips with comparable performance to China. During the year-long period in which A800 and H800 chip exports to China were legal, an enormous number of them were sold to China. Precise sales figures are not available, but Nvidia's disclosed sales to China between October 31, 2022, and October 31, 2023, exceeded 9 billion dollars.

If DeepSeek were to admit that it used H100s in its data center to train its models, it would be confessing to activity illegal under U.S. law, since the H100 has never been legally available in China and cannot even legally be purchased by Chinese-firm subsidiaries outside of China. Thus, if DeepSeek did engage in illegal activity, it would have an incentive to conceal this fact. This raises the question of whether DeepSeek lied in its publications when it claimed to have used H800 chips exclusively.

While there is no public evidence that large-scale Nvidia chip smuggling occurred prior to October 2023, industry sources told me that, by early 2024, large-scale H100 chip smuggling operations were underway. In mid-2024, journalists at The Information interviewed participants in eight distinct H100 smuggling networks, each of which provided evidence that they had completed

H100 smuggling transactions worth more than $100 million.[1] These networks continue to be active, with increasingly sophisticated techniques for evading detection. For example, a December 2024 investigation by The Information found that

> When notified of an upcoming inspection, smugglers have duplicated the serial numbers of the servers with Nvidia chips they've purchased from Supermicro [and already smuggled to China] and attached them to other servers they had access to.[2]

More recent reporting by the Wall Street Journal suggests that this smuggling now includes the latest generation of Nvidia Blackwell AI chips.[3] China is betting that its network of smugglers and shell companies can find the leaks in the Commerce Department's Bureau of Industry and Security (BIS) export control enforcement barrier. As long as Congress continues to neglect BIS by providing grossly inadequate resources compared to the size and importance of its mission, China has a reasonable expectation of success.[4] BIS needs not only more money, but also more skilled staff, more enforcement agents, and better enabling technology, especially in data analysis.

Industry analyst Ben Thompson has pointed to strong evidence that DeepSeek did in fact use H800s as it claimed: Many of DeepSeek's algorithmic and architectural improvements are ideal for maximizing the effective use of computing resources under conditions of limited interconnect bandwidth. At a minimum, this strongly suggests that DeepSeek uses many H800s in its computing infrastructure and that a model with the performance of V3 can indeed be trained exclusively on H800s.

This does not prove, however, that DeepSeek exclusively uses H800s in its overall computing infrastructure. Indeed, some reporting in Chinese news media claims that DeepSeek did train its slightly more recent R1 model on Nvidia H100 chips that are banned from China under U.S. export controls. The semiconductor consulting firm SemiAnalysis, citing anonymous industry sources, recently wrote that DeepSeek has a total of 50,000 Hopper generation GPUs, a category that includes H100s, H800s, and H20s. SemiAnalysis specifically claimed that it has evidence that High-Flyer's/DeepSeek's computing infrastructure includes at least 10,000 H100s, 10,000 H800s, 30,000 H20s, and 10,000 A100s as part of its total computing stack. SemiAnalysis further estimated that High-Flyer/DeepSeek spent $1.63 billion in GPU server capital expenditures alone (i.e., excluding other data center construction and operating costs).[5]

In terms of absolute performance DeepSeek's January 2025 AI models are best understood as being comparable to the best American models available in the summer of 2024. DeepSeek did introduce, however, many improvements related to computational efficiency for both AI training

[1] https://www.theinformation.com/articles/nvidia-ai-chip-smuggling-to-china-becomes-an-industry
[2] https://www.theinformation.com/articles/u-s-prompts-nvidia-supermicro-probe-into-how-chips-ended-up-in-china
[3] https://www.wsj.com/tech/china-nvidia-blackwell-chips-ai-531fed0c
[4] https://www.csis.org/analysis/mismatch-strategy-and-budgets-ai-chip-export-controls
[5] https://semianalysis.com/2025/01/31/deepseek-debates/

and inference. Some of these were already well understood by American firms but not disclosed publicly, while others were genuinely new innovations. To the general public and media, both types were perceived as new.

While I do believe that DeepSeek should serve as a wakeup call for America, the extent of the media coverage on DeepSeek was out of proportion to its technical achievements. This reflects in part the fact that users of free AI models had not experienced an exposed chain-of-thought reasoning feature, which customers clearly enjoyed, as well as interest in the fact that the model came from China. This helped drive DeepSeek as a major topic of consumer and media interest, which in turn drove the tech stock price drop.

Beyond chip smuggling, the greatest strategic challenge for the United States is the potential for China to produce AI chips domestically at sufficient quantity and quality to build AI data center infrastructure that is competitive with the United States.

DeepSeek is not in and of itself the most significant threat to U.S. leadership in AI. Instead, the greater challenge arises from the possibility of China having a domestic ecosystem for producing its own AI chips at large scale and integrating them into Chinese data center training, as well as running inference for DeepSeek and other AI models.

As U.S. technology firms are planning hundreds of billions of dollars in AI data center infrastructure investments, it is worth remembering that—for those investments to be possible—companies like TSMC must manufacture enough AI chips to fill those data centers. In the case of companies such as Nvidia, their revenue growth in recent years is less than it would have otherwise been due to shortages of TSMC production capacity.

The Biden administration took many steps designed to definitively cut China's AI chip designers off from TSMC production capacity. Most recently, on January 15, the Commerce Department announced the final tranche of Biden administration export controls, often referred to as "the Foundry Rule." The Foundry Rule moved advanced chip production to a white-list system that will likely make it impossible for Chinese AI firms to access TSMC capacity to produce chips above export control performance thresholds even when operating through complex shell company arrangements. However, TSMC manufactured a strategically significant quantity of chips on behalf of Huawei via shell companies prior to the rule going into effect.

That effectively means that China's long-term future in AI is closely tied to its ability to produce AI chips domestically. The Biden administration sought to hamstring China's domestic production of advanced chips by restricting the sale of advanced semiconductor manufacturing equipment, including from other countries.

China's alliance of Huawei (AI chip designer), SMIC (AI chip manufacturer), and CXMT/XMC (high-bandwidth memory manufacturers) have recently made strategically significant progress in advancing domestic production of AI chips.

Domestically producing large quantities of AI chips will require China to domestically replicate multiple segments of the AI chip value chain. The most important links are AI chip design, advanced node logic chip manufacturing, and advanced node high-bandwidth memory (HBM) manufacturing.

Like the United States, China has many different companies working on AI chip design, including Huawei, Cambricon, Biren, and more. However, Huawei is unambiguously in the strongest position with its Ascend AI chip product line.

The most advanced logic chip manufacturer in China is SMIC. SMIC's SN2 facility in Shanghai is the sole facility in China with an active 7 nm logic chip production line and has been producing 7 nm chips since July 2021, more than a year before the first tranche of the Biden administration's semiconductor equipment export controls went into effect. SMIC and Huawei are now working to bring a 5 nm node into scaled production but must do so without access to Extreme Ultraviolet (EUV) lithography equipment, since China has no local producer of EUV lithography machines and since export controls have prevented such machines from ever being exported to China. Industry sources told me that, in early 2020, ASML was poised to ship EUV tools to China and that SMIC was planning to work with key research labs in Europe, such as the Interuniversity Microelectronics Centre (IMEC), to help develop their EUV-based manufacturing process.

In December 2024, industry sources told me that SMIC currently has enough immersion deep ultraviolet (DUV) lithography equipment supplied by the Dutch company ASML to produce 85,000 FinFET wafers per month (WPM) across both SN1 (which focuses on 14 nm node production) and SN2 (which focuses on 7 nm and 5 nm production). This acquisition of lithography tools reportedly took effect before Dutch DUV lithography export controls went into effect in mid-2023.

However, the bottleneck in expanding 7 nm (which in SMIC's node naming system is called "N+2") production capacity has not been lithography but rather U.S. tools for etching, deposition, inspection, and metrology. Some of this equipment is restricted on a country-wide basis, meaning that it cannot be legally sold anywhere in China. However, other types of this equipment were restricted only on an end-use and end-user basis. This means that the equipment can be sold to some customers in China but not others. In some cases, this even means that it can be sold to some facilities of a particular customer, but not others. In such cases, relocating the equipment from one facility to another would require a new export license in order to be legal. But SMIC's production of 7 nm chips using U.S. equipment is already illegal, and both SMIC and other Chinese firms always have the option to choose illegal activity, particularly since such activity frequently enjoys the active support of the Chinese government.

Industry sources told me that SiEn, Pensun, and Huawei's fab in Dongguan all were able to legally acquire the needed etching, deposition, and inspection/metrology equipment that SMIC needs for two reasons: (1) the equipment was not restricted on a country-wide basis to all of China and (2) the equipment was restricted on an end-use and end-user basis, but SiEn and Pensun told U.S. firms that it would exclusively be used for producing chips less advanced than 14 nm. These firms also denied any affiliation with Huawei. Government officials told me that in such circumstances, the equipment can often be sold under a no-license required status.

According to the sources, SiEn and Pensun, however, did not have sufficient customers providing demand for using all of the equipment they had purchased, and so some of it was never used operationally in their fabs. They purchased the equipment as a stockpiling move in anticipation of future export controls. The source described this as a "buy everything you can, while you can" strategy.

The SMIC SN2 facility needed the equipment. Since SiEn and Pensun were not making economically productive use of the equipment, they were amenable to a sale. This sale was negotiated in Q4 of 2024 and completed in Q1 of 2025. The sources are under the impression that all of the desired equipment is currently either installed at SMIC SN2 or on-site awaiting installation.

As a result of the successful in-country equipment transfer, SMIC expects to achieve 50,000 7 nm wafers per month (WPM) by the end of 2025. If all of this capacity was devoted to manufacturing Ascend AI chips that would imply the production of millions of Ascend 910C chips annually. However, SMIC is unlikely to devote all of its 7 nm capacity to Ascend chips. Huawei needs that 7 nm capacity for its chips for smartphones, laptops, data centers, and telecommunications equipment. Moreover, SMIC has other customers besides Huawei. Still, the point remains that Huawei is likely poised to dramatically expand Ascend production in the near future.

Huawei's Ascend chips continue to face challenges in terms of a lack of compatible AI software that is driving low utilization of purchased chips. However, this could change if DeepSeek's open-source community enthusiasm improves Huawei's CANN software ecosystem competitiveness with Nvidia's Compute Unified Device Architecture (CUDA). DeepSeek may have both the technical knowledge and the open-source community enthusiasm to finally start generating momentum around Huawei's competitor to CUDA, which Huawei refers to as its Compute Architecture for Neural Networks (CANN). For a company like DeepSeek, migrating all AI workloads from CUDA to CANN would likely be a multiyear project. The greater maturity of the CUDA software ecosystem currently makes Nvidia chips more attractive, but this could change over the next few years. If it does, it would have major implications for U.S. AI competitiveness both inside and outside of China.

Thank you for the opportunity to testify today, and I look forward to your questions.

MARCH 2025

DeepSeek, Huawei, Export Controls, and the Future of the U.S.-China AI Race

By Gregory C. Allen

Introduction

Six months ago, few in the West aside from obsessive AI professionals had heard of DeepSeek, a Chinese AI research lab founded barely more than a year and a half ago. Today, DeepSeek is a global sensation attracting the attention of heads of state, global CEOs, top investors, and the general public.

