OXFORD SEMICONDUCTOR CONFERENCE 2025
POST-CONFERENCE COMPENDIUM

SECURING THE FUTURE OF TRUSTED SEMICONDUCTOR SUPPLY CHAINS

April 2026

EDITORS
Oxford Semiconductor Conference Committee

CONTRIBUTORS
Jonah Allen
Sam Bresnick
Hanna Dohmen
Ian Lange
Tsaiying Lu
Jeffrey Payne
Shane Sethi
Cartus Bo-Xiang You

OXFORD SEMICONDUCTOR CONFERENCE 2025

POST-CONFERENCE COMPENDIUM

SECURING THE FUTURE OF TRUSTED SEMICONDUCTOR SUPPLY CHAINS

APRIL 2026

EDITORS

Oxford Semiconductor Conference Committee

CONTRIBUTORS

Jonah Allen

Sam Bresnick

Hanna Dohmen

Ian Lange

Tsaiying Lu

Jeffrey Payne

Shane Sethi

Cartus Bo-Xiang You

ABOUT OSC

The Oxford Semiconductor Conference (OSC) is an independent forum that unites global technology, policy, and investment leaders to sustain partner and allied alignment on critical technologies amid growing geopolitical volatility.

Established in 2024, OSC convenes an annual closed-door, invite-only gathering of senior leaders from government, industry, finance, and academia to examine the strategic challenges shaping the future of the semiconductor ecosystem.

By maintaining lines of communication across these communities, the conference aims to strengthen policy coordination, surface emerging risks, and support practical collaboration on issues ranging from supply chain resilience and industrial strategy to national security and long-term competitiveness.

The 2025 conference brought together participants from across North America, Europe, and Asia for closed-door discussions on the opportunities and constraints facing the global semiconductor sector. This compendium responds to those discussions and presents a set of essays on the most pressing issues shaping the international debate.

© 2026 Oxford Semiconductor Conference. All rights reserved.

Oxford Semiconductor Conference
1200 Pennsylvania Ave NW STE 18 Washington 20004 District of Columbia United States
Oxford Semiconductor Conference, Inc is a registered 501(c)(3)
(EIN: 33-2328509), incorporated in Washington, DC
admin@oxfordsemi.org | oxfordsemi.org

OXFORD
SEMICONDUCTOR CONFERENCE

OSC Editorial Committee

Editorial Lead: Aidan Powers-Riggs
Editorial Co-Lead: Tommy Hall
Post-Production and Design Lead: Feifei Hung
Copywriting Lead: Rory Kind
Editorial Team: Margaret Siu, Jaebok Lee, Jacob Lapin

OSC Advisory Board

Cristina Brayton-Lewis
White & Case

R. Clarke Cooper
Atlantic Council

Melanie Garson
University of London

Martin Kuball
University of Bristol

Evan Medeiros
Georgetown University

Chris Miller
Tufts University

Sarah O'Hare O'Neal
Microsoft

Robert Trager
Oxford Martin School

W. Patrick Wilson
Founding Principal, Semiconductor and Innovation Group LLC (Former MediaTek Inc)

Wenchi Yu
Harvard Kennedy School

OXFORD
SEMICONDUCTOR CONFERENCE

ACKNOWLEDGEMENTS

The editors thank the authors, rapporteurs, speakers, and participants whose contributions made the 2025 Oxford Semiconductor Conference and this publication possible.

We are especially grateful to OSC