With the **release** of its R1 model on January 20, 2025—the same day as President Trump's second inauguration—DeepSeek has cemented its **reputation** as the top frontier AI research lab in China and caused a reassessment of assumptions about the landscape of global AI competition. By January 27, DeepSeek's iPhone app had overtaken OpenAI's ChatGPT as the **most-downloaded** free app on Apple's U.S. App Store. The stock prices of some U.S. tech companies briefly tumbled, including the AI chip designer Nvidia, which lost more than **$600 billion** off its valuation in a single day. (AI chips are also known as graphics processing units, or GPUs, and the terms are used interchangeably in this report.)

ChatGPT has again overtaken DeepSeek in app store rankings, and Nvidia's stock price has **since mostly recovered**. However, investor interest in Chinese tech companies has **grown significantly** and remains elevated. DeepSeek is now even **reportedly** seeking investment from venture capital firms.

As a sector, AI is prone to overreactions and wild swings in perception. One might think that the story of DeepSeek is just another overblown AI hype cycle. However, the extraordinary attention focused on DeepSeek is justified, even if the conclusions some have drawn from its success are not. It would be a great mistake for U.S. policymakers to ignore DeepSeek or to suggest that its accomplishments are merely a combination of intellectual property theft and misleading Chinese propaganda. Policymakers need to understand that—even while DeepSeek has in some cases simply implemented innovations **already known** to U.S. AI companies—DeepSeek has also demonstrated **genuine technological**

CSIS | CENTER FOR STRATEGIC & INTERNATIONAL STUDIES

breakthroughs of its own. These facts deserve careful consideration as the second Trump administration sets its AI policy agenda.

This paper provides an overview of DeepSeek's origins and achievements, their geopolitical implications, and the key challenges facing U.S. and allied policymaking as a result. The paper pays particular attention to DeepSeek's implications for the future of AI and semiconductor export control policy. It concludes that—while DeepSeek's success does partly reflect failures of earlier implementations of U.S. export controls—these controls can continue to play a critical role in supporting the United States' strategy for winning the AI race against China. Success is not guaranteed and will require two things:

1. Preventing large-scale AI chip smuggling and
2. Preventing the team of Huawei and SMIC from providing a viable Chinese AI chip alternative to the leading incumbent international team of Nvidia and TSMC.

Both will be exceedingly difficult, and in both cases export controls can at best slow and disrupt Chinese efforts, not stop them.

Regarding the second point, Chinese sources and leaders are expressing renewed optimism. At a meeting on February 17 between Chinese Community Party (CCP) Chairman Xi Jinping and Chinese technology executives (including DeepSeek CEO Liang Wenfeng), Huawei founder Ren Zhengfei **told** Xi that his **previous concerns** about the lack of domestic advanced semiconductor production and the damaging impacts of U.S. export controls had eased because of recent breakthroughs by Huawei and its partners. Ren further **said** that he is leading a network of more than 2,000 Chinese companies who are collectively working to ensure that China achieves self-sufficiency of more than 70 percent across the entire semiconductor value chain by 2028. These predictions should be taken seriously. While China's economy faces many challenges, Gerard DiPippo of RAND is correct to **argue**,

> Even if overall economic growth remains comparatively weak and many Chinese firms continue to struggle, China's central and local governments will keep supporting high-tech industries and emerging stars like DeepSeek. . . . Western policymakers shouldn't make the mistake of believing China is down and out, and they shouldn't be surprised when China continues to catch up or make breakthroughs in critical and emerging technologies.

What follows are a series of 21 key judgments regarding DeepSeek, Huawei, export controls, and the future of U.S.-China AI competition.

Readers are advised to pay special attention to the discussion of Huawei and SMIC, as this report contains a great deal of new and previously private information regarding Huawei's progress toward indigenizing China's AI and semiconductor technology stack.

1. **DeepSeek did not come out of nowhere. Its parent company, High-Flyer Capital Management, has roots in AI-enabled high frequency trading that provided a strong technical foundation in terms of both computing infrastructure and workforce talent.**

DeepSeek is a subsidiary of High-Flyer Capital Management, a Chinese quantitative hedge fund focused on algorithmic trading powered by deep learning. Deep learning is the **core technology paradigm** underlying the past two decades of rapid AI advancements. Cofounded in 2015 by Liang Wenfeng, who is also the founder and CEO of DeepSeek, High-Flyer was already **ranked** as the top Chinese hedge fund in 2019.

DeepSeek's heritage via High-Flyer in the algorithmic trading world helps to explain its success in frontier AI research and especially its success with its most recent model releases. Algorithmic trading firms **frequently own and operate** their own data center infrastructure, which they use to develop proprietary investment strategies **powered by machine learning artificial intelligence**.

High-Flyer has **admitted** that its activities include "High-Frequency Trading," though High-Flyer de-emphasizes this work since it is **politically controversial** (and was for a time banned) in China. High-frequency trading firms race against each other to be the fastest asset traders reacting to new market developments and are obsessively focused on optimizing and accelerating their computing infrastructure. These efforts **routinely include** intense research on developing proprietary and unorthodox data center engineering methods for improved speed and efficiency. Firms can justify making these expensive investments in computing and network infrastructure optimization because the success or failure of trades collectively worth billions of dollars depends upon reducing the time between learning of a fact and making a stock trade based on that new fact by mere nanoseconds (billionths of a second).

High-Flyer is among the leaders of such firms in China and has long been among the most focused on both AI algorithms for trading and data center engineering. The *Financial Times* **reported** that, by mid-2022, High-Flyer had already acquired more than 10,000 Nvidia A100 chips (which at the time were the most advanced AI chips in the world) and spent more than 1.2 billion RMB (roughly $180 million at the time) to build two data centers. These infrastructure investments distinguished High-Flyer from most other Chinese companies: Only **four** other companies in China possessed such a large quantity of chips at the time, and all were major tech companies, not finance companies. In a 2023 interview, Liang **said**:

> We always wanted to carry out larger-scale experiments, so we've always aimed to deploy as much computational power as possible. . . . We wanted to find a paradigm that can fully describe the entire financial market.

In another 2023 **interview** with a Chinese media outlet, Liang provided a timeline for High-Flyer's AI chip compute reserves as a gradual "progression from one GPU in the beginning, to 100 GPUs in 2015, 1,000 GPUs in 2019." He stated that, from the day of DeepSeek's 2023 founding onward, High-Flyer had both the computational power and "ample R&D budgets" to support DeepSeek's mission of "research and exploration" in pursuit of artificial general intelligence (AGI).

DeepSeek's heritage from High-Flyer's machine learning research and intensive computing optimization techniques are evident in some of the strategies that it has pursued to develop its AI models. More

broadly, the synergy of technical skills between the world of algorithmic trading and AI research is well known. For example, industry sources at leading U.S. AI labs told CSIS that they have also focused on recruiting talent, especially compute infrastructure engineers, with experience in the high-frequency trading industry. Both sectors **pay** such talented individuals handsomely.

2. **DeepSeek's technological achievements in terms of performance and cost were not surprising. They reflected a continuation of longstanding trends. However, the fact that such achievements came from a Chinese lab was a surprise.**

DeepSeek's recent technological achievements (reflected in the **DeepSeek-V2**, **DeepSeek-V3**, **DeepSeek-R1**, **DeepSeek Math**, and **Native Sparse Attention** technical papers) can most easily be understood through two basic facts:

First, DeepSeek has demonstrated a suite of algorithmic and architectural improvements that significantly reduce the amount of computing power (and therefore financial investment) required for any AI model to reach and operate at a given level of performance. This is true for both the training stage (in which AI models are created) and the inference stage (in which their capabilities are used to serve internal or external customers).

Second, DeepSeek has now developed AI models in China that are broadly comparable to the best U.S.-developed AI models introduced in mid-2024 (e.g., **OpenAI's o1** or **Anthropic's Claude 3.5 Sonnet**).

Of the two demonstrations, the fact that these advances came from China is more surprising than the fact that they happened at all. DeepSeek's improvements in technical performance were consistent with **preexisting trend lines** of AI performance improvements. They represent expected performance breakthroughs, not shocking ones.

To understand why, it is helpful to consider an analogy in the form of Moore's Law. **In economic terms**, Moore's Law essentially means that computer chips improve to be roughly twice as good in performance for the same amount of money roughly every two years. Though the nature of that performance improvement has changed, this trend has been consistently observed for almost **six decades** (with **some caveats**).

As Moore's Law has been repeatedly validated since the 1960s, it has become steadily easier to confidently predict that computing performance per dollar will continue to double every two years. But, as the fortunes of Intel vividly demonstrate, that is not synonymous with being able to state with confidence which specific company or which specific technical innovations will deliver that improved performance. The same is true of AI today: It is easier to forecast the degree of AI performance growth than to forecast who will invent that performance growth or how.

AI has its own versions of Moore's Law that are **driving exponential growth** in both absolute performance and performance available at a given cost. At a high level, these sources of improvement can be grouped into three types:

- **Increasing available computing resources** (either by using **better computer chips**, **more chips**, or both);

- **Improving algorithms and architectures** (for example, to deliver more "intelligence" for a given compute budget); and
- Improving the quality and/or quantity of available **training data**.

Epoch AI, an AI-focused nonprofit research institute, has **estimated** that, while the amount of human-generated training data will likely plateau in 2028, algorithmic and architectural improvements and increased computing resources have been improving at 3x per year and 4.6x per year respectively—for a combined multiplier effect of 13.8x per year. In short, AI performance is increasing at a rate far faster than the familiar (though still extraordinary) pace of exponential progress described by Moore's Law. Even ignoring the possible remaining improvements from harvesting additional training data, a (highly oversimplified) back-of-the-envelope estimate of future performance based on the compute trends and algorithmic architectural trends would suggest that AI performance at the end of 2028 will be more than 36,000 times better than at the beginning of 2025, rather than just the four or eight times better that would be true if improvement were only occurring at the familiar pace of Moore's Law.

For those who only began paying attention to AI with the late 2022 launch of ChatGPT, this pace of progress can be both breathtaking and difficult to believe. However, for those familiar with the AI industry's history over the past two decades, the fact that an AI model that requires a massive data center one year can run on a laptop or a single AI chip a **few years later** is well-known as a **recurring**, though still quite incredible, phenomenon. The dramatic reduction in cost-per-performance DeepSeek demonstrated is not a surprise to Western AI researchers. They have been leaders in driving—and benefitting from—the exact same trend for more than a decade. One might ask: "If that is the case, why are U.S. AI labs still building ever larger and more powerful data centers?" Put simply, such companies are reinvesting all efficiency gains toward improved overall performance. Dario Amodei, CEO of the U.S. AI lab Anthropic, put it **this way**:

> Because the value of having a more intelligent system is so high, this [reduced cost] shifting of the curve typically causes companies to spend *more*, not less, on training models: the gains in cost efficiency end up entirely devoted to training smarter models, limited only by the company's financial resources.

Moreover, AI researchers are clearly nowhere near the absolute ceiling of computational performance and efficiency. The human brain is a real-world existence proof of human-level intelligence that requires **only 20 watts** of electricity, compared to the megawatts or **gigawatts** contemplated for future AI data centers. Despite the many **differences** between human brains and machine intelligence, the existence of the brain hints at future intelligence efficiency possibilities in much the same way that bird wings **inspired** the Wright brothers.

Thus, it is not at all surprising that somewhere on earth, an AI research lab managed to deliver the improved AI performance-per-dollar that DeepSeek's technical papers demonstrate. The only surprise is that the research lab that discovered some of the relevant technical innovations is a Chinese one, and notably one whose **technical team** appears to have been **educated** and trained almost entirely in China, rather than at Western universities or Western companies.

While Chinese researchers have been at or close to **world class** in many **domains of AI research** for many years, DeepSeek represents the **first time** that a Chinese AI lab has demonstrated breakthroughs

at or near the absolute frontier of foundational AI research. Many of the techniques that DeepSeek demonstrated are now the new state-of-the-art. Of note, this is exactly what the Chinese government set as a goal in China's 2017 **official AI strategy**:

> By 2025, China will achieve major breakthroughs in basic theories for AI, such that some technologies and applications achieve a world-leading level and AI becomes the main driving force for China's industrial upgrading and economic transformation.

3. **DeepSeek's technological innovations are real, not propaganda. They have been in all cases proven to work by Western researchers who replicated DeepSeek's approach.**

In the first week after the publication of DeepSeek's R1 paper, **some commentators** argued that DeepSeek's results were fabricated as part of a Chinese propaganda effort or "psyop." This is simply false. Researchers at U.S. universities have **already reproduced** some of the research results described in the DeepSeek V3 and R1 papers, and U.S. AI firms are **already working** on adapting DeepSeek's novel techniques to utilize them in their own AI efforts.

As will be discussed further in this paper, essentially all DeepSeek's technical innovations relate to algorithmic and architectural improvements. DeepSeek did not have superior computing hardware, nor did they have access to uniquely important datasets (excluding synthetic data) that might have provided an advantage over Western researchers. The key technical innovations that DeepSeek demonstrated across the **DeepSeek-V2** (May 2024), **DeepSeek-V3** (December 2024), **DeepSeek-R1** (January 2025), and **other** research papers are significant. Some of these innovations had already been independently discovered by U.S. firms (and not disclosed publicly), but others were genuinely new.

4. **DeepSeek's success in large part reflects the lagging impact of the flawed first package of U.S. AI chip export controls in October 2022. The U.S. government acknowledged and partially remedied these flaws in its October 2023 update.**

Many **media** and **policy** commentators in the United States and China have argued that DeepSeek's efficiency gains are evidence that U.S. export controls have failed to restrain China's AI sector and that DeepSeek proves that huge numbers of cutting-edge U.S. AI chips are not required for developing and serving high-performance AI capabilities. There is some merit to aspects of the first claim, but the second is wrong.

To begin, DeepSeek's **V3 research paper** states that their models were trained on 2,788,000 GPU-hours using Nvidia H800 chips, which, at an estimated cost of 2 dollars per GPU-hour, equates to $5.576 million.

The H800 is a degraded version of the H100 chip that Nvidia **specifically developed** for the Chinese market to comply with the U.S. export controls imposed on **October 7, 2022**. In particular, **those controls** restricted the sale of chips that exceeded performance thresholds across two metrics: total processing power and interconnect speed.

Government sources told CSIS that the U.S. government knew that these restrictions would prevent Nvidia from shipping its (at the time) market-leading A100 chip and upcoming H100 chips. They inferred that Chinese customers would be restricted to using the V100, which was first **introduced** in

Figure 1: Timeline of Major Events in DeepSeek's Development

Event Type Corporate Government GPU acquisition Technology

2015

Liang Wenfeng and two classmates from Zhejiang University found High-Flyer, a technology-focused algorithmic trading hedge fund.

2019

High-Flyer develops Firefly One, a proprietary deep learning training platform with around 1,000 graphics processing units (GPUs) of unknown type.

JULY 17, 2023

High-Flyer founds DeepSeek—an AGI research lab dedicated to open-source releases—as a wholly owned subsidiary with Liang Wenfeng as CEO.

NOVEMBER 29, 2023

DeepSeek releases DeepSeek LLM (aka "DeepSeek-V1"), its first general-purpose LLM.

- DeepSeek conducted a comprehensive study of the scaling laws of the data and model scale, "successfully" revealing the optimal model/data scaling-up allocation strategy and predicting the "expected performance" of future models.
- DeepSeek further found that "data quality significantly influences the optimal model/data scaling up allocation strategy" and that the "higher the data quality, the more the increased compute budget should be allocated to model scaling."
- V1 reportedly outperformed Meta's LLaMA2-70B in general capabilities and GPT-3.5 in certain open-ended evaluations.

FEBRUARY 5, 2024

DeepSeek releases DeepSeekMath, a small, math-specific language model with DeepSeek's first implementation of reasoning capabilities.

- DeepSeek trained this model with a novel and more efficient reinforcement learning algorithm, which they would later reuse in training DeepSeek-R1.
- They also found that, for math capabilities, starting from a code model results in better performance than starting from a more general-purpose language model.
- DeepSeekMath was the first open-source model to score above 50 percent on the MATH benchmark, and DeepSeek claimed that it matched GPT-4 on solving math problems.

NOVEMBER 2024

By this time, DeepSeek had reportedly acquired at least 50,000 Nvidia Hopper GPUs (10,000 H100, 10,000 H800, and 30,000 H20).

DECEMBER 26, 2024

DeepSeek releases DeepSeek-V3, which leverages the innovations of V2 while adding new ones.

- DeepSeek designed V3 to generate multiple words at a time during training, improving performance by making the final model "think ahead" more in its internal representations.
- DeepSeek also applied an in-house strategy for maintaining model performance while preventing over-reliance on any single expert, further improving performance and reducing costs.
- V3 had comparable performance to more expensive Western models, including GPT-4o and Anthropic's Claude-3.5-Sonnet.
- DeepSeek claimed a total training cost (final model only, not including R&D or infrastructure) of $5.6 million.

JANUARY 20, 2025

DeepSeek CEO Liang Wenfeng delivers remarks at Premier Li Qiang's (third-highest ranking official) symposium to inform China's "Government Work Report (Draft for Solicitation of Comments)."

JANUARY 21, 2025

Chinese Vice Premier Ding Xuexiang (fourth-highest ranking official) delivers a speech in Davos indirectly praising open-source AI, DeepSeek, and R1 and its geopolitical implications.

FEBRUARY 17, 2025

Liang and other Chinese entrepreneurs, including the CEOs of Alibaba and Huawei, meet with President Xi Jinping.

OCTOBER 21, 2016

High-Flyer launches its first deep learning AI model for trading.

2021

High-Flyer acquires over 10,000 Nvidia A100 chips to construct Firefly Two, with some estimates up to 50,000. At most four other companies in China had this many chips.

NOVEMBER 2, 2023

DeepSeek releases its first models, the DeepSeek-Coder model series.

- With these models, DeepSeek introduced new methods for improving code generation and conducted one of the first empirical studies of how to best train a large language model (LLM) to "fill in the blank" between two blocks of code in the same file.
- These innovations enabled the Coder models to reportedly achieve state-of-the-art performance among open-source models and superior performance to closed-source models like OpenAI's Codex and GPT-3.5.

JANUARY 10, 2024

DeepSeek releases DeepSeekMoE, which uses novel improvements on the Mixture-of-Experts (MoE) architecture.

- This novel architecture improves performance at a given compute budget by using a greater number of smaller, specialized expert subnetworks while adding a few generalist subnetworks.
- These optimizations allowed DeepSeekMoE 2B to match the performance of other MoE architectures while using 67 percent fewer parameters and match the performance of LLaMA-2 7B while being 40 percent more efficient.

MAY 6, 2024

DeepSeek releases DeepSeek-V2, which leverages DeepSeekMoE and introduces multi-head latent attention (MLA).

- MLA made the model's "working memory" 93 percent more efficient, dramatically reducing compute requirements for training and execution.
- MLA would later play a key role in DeepSeek-V3 and R1.
- V2 performed better than V1 while reducing training costs by 42.5 percent.
- DeepSeek also claimed that it matched LLaMA-3 70B on language understanding while using only 28.5 percent of the compute power.
- Its low costs initiated a price war between language model providers in China, which DeepSeek reportedly won.

JANUARY 20, 2025

DeepSeek releases DeepSeek-R1, the first open-source reasoning model to rival OpenAI's o1.

- R1 was built by training reasoning capabilities into V3.
- DeepSeek used their trial-and-error reinforcement learning (RL) algorithm from DeepSeekMath to discover successful reasoning strategies.
- They also demonstrated that RL could significantly improve performance without any initial human reasoning data (i.e., no supervised fine-tuning).
- In a significant result for the diffusion of AI capabilities, DeepSeek demonstrated the feasibility of upgrading small, non-reasoning models into powerful reasoners by training on the outputs of reasoning models for minimal computing costs.
- By matching a closed-source Western model while costing 94 percent less per million input tokens, R1 triggered Nvidia's historic single-day $600 billion market cap erasure.

FEBRUARY 16, 2025

DeepSeek introduces Native Sparse Attention (NSA), its first release since R1.

- NSA achieves significant speedups over traditional attention mechanisms while maintaining performance by leveraging three methods in parallel: compressing the internal representations of groups of tokens, consulting the most relevant groups, and confining attention to analyze local patterns. The fact that DeepSeek designed the algorithm with modern GPU architecture in mind will allow them to make even more efficient use of their hardware in the future.

2017 and has significantly lower performance than its successors. Additionally, the U.S. government assumed that, if Nvidia were to develop a new chip specifically for the Chinese market that exceeded export control performance thresholds in one metric but not the other, this would require the typical **multiyear** AI chip development timeline.

What the U.S. government architects of the October 2022 export control policy did not realize, however, was that Nvidia had a mechanism for post-manufacturing modification of its existing chip products. Specifically, each Nvidia chip is designed for **redundancy, recoverability, and defect tolerance** to minimize the impact of manufacturing defects. Such defects are a significant concern in the advanced node semiconductor industry, and designing chips to mitigate their potential impact is near universal. Nvidia can blow one or more strategically placed electrical fuses on the chip so that a manufacturing defect would usually only reduce the chip's performance or deactivate a redundant component, rather than rendering the entire chip nonfunctional. Industry sources confirmed to CSIS that Nvidia **blew fuses** on A100 chips to **reduce their interconnect speed** (but not their processing power) below the export control performance thresholds, thus creating the A800 product lines.

In essence, the A100 was born with the capability to become an A800 if needed, via fuse settings. The H800, by contrast, did involve some minor Nvidia **re-engineering** and design optimization in the face of export controls. This allowed Nvidia to continue shipping large volumes of advanced AI chips (with performance near the global state of the art) **to China** without violating export controls and without taking multiple years to design new chips specifically for the Chinese market. According to **reporting** by *The Information*, the performance of the A800 and H800 chips was close enough to the performance of the originals that there was effectively no demand for large-scale AI chip smuggling to China during the period in which A800s and H800s were available. Why smuggle when the legally purchased chips are good enough?

The Biden administration ultimately realized that continued sales of the A800 and H800 meant that their policy would not have the intended impact on China's AI ecosystem. Former government officials told CSIS that this was clear internally by December 2022. However, it took the administration 12 months after their initial package of export controls to act. The U.S. government **modified** the export control performance thresholds in October 2023 to block exports of A800s, H800s, and any non-Nvidia chips with comparable performance to China.

During the year-long period in which A800 and H800 chip exports to China were legal, an enormous number of them were sold to China. Precise sales figures are not available, but Nvidia's disclosed sales to China between October 31, 2022, and October 31, 2023, exceeded 9 billion dollars. This does include revenue for chips that were not widely used in AI training or inference, however. See Table 1 for Nvidia's quarterly revenue by billing locations.

If DeepSeek were to admit that it used H100s in its data center to train its models, it would be confessing to activity illegal under U.S. law, since the H100 has never been legally available in China and cannot even legally be purchased by Chinese-firm subsidiaries outside of China. Thus, if DeepSeek did engage in illegal activity, it would have an incentive to conceal this fact. This raises the question of whether DeepSeek lied in its publications when it claimed to have used H800 chips exclusively.

Table 1: Nvidia Quarterly Revenue by Billing Location

Calculated Values Given Values

	FY 2022				FY 2023			
Nvidia Fiscal Year/Quarter	Q1 FY 2022	Q2 FY 2022	Q3 FY 2022	Q4 FY 2022	Q1 FY 2023	Q2 FY 2023	Q3 FY 2023	Q4 FY 2023
Quarter Ended	1-May-21	1-Aug-21	31-Oct-21	30-Jan-22	1-May-22	31-Jul-22	30-Oct-22	28-Jan-23
United States	$ 768	$ 996	$ 1,126	$ 1,459	$ 1,932	$ 1,988	$ 2,148	$ 2,224
Taiwan	$ 1,784	$ 1,961	$ 2,187	$ 2,612	$ 2,777	$ 1,204	$ 1,153	$ 1,852
Singapore	$ 529	$ 564	$ 546	$ 489	$ 454	$ 495	$ 536	$ 649
China (including Hong Kong)	$ 1,391	$ 1,720	$ 2,017	$ 1,983	$ 2,081	$ 1,602	$ 1,148	$ 954
Other countries (ex-Singapore)	$ 1,189	$ 1,266	$ 1,227	$ 1,100	$ 1,044	$ 1,415	$ 946	$ 372
Other include Singapore	$ 1,718	$ 1,830	$ 1,773	$ 1,589	$ 1,498	$ 1,910	$ 1,482	$ 1,021
Total (Given)	**$ 5,661**	**$ 6,507**	**$ 7,103**	**$ 7,643**	**$ 8,288**	**$ 6,704**	**$ 5,931**	**$ 6,051**
Total (Calculated)	**$ 5,661**	**$ 6,507**	**$ 7,103**	**$ 7,643**	**$ 8,288**	**$ 6,704**	**$ 5,931**	**$ 6,051**

	FY 2024				FY 2025			
Nvidia Fiscal Year/Quarter	Q1 FY 2024	Q2 FY 2024	Q3 FY 2024	Q4 FY 2024	Q1 FY 2025	Q2 FY 2025	Q3 FY 2025	Q4 FY 2025
Quarter Ended	30-Apr-23	29-Jul-23	29-Oct-23	28-Jan-24	28-Apr-24	28-Jul-24	27-Oct-24	26-Jan-25
United States	$ 2,385	$ 6,043	$ 6,302	$ 12,236	$ 13,496	$ 13,022	$ 14,800	$ 19,939
Taiwan	$ 1,796	$ 2,839	$ 4,333	$ 4,437	$ 4,373	$ 5,740	$ 5,153	$ 5,307
Singapore	$ 762	$ 1,732	$ 2,702	$ 3,370	$ 4,037	$ 5,622	$ 7,697	$ 6,328
China (including Hong Kong)	$ 1,590	$ 2,740	$ 4,030	$ 1,946	$ 2,491	$ 3,667	$ 5,416	$ 5,534
Other countries (ex-Singapore)	$ 659	$ 153	$ 753	$ 115	$ 1,647	$ 1,989	$ 2,016	$ 2,223
Other include Singapore	$ 1,421	$ 1,885	$ 3,455	$ 3,484	$ 5,684	$ 7,611	$ 9,713	$ 8,551
Total (Given)	**$ 7,192**	**$ 13,507**	**$ 18,120**	**$ 22,103**	**$ 26,044**	**$ 30,040**	**$ 35,082**	**$ 39,331**
Total (Calculated)	**$ 7,192**	**$ 13,507**	**$ 18,120**	**$ 22,103**	**$ 26,044**	**$ 30,040**	**$ 35,082**	**$ 39,331**

Source: Nvidia Quarterly and Annual SEC Filings; CSIS analysis.

Industry analyst Ben Thompson has pointed to **strong evidence** that DeepSeek did in fact use H800s as it claimed: Many of DeepSeek's algorithmic and architectural improvements are ideal for maximizing the effective use of computing resources under conditions of limited interconnect bandwidth. At a minimum, this strongly suggests that DeepSeek uses many H800s in its computing infrastructure and that a model with the performance of V3 can indeed be trained exclusively on H800s.

This does not prove, however, that DeepSeek exclusively uses H800s in its overall computing infrastructure. Indeed, some **reporting** in Chinese news media claims that DeepSeek did train its slightly more recent R1 model on Nvidia H100 chips that are banned from China under U.S. export controls. The semiconductor consulting firm SemiAnalysis, citing anonymous industry sources, recently **wrote** that DeepSeek has a total of 50,000 Hopper generation GPUs, a category that includes H100s, H800s, and H20s. SemiAnalysis specifically claimed that it has evidence that High-Flyer's/DeepSeek's computing infrastructure includes at least 10,000 H100s, 10,000 H800s, 30,000 H20s, and 10,000 A100s as part of its total computing stack. SemiAnalysis further estimated that High-Flyer/DeepSeek spent $1.63 billion in GPU server capital expenditures alone (i.e., excluding other data center construction and operating costs).

Nathan Lambert, a machine learning researcher at the Allen Institute of Artificial Intelligence, has correctly **pointed out** that

> Tracking the compute used for a project just off the final pretraining run is a very unhelpful way to estimate actual cost. It's a very useful measure for understanding the actual utilization of the compute and the efficiency of the underlying *learning*, but assigning a cost to the model based on the market price for the GPUs used for the final run is deceptive.

In this case, it is not DeepSeek that is being deceptive, however, since their paper was extremely clear about what was and was not included in the $5.6 million figure. It was only the final, successful pretraining run. It did not include the compute cost of the hundreds of experiments that were run beforehand to generate the insights necessary for that final run, nor did it include the compute costs associated with post-training fine-tuning or inference compute workloads.

While there is no public evidence that large-scale Nvidia chip smuggling occurred prior to October 2023, industry sources told CSIS that, by early 2024, large-scale H100 chip smuggling operations were underway. In mid-2024, journalists at ***The Information*** interviewed participants in eight distinct H100 smuggling networks, each of which provided evidence that they had completed Chinese transactions worth more than $100 million. These networks continue to be active, with increasingly **sophisticated techniques** for evading detection. For example, a December 2024 investigation by *The Information* found that

> When notified of an upcoming inspection, smugglers have duplicated the serial numbers of the servers with Nvidia chips they've purchased from Supermicro [and already smuggled to China] and attached them to other servers they had access to.

There is some evidence that governments are becoming more serious about cracking down on AI chip smuggling. Singapore was the billing location for 18 percent of Nvidia's revenue in the most recently **concluded fiscal year** but the shipping location for only 2 percent of revenue. Officials in Singapore recently **arrested** three individuals accused of facilitating AI chip smuggling to China. The Singaporean

government's increased emphasis on cracking down on smuggling may reflect its **concerns** about being relegated to tier 2 status under the Biden administration's **AI Diffusion Framework**.

5. **However, DeepSeek's challenges also reflect the successes of the AI chip export control policy, which DeepSeek's CEO has described as the greatest challenge facing his company and the wider Chinese AI ecosystem.**

As with the introduction of **Huawei's Mate60 Pro** 5G smartphone in August 2023, many commentators were quick to highlight how DeepSeek reflected the failure of U.S. export controls. This included Chinese state-owned media, one outlet of which, **GlobalVision**, **broadcast** that "DeepSeek and Huawei join forces! . . . breaking through U.S. tech restrictions is not impossible!"

Angela Zhang, a professor at the Gould School of Law, went further, **arguing**:

> China's achievements in efficiency are no accident. They are a direct response to the escalating export restrictions imposed by the US and its allies. By limiting China's access to advanced AI chips, the US has inadvertently spurred its innovation.

This argument is in one sense true: Chinese firms like DeepSeek prioritized innovation in efficiency—and particularly efficiency innovations that are useful given the limitations of the H800—because that was the most attractive AI research path available to them.

However, this argument is deeply flawed for three reasons. First, every leading AI firm, including both U.S. and Chinese ones, is pursuing enhanced efficiency. As mentioned above, DeepSeek's efficiency improvements represented a continuation of preexisting industry trends in AI computation efficiency improvements, not a drastic acceleration of them. Even before DeepSeek was founded, and even before the U.S. imposed export controls, DeepSeek's parent company was obsessively focused on improving computing efficiency (though more focused on optimizing latency). It is strange to say that export controls caused AI firms to do something that they—both U.S. and Chinese—were already doing (and already had a massive financial incentive to do) before export controls were imposed.

Second, the argument is flawed because it does not engage with the question of "how much further along might China be in the absence of export controls?" The largest AI computing cluster currently operating on Earth is xAI's Colossus supercomputer in Memphis, Tennessee, which was recently upgraded to use **200,000 H100 chips**. Given that leading Chinese firms were already **producing more** AI-related patents, publishing nearly as many highly cited research papers, and **spending nearly as much** on data center capital expenditures before the export controls went into effect, it is entirely possible that China would already be ahead of the United States in AI in the absence of export controls. Jimmy Goodrich, a non-resident senior associate at the Wadhwani AI Center, put it well when he **said**, "It's been long known that DeepSeek has a really good team, and if they had access to even more compute, God knows how capable they would be." Goodrich's statement on DeepSeek applies to China as a whole.

Third, this argument does not take account of the benefits of reallocating the supply of high-performance chips that would otherwise have gone to China. Nvidia, which is estimated to have **90 percent** global market share in AI chips, is supply constrained, not demand constrained. For every Nvidia chip that TSMC has been able to manufacture since 2022, there are many customers willing to

buy them. Larry Ellison, the cofounder of Oracle, colorfully described how Nvidia's customers have to beg for greater **chip allocation** in September 2024 when he **recounted** a dinner between himself, xAI CEO Elon Musk, and Nvidia CEO Jensen Huang:

> I would describe the dinner as Oracle—me and Elon begging Jensen for GPUs, [saying] "Please take our money. Please take our money. By the way, I got dinner. No, no, take more of it. We need you to take more of our money please."

What this means is that, since the AI chip shortage began in 2022, all the chips that Nvidia would have sold to China in the absence of export controls would have come at the expense of customers elsewhere, which in this case is overwhelmingly the United States. In other words, eliminating the export controls would have not only helped China, it would have also hurt the United States, at least in terms of Nvidia chip allocation.

For his part, DeepSeek CEO Liang Wenfeng is under no illusions that the chip export controls have helped China or DeepSeek. In a July 2024 interview, he **said**, "We do not have financing plans in the short term. Money has never been the problem for us; bans on shipments of advanced chips are the problem." Liang has also said in an interview with a Chinese media outlet that U.S. restrictions on AI chips mean that Chinese companies must use **two to four times** the computing power to achieve the same results, referring to the penalty of using the H800 instead of the H100 for large model training.

Further evidence of the impact of U.S. export controls comes from the fact that DeepSeek experienced widespread outages and had to **halt new user sign-ups** in the wake of its massive user adoption growth. While DeepSeek claimed that the outages were the result of cyberattacks, the far more likely explanation is that they simply lacked sufficient computing capacity to serve their growing user base. There is a big difference between the amount of computing power required to serve an AI large language model to a small number of users and doing what OpenAI does, which is serve its AI models to **400 million active users** every week. More recently, DeepSeek has **once again** reopened new user sign-ups, which likely reflects the company successfully purchasing or renting additional computing capacity. Nevertheless, the point remains: Compute capacity matters.

6. **DeepSeek's discovery of techniques for increased AI computational efficiency could benefit U.S. firms more than Chinese ones, as U.S. firms can apply such techniques to their much larger computing resources and thus deliver better AI to more customers.**

In describing its goals toward creating AGI, a recent DeepSeek job advertisement **said** that "We believe AGI is the violent beauty of model x data x computing power." Though a bit oversimplified, that formula is still essentially correct and helps explain why having larger computing resources is a useful advantage at any level of computational efficiency.

All of DeepSeek's innovations were algorithmic and architectural. DeepSeek appears to have described all or nearly all of them in great detail in its research papers. That means that U.S. AI labs are free to apply these same innovations in training and deploying their own AI models. Indeed, U.S. firms are **already doing so**.

But while DeepSeek's algorithmic innovations are replicable by U.S. firms, DeepSeek will struggle to replicate U.S. AI chip and compute advantages. In the **absence** of extreme ultraviolet lithography

(EUV) technology, China's most advanced AI chip designer and logic chip manufacturer, Huawei and SMIC, respectively, will most likely remain stuck at 7 nanometers (nm) or perhaps a **flawed 5 nm** technology node for many years. The CEO of ASML, which is the sole firm in the world manufacturing EUV lithography machines, **said** "By banning the export of EUV, China will lag 10 to 15 years behind the West. That really has an effect."

By contrast, U.S. chip firms such as Nvidia will in coming years press forward past the 4 nm era into the **2 nm and beyond** future, delivering chips that might be thousands or tens of thousands of times better performing integrated into data centers using **tens** or hundreds of times more chips. While there are still many ways in which the U.S. government could bungle export control implementation and enforcement—which will be discussed later in this paper—the greatest potential for impacts from the export controls are still well ahead in the future as the United States races forward and China's progress continues to slow. Implemented properly, they offer the opportunity for a meaningful, though far from permanent, edge over China.

7. **Just as the past six decades of Moore's Law have driven increased demand for computer chips, DeepSeek and other AI efficiency innovations will continue to do the same for AI chips. There is no ceiling on the demand for intelligence.**

Much of the initial **commentary** on DeepSeek suggested that demand for AI chips might plummet, as efficiency improvements meant that fewer chips were needed to achieve the same level of intelligence. While this is a generally accurate description of efficiency, it is a terrible description of how computing economics works.

As mentioned above, Moore's Law has driven a more than billionfold improvement in computing efficiency over the past seven decades. Does this mean that all of the world's computing needs are handled by a single computer that costs only a fraction of a penny? Of course not. Instead, the increased efficiency has made applying digital computation throughout the economy far more attractive, and so more governments, businesses, and consumers have invested in more computation across countless applications and use cases. This is the essence of **Jevon's Paradox**, which describes how increasing efficiency can increase aggregate demand.

The same is obviously going to be true of increased AI computational efficiency. Indeed, after the DeepSeek revelations, the largest data center buyers in the world—Microsoft, Google, Meta, and Amazon—**announced** that they were going to spend hundreds of billions of dollars in 2025 on AI chips and data centers, up by nearly 50 percent over the already record amount invested in 2024. Chinese firms such as Alibaba have **announced** that they too will make massive chip purchases in the wake of DeepSeek.

Nor is this merely a long-term phenomenon. Even in the short term, rental prices for H100 chips hosted by Amazon's cloud computing **spiked** upward in the wake of DeepSeek's publication as customers worked to incorporate DeepSeek models into their operations. More recently, OpenAI CEO Sam Altman **said** that OpenAI would have to temporarily limit usage of its GPT 4.5 model because the company did not have sufficient GPUs to meet all of the demand.

Moreover, there is almost certainly no "ceiling" for a desirable amount of intelligence and the computing required to power that intelligence. Miles Brundage, who previously led policy research at OpenAI, **described** it this way:

> To make a human-AI analogy, consider [Albert] Einstein or John von Neumann as the smartest possible person you could fit in a human brain. You would still want more of them. You'd want more copies. That's basically what inference compute or test-time compute is—copying the smart thing. It's better to have an hour of Einstein's time than a minute, and I don't see why that wouldn't be true for AI.

This again relates to how the most significant impacts of the export controls are likely in the future, as the evermore extremely scaled deployment of evermore powerful chips begins to bear fruit. It is the next few rounds of increasing AI computation—by tenfold, a hundredfold, a thousandfold—where China will be most severely blocked and the impact of the export controls will be most severely felt.

Some Chinese AI researchers have made similar points. On February 13, 2025, Li Guojie, a scholar at the Chinese Academy of Engineering, **told** Chinese media outlet *ChinaFund* that DeepSeek's success represents a step-change in Chinese AI capabilities but said that "due to the blockade of the U.S. government, China is currently unable to obtain the most advanced chip process technology." Li went on to **say** [machine translation]:

> The success of DeepSeek does not deny the key role of computing power in the development of artificial intelligence. In fact, since there are much more equipment for reasoning than for training, the computing power required for reasoning [aka inference] will become the main requirement in the future. It is very important to improve the efficiency of the model through algorithm optimization. It is our basic choice to take the green development path of saving computing power, but computing power is definitely a necessary condition for solving artificial intelligence problems and cannot be ignored.

8. **DeepSeek's success in distilling U.S. AI models and replicating closed source algorithmic innovations does raise strategic questions about the nature of competitive advantage in AI in the absence of strong intellectual property protections. A worst-case scenario would be if AI is structurally similar to pharmaceuticals.**

On January 29, OpenAI **told** the *New York Times* that

> We know that groups in the P.R.C. are actively working to use methods, including what's known as distillation, to replicate advanced U.S. A.I. models. . . . We are aware of and reviewing indications that DeepSeek may have inappropriately distilled our models.

Distillation is a well-known technique that **allows** "the transfer of knowledge from a larger, more complex language model (the 'teacher') to a smaller, more efficient version (the 'student')." In short, the inputs and outputs from the teacher model are used as very high-quality synthetic training data by the student model. Distillation is utilized by every major U.S. AI lab to offer smaller, more efficient versions of its models that are faster and cheaper to serve to customers. For example, GPT-4o-mini is the distilled version of GPT-4o, and both were made by OpenAI. There are multiple types of distillation,

however, and the more advanced types, which AI labs use to distill their own models, require access to the internals of the models.

While a firm distilling its own models is clearly harmless, there is a strategic consideration if one firm can distill another firm's models, incurring most or all of the benefits of acquiring a highly capable AI model while incurring far fewer of the costs of creating such a model.

In the case of OpenAI and DeepSeek, the distillation evidence appears quite strong: DeepSeek's initially released models, when asked by users "what model are you?," will **often respond**, "I'm ChatGPT." DeepSeek's updated versions of the model, likely modified using fine-tuning techniques, no longer produce such responses, but this does not reduce the strength of the evidence that DeepSeek relied on distillation. Such unauthorized distillation would be a violation of OpenAI's terms of service, but the law firm Fenwick has **written** that "whether remedies against such [a Terms of Service] breach would sufficiently protect teacher-model owner's IP remains to be determined."

If OpenAI were to file a lawsuit against DeepSeek, for example, perhaps it would win in U.S. courts. But how would the size of judicial penalties compare to the strategic benefit for DeepSeek as a firm or for China as a country? And the access to judicial remedies is almost certainly limited to the United States and other countries with strong intellectual property protections. Even if DeepSeek were hypothetically clearly guilty of not only distillation but an even more extreme case of outright corporate espionage, Chinese courts would likely find OpenAI guilty of stealing IP from DeepSeek rather than the reverse.

This is what happened to U.S. chipmaker Micron when it accused China's Fujian Jinhua of stealing memory chip trade secrets and when U.S. prosecutors charged Fujian Jinhua with economic espionage. In retaliation, Jinhua sued Micron in a Chinese court for patent infringement in early 2018. The Fuzhou Intermediate People's Court in Fujian province **sided with the Chinese side** and issued an injunction blocking Micron from selling 26 chip products in China.

Similarly, when U.S. semiconductor equipment firm Veeco found that a Chinese competitor, Advanced Micro-Fabrication Equipment (AMEC), was using Veeco's proprietary designs, Veeco secured a **preliminary injunction** in a U.S. court to stop a supplier from providing parts to AMEC. However, in response, AMEC filed a patent suit in China, and the Fujian High Court (the same province as above) acted with remarkable speed. On December 7, 2017—just weeks after AMEC's filing—the court granted AMEC a preliminary injunction **without even hearing** Veeco's defense.

Thus, one should separate the questions of what is "fair" or "true" from what the strategic consequences are likely to be. Neither geopolitics nor business is a contest where first prize always goes to the most innovative. Many companies, including **U.S. tech companies**, have a perfectly viable business strategy of closely monitoring their innovative competitors and then either buying the competitor or replicating its innovation. Chinese firms can do the same, including in ways that both do and do not violate intellectual property.

Thus, the meaningful strategic question for OpenAI and other U.S. AI firms is not whether DeepSeek is failing to play by the rules, it is whether U.S. firms have a mechanism for effectively preventing DeepSeek or any competitor from extracting the benefits of an AI lab's investments without bearing the costs. What enduring sources of competitive advantages could the large U.S. AI labs secure if Chinese

companies or other copycats (potentially even U.S. ones) can simply piggyback off their work to build cheaper and equivalently good models while incurring only a fraction of the costs?

One possible answer comes in the form of Meta's strategy, which is to invest in developing open-source AI models justified not by the direct benefits of increasing monthly subscriptions, but by the indirect benefits of enhancing their existing social media and advertising business lines.

Among the worst outcomes for U.S. AI companies would be if the fundamentals of AI research and development were structurally similar to those of the global pharmaceutical industry. In 2023, the top 20 global pharmaceutical companies invested approximately **$2.4 billion** over **10 to 15 years** into developing a typical new drug from start to finish. The costs of research and development dwarf those of drug manufacturing, marketing, and distribution, so the final price of the product must reflect the costs of research and development for pharmaceutical firms to have any hope of recovering their investments and earning a profit.

To incentivize drug R&D, the U.S. government offers companies **20-year patents** that allow them the exclusive right to sell their products in the United States at a monopoly premium to recover upfront R&D costs. However, once a patent expires, so-called generic manufacturers can reproduce and sell the same drug for a far lower price—avoiding the burden of R&D costs altogether. Generic manufacturers can thereafter compete by selling identical drugs while incurring only a tiny fraction of the costs faced by the original drug developer. In a world where there was no patent protection on new pharmaceuticals, investment in new drug discovery (and thus the pace of innovative new drugs) **would be radically limited** because governments and philanthropic actors might be the only ones willing to invest, and they would almost certainly do so at far lower levels than the current private sector does.

Government and business-makers alike must now ask the question: What if frontier AI model research is structurally identical to that of the pharmaceutical industry, meaning extremely high upfront R&D costs and very low replication costs? The implications for AI investment and for U.S.-China AI competition could be significant.

The U.S.-China AI race is very likely not one in which gentlemanly concepts of "fairness" matter. Chinese companies have **systematically stolen American IP** for **years** and will almost certainly continue to do so. Beyond Chinese firms, the Chinese government has a **very clear history** of either not enforcing or **outright supporting** Chinese firms that illegally violate U.S. patents in cases where the Chinese government has identified the industry as strategically significant. AI certainly falls into that category.

Kai-Fu Lee, arguably the top venture capitalist in China's AI sector, wrote in his 2018 book ***AI Superpowers*** that imitation does not have the same stigma in China that it does in the United States. He went further to argue that China's unabashed willingness to copy innovation and even ignore intellectual property makes its companies more competitive, not less. In describing the post-2010 rise of hypercompetitive and cutthroat Chinese tech firms as "copycat gladiators," Lee's writing on this topic is worth quoting at length:

> Silicon Valley may have found the copying undignified and the tactics unsavory. In many cases, it was. But it was precisely this widespread cloning—the onslaught of thousands of mimicking

> competitors—that forced companies to innovate. Survival in the internet coliseum required relentlessly iterating products, controlling costs, executing flawlessly, generating positive PR [public relations], raising money at exaggerated valuations, and seeking ways to build a robust business 'moat' to keep the copycats out. Pure copycats never made for great companies, and they couldn't survive inside this coliseum. But the trial-by-fire competitive landscape created when one is surrounded by ruthless copycats had the result of forging a generation of the most tenacious entrepreneurs on earth. As we enter the age of AI implementation, this cutthroat entrepreneurial environment will be one of China's core assets in building a machine-learning-driven economy.

Taking Lee's argument one step further: If the future global competitive landscape for AI is one of ruthless copying in the absence of robust intellectual property protections, that is an environment where Chinese entrepreneurs ought to do extremely well, because it resembles the domestic Chinese tech ecosystem in which they grew up and thrived.

9. Unfortunately, developments in recent years do not give confidence that traditional approaches to protecting intellectual property and technology secrets are likely to be effective when it comes to China and AI.

Even over just the past three years, a diverse set of examples suggests that robustly protecting key knowledge in the AI field is likely to be extremely difficult:

- In February 2025, a federal grand jury **charged** a Chinese national and former Google employee with "seven counts of economic espionage and seven counts of theft of trade secrets in connection with an alleged plan to steal from Google LLC (Google) proprietary information related to AI technology."
- In February 2023, ASML, the most important provider of the lithography equipment used to manufacture advanced chips, **accused** a former employee in China of stealing proprietary technology data in an attempt to circumvent export controls.
- In January 2025, Eindhoven University of Technology, a school only five miles away from ASML's headquarters and with close ties to the company, had to **suspend** all activities as it worked to analyze the impact of a Chinese cyberattack.
- In 2022, WIRED **reported** that a group of cyber criminals breached Nvidia and stole "a significant amount of sensitive information about the designs of Nvidia graphics cards, source code for an Nvidia AI rendering system called DLSS, and the usernames and passwords of more than 71,000 Nvidia employees." Nvidia claims to have since significantly upgraded its cybersecurity, and this particular attack does not seem to have originated in China. However, it is concerning that a hacking group got so close to the technology crown jewels of such an important U.S. AI company.
- In 2023, Chinese intelligence services successfully **hacked** the email accounts of then-Secretary of Commerce Gina Raimondo and other U.S. officials, apparently looking for intelligence related to U.S. export controls.

- In 2025, *The Information* **reported** that DeepSeek CEO Liang had visited San Francisco in mid-2024 and "met with researchers he knows, including some OpenAI employees, to stay up to date." This is consistent with Center for AI Safety Director Dan Hendrycks' **claim** that

> If you know which research dead ends to avoid (e.g., MCTS) and roughly know what research direction to pursue, replication is much easier. These algorithmic insights likely slowly leaked through SF party conversations and other channels. It probably takes a few months once you have the idea for [OpenAI's reasoning AI model] o1 to replicate it. That's how DeepSeek did it without much compute.

While there is ongoing **research** into **techniques** to prevent model distillation, and AI companies will no doubt increase these efforts in the wake of DeepSeek, this remains an open question with significant implications. In an interview published in *ChinaTalk*, Anthropic CEO Dario Amodei **said**,

> Like all problems in cybersecurity . . . [distillation is] going to be an ongoing issue. That's why some companies have chosen to kind of hide the chain of thought in their reasoning models, because it makes it much harder to distill. Those can be jailbroken. But folks are working on antidotes to jailbreaking. We just released something today that makes it **much harder to jailbreak models**.

Even in a case where distillation and more explicit cases of technological espionage are difficult or impossible to prevent, however, there will remain the issue inference computation, which involves applying more computational resources so that the model can "think" harder about the answer to a question or so that the model can be used more times and serve more users. One way of thinking about this is that the nature of competitive advantage in AI could be shifting from a secret that one keeps (i.e., the **model weights**) toward an asset that one controls (i.e., massive data center compute infrastructure). This means that export controls—if implemented and enforced effectively—still have an opportunity to have a strategic impact.

Nevertheless, the U.S. government and U.S. AI firms need to take a tough look at whether and how they can remain competitive in a world where China does not respect intellectual property and is committed to devoting enormous state resources to technical espionage in the AI and semiconductor sectors.

10. The effectiveness of export controls will depend upon effective implementation and enforcement to prevent chip smuggling. There is cause for significant concern on this point if the Trump administration reduces relevant government resources and staff or causes allied countries to not cooperate.

The Bureau of Industry and Security (BIS) at the U.S. Department of Commerce is the agency charged with adjudicating license applications and enforcing export controls, not just on semiconductors bound for China, but for all U.S. dual-use technology exports that might end up in Russia, Iran, North Korea, and many other restricted destinations. To implement its work overseeing trillions of dollars in economic activity and policing smuggling operations worldwide, BIS has fewer than 600 employees and a relatively paltry budget of just under $200 million. Semiconductors are just one technology category out of hundreds that this organization is responsible for enforcing.

Reporting by *The Information* **found** at least eight Chinese AI chip-smuggling networks, with each engaging in transactions valued at more than $100 million. More recent **reporting** by the *Wall Street Journal* suggests that this smuggling now includes the latest generation of Nvidia Blackwell AI chips. China is betting that its network of smugglers and shell companies can find the leaks in the BIS export control enforcement barrier. As long as Congress continues to neglect BIS by providing grossly inadequate resources compared to the size and importance of its mission, China has a reasonable expectation of success. BIS needs not only more money, but also more skilled staff, more enforcement agents, and better enabling technology, especially in **data analysis**.

Moreover, the Department of Commerce needs more help from the rest of the government, in particular the U.S. intelligence community. Declassified Central Intelligence Agency **documents** show that the intelligence community was deeply involved in assisting export control enforcement during the Cold War and delivered solid results by doing so. These are capabilities and priorities that have significantly atrophied in the post-Cold War era but urgently need to be restored.

Recent efforts by the Trump administration to reduce government spending risk making a bad situation worse. On February 5, 2025, U.S. Attorney General Pam Bondi signed a Department of Justice **memorandum** that, among many other actions, disbanded the National Security Division's Corporate Enforcement Unit, which **was** "created with the goal of investigating and prosecuting corporate actors involved in sanctions evasion, export control violations, and other national security-related economic crimes." The memo did not provide any explanation for how the duties of the unit would be performed elsewhere in the Department of Justice, which could mean that this will be implemented as a straightforward reduction in U.S. government capacity to enforce export controls.

Of all the choices available, the worst policy choice that the U.S. government could make is to regulate aggressively but with large loopholes and weak enforcement. Such an approach—which fairly describes the approach taken in the first Trump administration and large chunks of the Biden administration—incurs nearly all the costs of an aggressive export control policy but does so without providing any of the possible strategic benefits in terms of slowing Chinese progress in semiconductors and AI.

11. Beyond smuggling, the greatest strategic challenge for the United States is the potential for China to produce AI chips domestically at sufficient quantity and quality to build AI data center infrastructure that is competitive with the United States.

As described above, the nature of AI geopolitical advantage appears to be moving away from software secrets that are difficult to protect and toward hardware chip and data center assets that are (hopefully) easier to protect. In this sense, DeepSeek is not in and of itself the most significant threat to U.S. leadership in AI. Instead, the greater challenge arises from the possibility of China having a domestic ecosystem for producing its own AI chips at large scale and integrating them into Chinese data center training, as well as running inference for DeepSeek and other AI models.

As U.S. technology firms are planning hundreds of billions of dollars in AI data center infrastructure investments, it is worth remembering that—for those investments to be possible—companies like TSMC must manufacture enough AI chips to fill those data centers. In the case of companies such as Nvidia, their revenue growth in recent years is less than it would have otherwise been due to **shortages** of TSMC production capacity.

The Biden administration took many steps designed to definitively cut China's AI chip designers off from TSMC production capacity. Most recently, on January 15, the Commerce Department announced the **final tranche** of Biden administration export controls, often referred to as "the Foundry Rule." The Foundry Rule moved advanced chip production to a **white-list** system that will likely make it impossible for Chinese AI firms to access TSMC capacity to produce chips above export control performance thresholds even when operating through complex shell company arrangements. However, as discussed below, TSMC manufactured a strategically significant quantity of chips on behalf of Huawei via shell companies prior to the rule going into effect.

That effectively means that China's long-term future in AI is closely tied to its ability to produce AI chips domestically. The Biden administration sought to hamstring China's domestic production of advanced chips by restricting the sale of advanced semiconductor manufacturing equipment, including from other countries.

12. DeepSeek stands out among Chinese AI model developers because of its demonstrated ability to make architectural innovations below the level of Nvidia's CUDA software ecosystem. If DeepSeek were devoted to strengthening Huawei's Ascend chips and CANN software ecosystem, it would pose a much more significant threat to Nvidia.

Historically, one of the biggest sources of **competitive advantage** for Nvidia has been its CUDA (Compute Unified Device Architecture) software ecosystem. CUDA makes it **much easier** for programmers to write massively parallelized software (as all modern AI software is) and ensures backward and forward compatibility so that older chips can still run newer software and vice versa. Any customer who seeks to stop using Nvidia chips has to leave the CUDA ecosystem, which requires solving a lot of incredibly hard software problems for which CUDA already provides free answers. Those free answers reflect billions of dollars of investment in the CUDA platform by both Nvidia and its customers.

DeepSeek stands out among AI model developers because it implemented software changes below the CUDA level. Most AI researchers rely on CUDA, a higher-level programming language for NVIDIA GPUs, because it simplifies the development process. However, CUDA does not allow fine-grained control over GPU operations at the level DeepSeek required. DeepSeek modified software instructions written in a language called PTX (Parallel Thread Execution). By programming directly in PTX, DeepSeek engineers were able to optimize how the GPU handled specific workloads, particularly in managing expert specialization in the Mixture-of-Experts model and improving memory efficiency. This is an "assembly-level" approach to GPU programming—one that is both extremely difficult and (at least historically) rarely attempted. The trade-off, however, was a major increase in performance efficiency, particularly for large-scale AI training tasks.

This suggests that DeepSeek may have both the technical knowledge and the open-source community enthusiasm to finally start generating momentum around Huawei's competitor to CUDA, which Huawei **refers to** as its Compute Architecture for Neural Networks (CANN). The greater maturity of the CUDA software ecosystem currently makes Nvidia chips more attractive, but this **could change** over the next few years. If it does, it would have major implications for Nvidia's competitiveness both inside and outside of China.

At present, the combination of Ascend chips (discussed more below) and CANN software appears far from being competitive to CUDA. A *Financial Times* report from September 2024 **found** that even Huawei employees found the product "difficult and unstable to use" and prone to crashing frequently. Industry sources told CSIS that DeepSeek's more recent evaluation of CANN was very negative and that DeepSeek assessed that it would be years before the combination of Ascend chips and CANN-compatible software was a viable alternative.

Still, this could change with time, effort, and investment. Huawei recently **joined** the open-source PyTorch foundation in an effort to increase Ascend and CANN compatibility of the PyTorch AI development framework.

There is at least one meaningful precedent regarding the difficulty of leaving the Nvidia CUDA ecosystem, which is Google's shift from TensorFlow on Nvidia GPUs to JAX on their proprietary chips, called TPUs. In this case, it took several years for JAX to mature and for a critical mass of software libraries to appear—even with heavy Google investment and the payoff opportunity in terms of both efficiency and competitive differentiation. Even by 2023, Google was still heavily investing to **improve** JAX (for example, addressing pain points like multi-host setup and memory efficiency) and encouraging its use through education (tutorials, conference workshops). For a company like DeepSeek, migrating all AI workloads from CUDA to CANN would likely be a multiyear project. Google's internal shift took 2–3 years to bear significant fruit, and even then Google maintained support for both for a long while.

13. China's alliance of Huawei (AI chip designer), SMIC (AI chip manufacturer), and CXMT/XMC (high-bandwidth memory manufacturers) have recently made strategically significant progress in advancing domestic production of AI chips.

Domestically producing large quantities of AI chips will require China to domestically replicate multiple segments of the AI chip value chain. The most important links are AI chip design, advanced node logic chip manufacturing, and advanced node high-bandwidth memory (HBM) manufacturing. Each of these will be addressed in turn.

14. Huawei's AI Chip design has long been strong. Before the 2020 export controls, it was poised to meaningfully challenge Nvidia in China with TSMC manufacturing and Chinese government pressure to purchase Huawei products over Nvidia ones.

Like the United States, China has many different companies working on AI chip design, including Huawei, Cambricon, Biren, and more. However, Huawei is unambiguously in the **strongest position** with its Ascend AI chip product line. Industry sources also told CSIS that Huawei—due to its highly influential position with the Chinese government and its size as a customer—has the greatest leverage over SMIC. In other words, even if Huawei were not the best AI chip designer in China, it would still be the most important since it is in a position to restrict the share of SMIC manufacturing capacity that goes toward its domestic competitors.

Huawei has been a player in the AI chip market for longer than is generally realized. Huawei had a 7 nm AI Accelerator chip planned for production at TSMC **in 2020**. At the time, Huawei was preparing to capitalize the so-called **3-5-2 policy**, in which the CCP's Central Office ordered all government agencies and **many state-owned enterprises** to eliminate the use of all non-Chinese technology within three

years. However, Huawei was unable to fulfill demand because of the first Trump administration's updated **2020 entity listing**, which temporarily cut Huawei's Ascend and Kirin lines off from TSMC. In this instance, Nvidia was the beneficiary of these early U.S. export controls, which delivered a setback to Chinese efforts to eliminate dependence on U.S. AI chip technology.

Today, Huawei's Ascend product line is focused on the Ascend 910B and the Ascend 910C, the latter of which includes two Ascend 910B logic dies per integrated unit (meaning more real estate on each silicon wafer is required for manufacturing an Ascend 910C than a 910B). DeepSeek has **reportedly** evaluated the Ascend chips and found that they are unattractive for training AI models but that each Ascend 910C delivers roughly 60 percent of the performance of an Nvidia H100 for inferencing AI models. This matters greatly since more and more of the computing requirements for advanced AI models are expected to be devoted to inference in the coming years. Barclays, an investment bank, **estimates** that by 2026, 70 percent of AI compute demand will come from inference. In February 2024, Nvidia CEO Jensen Huang **estimated** that utilization of its chips was "40 percent inference, 60 percent training," compared to 90 percent training-focused in the 2016 timeframe.

15. Sources told CSIS that Huawei's use of shell companies to gain access to TSMC manufacturing capacity allowed them to acquire more than 2 million AI chip dies and that Huawei has also stockpiled more than a year's worth of HBM.

Chinese firms like Huawei have historically had two major options for manufacturing their AI chip designs: outsource production abroad to TSMC of Taiwan or produce chips domestically in partnership with SMIC.

For a time, it seemed as though the 2020 entity listing of Huawei definitively cut the company off from accessing TSMC advanced node manufacturing capacity. The **October 2022** and **October 2023** export controls were supposed to have done the same for all advanced node AI chip designers in China.

However, TSMC manufactured large quantities of Huawei Ascend 910B chips on behalf of **Huawei shell companies** and shipped the chips to China in violation of U.S. export controls. In an interview with the *New York Times*, an unnamed Taiwanese official **acknowledged** that these chips "ended up in the hands of Huawei." Government officials told CSIS that TSMC manufactured more than 2 million Ascend 910B logic dies and that all of these are now with Huawei. If true, this is enough dies to make 1 million Ascend 910C units. However, the advanced packaging process by which two Ascend 910B dies and HBM are **combined** into a unified Ascend 910C chip also **introduces defects** that can compromise the functionality of the chip. Industry sources told CSIS that roughly 75 percent of the Ascend 910Cs currently survive the advanced packaging process.

This is a strategically significant stockpile of AI chips. For comparison, Nvidia **reportedly** shipped 1 million H20 chips (which are **inference specialized** and targeted at the Chinese market) to China in 2024. Nvidia's January 26 SEC filing states that it generated $17 billion in revenues from Chinese customers in the most recently ended fiscal year (which would obviously exclude revenues from chips later smuggled to China). Presumably, the vast majority of these revenues are from H20 sales.

Even though Huawei likely has the more than 2 million Ascend 910B logic dies made by TSMC, there is a question as to whether it has enough HBM to integrate with those dies in manufacturing 1 million

Ascend 910Cs. It seems likely that Huawei does, however, since the U.S. plan to restrict all advanced HBM sales to China on a country-wide basis was leaked to Bloomberg in **August 2024** and did not go into effect until **December** of that year, giving Huawei ample time to legally acquire HBM chips as part of a stockpiling strategy. Industry sources told CSIS that Huawei has stockpiled enough HBM to meet their internal expectations for at least a full year's worth of production, mostly in the form of **purchases from Samsung** of South Korea, potentially via shell companies, that occurred before the December controls went into effect.

It is as though the U.S. government was having an internal debate about whether or not to have a surprise attack, and in the spirit of compromise, the slow interagency process settled on attacking without the surprise. This was an egregious failure with strategic consequences.

With the adoption of the **Foundry Rule**, it will hopefully be impossible for Huawei or other Chinese AI chip designers to ever again access TSMC manufacturing capacity, even when using sophisticated shell company tactics. Huawei will instead be restricted to what Chinese domestic logic chip manufacturers can produce.

However, there is also the potential risk that an existing chip design company on the Foundry Rule white list could be tempted by massive revenue opportunities to act as a pass-through to give Huawei access to TSMC. Guarding against this risk will require making clear to TSMC that the negative consequences of continuing to supply Huawei with advanced AI chips would be colossal compared to the revenue and profit opportunity. Similarly, Huawei may attempt similar tactics with other chip foundries such as Samsung or even Intel.

16. **Huawei's AI chip manufacturing partner, SMIC, has been struggling with low production yield (~20 percent) and 20,000 7 nm monthly wafer production due to U.S. and allied export controls. SMIC has a difficult and uncertain path toward producing at nodes more advanced than 7 nm.**

The most advanced logic chip manufacturer in China is SMIC. SMIC's SN2 facility in Shanghai is the sole facility in China with an active 7 nm logic chip production line and has been producing 7 nm chips since **July 2021**, more than a year before the first tranche of the Biden administration's semiconductor equipment export controls went into effect. SMIC and Huawei are now working to **bring a 5 nm node** into scaled production but must do so without access to EUV lithography equipment, since China has no local producer of EUV lithography machines and since export controls have **prevented** such machines from ever being exported to China. Industry sources told CSIS that, in early 2020, ASML was poised to ship EUV tools to China and that SMIC was planning to work with key research labs in Europe, such as the Interuniversity Microelectronics Centre (IMEC), to help develop their EUV-based manufacturing process.

In December 2024, industry sources told CSIS that SMIC currently has enough immersion deep ultraviolet (DUV) lithography equipment supplied by the Dutch company ASML to produce 85,000 FinFET wafers per month (WPM) across both SN1 (which focuses on 14 nm node production) and SN2 (which focuses on 7 nm and 5 nm production). This acquisition of lithography tools reportedly took effect before **Dutch DUV lithography export controls** went into effect in mid-2023.

However, the bottleneck in expanding 7 nm (which in SMIC's node naming system is called "N+2") production capacity has not been lithography but rather U.S. tools for etching, deposition, inspection, and metrology. The shortage of inspection and metrology tools has also been an important factor preventing SMIC from improving its yield. SMIC acquired lithography equipment as a stockpiling measure in anticipation of future export controls by the Dutch government.

Low production rate also has an impact on yield, since manufacturing technicians and engineers tend to improve their techniques based on experimentation and observed results from production. Thus, more production is helpful in accelerating the pace of learning and thus improving the yield.

Based on the Ascend 910B die size of 665.61 mm^2, CSIS estimates that each 300-millimeter diameter silicon wafer produces roughly 80 chip dies. This is simply counting the number of small rectangles (chip dies) that can fit inside the large circle (the wafer). A source told CSIS that of these 80 chip dies:

- Roughly 20 percent (~16) are fully functional with either no defects or no defects that negatively impact performance. This is the same as saying that SMIC has 20 percent yield when making these chips.
- An unknown share of chips are functional with degraded performance, though customer demand for such degraded chips (in Huawei's case, though not Nvidia's) is low.
- An unknown share of chips are completely nonfunctional.

SMIC's 20 percent yield when producing Ascend 910Bs is quite low, but it is not as low as it might seem. In 2020, when Nvidia moved its GPU production to TSMC's 7 nm process, the defect-free yield rate was **roughly 41.5 percent**. Since manufacturing defects occur randomly on a defects per square centimeter of chip area basis, larger chips (such as AI chips) are more susceptible than smaller chips (such as smartphone application processors) to a high defect rate. Using TSMC during roughly the same time period and with roughly the same defect rate, Apple's A12 processor enjoyed 90 percent+ yield. SMIC also makes the application processor for Huawei's Mate smartphone line, and its yield when making these smaller chips is somewhere in the 50–70 percent range.

However, a February 24, 2025 **report** by the *Financial Times*, citing two anonymous individuals, claims that SMIC's AI chip yield has increased to roughly 40 percent. Industry sources told CSIS that the *Financial Times* report is not correct and that Huawei/SMIC's true yield remains at 20 percent. One explanation for the discrepancy would be if the *Financial Times*' sources were mistakenly including both the fully functional chips and the chips that are functional but with degraded performance when estimating yield.

17. Sources told CSIS that Huawei, SMIC's most important customer, has successfully managed to move stockpiled U.S.-built equipment from SiEn (芯恩), Pensun (鹏新旭), and Huawei's fab in Dongguan to SMIC SN2.

As mentioned above, SMIC's shortage of deposition, etching, inspection, and metrology semiconductor manufacturing equipment also been an important factor preventing expanded production capacity of its SN2 facility. Some of this equipment is restricted on a country-wide basis, meaning that it cannot be legally sold anywhere in China. However, other types of this equipment were restricted only on an end-use and end-user basis. This means that the equipment can be sold to some customers in China but

not others. In some cases, this even means that it can be sold to some facilities of a particular customer, but not others. In such cases, relocating the equipment from one facility to another would require a new export license in order to be legal. But SMIC's production of 7 nm chips using U.S. equipment is already **illegal**, and both SMIC and other Chinese firms always have the option to choose illegal activity, particularly since such activity frequently enjoys the active support of the Chinese government.

Industry sources told CSIS that SiEn, Pensun, and Huawei's fab in Dongguan all were able to legally acquire the needed etching, deposition, and inspection/metrology equipment that SMIC needs for two reasons: (1) the equipment was not restricted on a country-wide basis to all of China and (2) the equipment was restricted on an end-use and end-user basis, but SiEn and Pensun told U.S. firms that it would exclusively be used for producing chips less advanced than 14 nm. These firms also denied any affiliation with Huawei. Government officials told CSIS that in such circumstances, the equipment can often be sold under a no-license required status.

According to the sources, SiEn and Pensun, however, did not have sufficient customers providing demand for using all of the equipment they had purchased, and so some of it was never used operationally in their fabs. They purchased the equipment as a stockpiling move in anticipation of future export controls. The source described this as a "buy everything you can, while you can" strategy.

The SMIC SN2 facility needed the equipment. Since SiEn and Pensun were not making economically productive use of the equipment, they were amenable to a sale. This sale was negotiated in Q4 of 2024 and completed in Q1 of 2025. The sources are under the impression that all of the desired equipment is currently either installed at SMIC SN2 or on-site awaiting installation.

18. As a result of the successful in-country equipment transfer, SMIC expects to achieve 50,000 7 nm WPM by the end of 2025. If all of this capacity was devoted to manufacturing Ascend AI chips (which is unlikely), that would imply the production of millions of Ascend 910C chips annually.

This equipment will eliminate the near-term U.S. equipment bottleneck facing SMIC for SN2 7 nm production. SMIC's new 7 nm production bottleneck is the overall size of the facility. Prior public statements from SMIC indicated that the SN1 and SN2 facilities combined have a maximum capacity of 85,000 WPM and that SMIC intended to split production evenly between the two facilities, implying SMIC was targeting ~42,500 WPM of lithography capacity at SN2. Sources told CSIS that SMIC was targeting 50,000 WPM of 7 nm specifically by the end of 2025. It is unclear whether this means that 7 nm production will now also take place at SN1 or whether SMIC has figured out a production process and equipment configuration that allows them to squeeze more production capacity than expected into SN2. If, hypothetically, SMIC were to devote all 50,000 WPM to Ascend 910C manufacturing, this would imply producing 4 million Ascend 910B dies per month, of which 800,000 would be fully functional. 800,000 910B Ascend dies is enough to manufacture 400,000 910C chips, though the advanced packaging process would reduce yield further, and it is unclear how much HBM Huawei has on hand.

SMIC is unlikely to devote all of its 7 nm capacity to Ascend chips. Huawei needs that 7 nm capacity for its chips for smartphones, laptops, data centers, and telecommunications equipment. Moreover, SMIC has other customers besides Huawei. Still, the point remains that Huawei is likely poised to dramatically expand Ascend production in the near future.

Over time, SMIC's 7 nm yield will likely improve significantly above 20 percent due to the additional equipment and the increased production rate. It will likely not match the best TSMC 7 nm yields for AI chips, however, since TSMC's best 7 nm yields used EUV, and China has no credible near-term path for acquiring EUV technology.

Over the mid and long term, however, Huawei is working aggressively to produce a domestic alternative to ASML EUV technology. Huawei currently has two semiconductor manufacturing equipment tool research facilities, one in Shanghai and one in Shenzhen. According to a **report** by Nikkei Asia, Huawei is investing $1.66 billion in the Shanghai facility alone and has hired a large number of chip industry veterans with experience working at companies like Applied Materials, Lam Research, KLA, ASML, TSMC, Intel, and Micron.

Taylor Ogan, a venture capitalist based in Shenzhen, China, focused on investing in Chinese technology companies, **described** Huawei's EUV efforts as follows:

> There are teams at Huawei in two different parts of China that are working around the clock. . . . They are literally cut off from their friends and family just because the work they are doing is so sensitive, and imagine how quickly they're going to accelerate Chinese complete domestic chip development. . . . One of them is in Shenzhen. That's the bigger [EUV lithography] breakthrough lab, and you can look on Google Earth. There was nothing there and now there is something there, and there are thousands of people working there. The other one is around Shanghai and that is more for the next chips in the next phones. So, yeah, the bigger lithography breakthroughs are [being worked on] in Shenzhen. . . . They're working their asses off. . . . This is like some of the smartest people in the space, and they're all Chinese.

As mentioned above, these secretive efforts underway at Huawei are extremely likely to be benefitting from state-backed industrial espionage, including the **cyberattacks on ASML** and its **partners**.

19. Huawei's Ascend chips continue to face challenges in terms of a lack of compatible AI software that is driving low utilization of purchased chips. However, this could change if DeepSeek's open-source community enthusiasm improves Huawei's CANN software ecosystem competitiveness with Nvidia's CUDA.

A high-quality recent **analysis** by Nicholas Welch, Lily Ottinger, and Jordan Schneider of *ChinaTalk* concludes that China's current compute shortage has more to do with practical challenges and suboptimal deployment—many idle small and medium data centers rather than fewer fully-utilized superclusters—than it does with an absolute shortage of chips. A combination of recent Chinese government policy changes, **such as** "hand[ing] over their idle computing resources to cloud providers" and strategic moves by China's largest tech companies to favor DeepSeek models even over their in-house ones, suggests that this could change in the near future.

According to **reporting** by *Caijing Magazine*, the chips in China most likely to be underutilized are Huawei's Ascends, which have benefitted in terms of government pressure to buy domestic but remain underutilized because of a lack of useful AI models that are Ascend-compatible. This could change if DeepSeek adapts its models to run on Ascends, which it is reportedly **working toward** for model inference, though not model training.

20. Between the chip dies acquired from TSMC, the stockpiled HBM, and SMIC's increasing Ascend production, Huawei and DeepSeek have a credible path to a million-Ascend-chip AI supercluster should they seek to build one, though they will face challenges in large scale chip integration and software frameworks.

As mentioned above, sources told CSIS that TSMC has supplied Huawei with enough chip dies to manufacture a million Ascend 910C chips, each of which has roughly 60 percent the performance of an Nvidia H100 when used for inference. With the help of Huawei and the Chinese government, DeepSeek could plausibly succeed in driving a reallocation of the Nvidia H800s and H100s already in China toward itself for construction of a new AI model training data center. This could then be augmented by an additional, much larger cluster of Huawei Ascend 910Cs (or Nvidia H20s) to be used for AI model inference. Such a combination is a plausible, perhaps even likely, path for China to continue making significant progress in its AI development.

As discussed above, a million-chip cluster of Huawei Ascends would face the challenge of a shortage of AI models, software, and developer tools that are compatible with Huawei's CANN software ecosystem. A second challenge is in integrating all of the chips as part of a single cluster where the whole is more than the sum of the parts. *Financial Times* reporter Eleanor Olcott interviewed one Beijing-based chip investor who **said** "The bottleneck at the moment isn't getting the chips but figuring out how to make them work in a cluster. This is really complicated work." Evidently the true attractiveness of the Huawei Ascend 910C chips is worse than the "60 percent of H100" figure might suggest.

21. U.S. firms are still ahead of China in the race toward human-level artificial general intelligence and beyond human-level artificial superintelligence. However, the gap has narrowed significantly, and it is unrealistic to expect a lead of more than a year or two, even with extremely aggressive export controls.

As discussed above, in the absence of export controls, it is plausible that Chinese firms would have already surpassed their U.S. competitors in developing and deploying frontier AI models. It is likewise plausible that China might have already succeeded in driving widespread adoption of the Huawei Ascend chips, just as Huawei's smartphone division, bolstered by internal chip designs, was looking poised to overtake Apple's iPhone in China prior to the introduction of effective export controls in 2020. The export controls have done much to slow and complicate China's technological rise and march toward their longstanding objective of technological self-sufficiency. They are still having a significant effect even now and are poised to have even greater impact in the future as U.S. firms benefit from Nvidia's next-generation Blackwell chips and as China likely remains stuck at the 7 nm technology node until and unless they develop a domestic EUV lithography capability.

However, China's successes in substantial government investment, chip smuggling, exploiting gaps in U.S. export control coverage, completing in-country transfers of equipment, hiring talent with experience from leading international firms, reverse-engineering foreign technology, harnessing state-backed economic espionage, and producing genuine domestic innovation are a formidable combination. U.S. firms, working at maximum effort and with the benefit of all their own advantages, will still likely be the first to develop artificial general intelligence or artificial superintelligence. This should not, however, be taken for granted.

At a meeting on February 17 between CCP Chairman Xi Jinping and Chinese technology executives, including DeepSeek CEO Liang Wenfeng, Huawei founder Ren Zhengfei **told** Xi that his **previous concerns** about the lack of domestic advanced semiconductor production and the damaging impacts of U.S. export controls had eased because of recent breakthroughs by Huawei and its partners. It is possible that he is referring to the aforementioned domestic equipment transfer to SMIC's facility. Ren further **said** that he is leading a network of more than 2,000 Chinese companies that are collectively working to ensure that China achieves self-sufficiency of more than 70 percent across the entire semiconductor value chain by 2028. These predictions should be taken seriously and have enormous strategic implications beyond just the AI sector.

However, for AI, the key question remains the one that Dario Amodei, CEO of Anthropic, has succinctly **posed**: "whether China will also be able to get millions of chips."

China's success to date suggests that, at least for Huawei Ascend chips, the answer is that they will have millions of chips within the next year or two. Thankfully, these chips are, at present, dramatically lower performing than Nvidia ones for training advanced AI models; they are also supported by a much weaker software ecosystem with many complex issues that will likely take years to sort out. This is the time that the export controls have bought for the United States to win the **race to AGI** and then use that victory to try and build more durable strategic advantages. At this point, all the margin for sloppy implementation of export controls or tolerance of large-scale chip smuggling has already been consumed. There is no more time to waste. ■

Gregory C. Allen *is the director of the Wadhwani AI Center at the Center for Strategic and International Studies in Washington, D.C.*

This report is made possible through general support to CSIS. No direct sponsorship contributed to this report.

The author would like to thank Georgia Adamson, Joshua Turner, and Ryan Featherston for their research support. The author would also like to thank Richard Danzig, Lennart Heim, William Reinsch, Scott Kennedy, Navin Girishankar, and others who wish to remain anonymous for their helpful feedback on earlier drafts of this report.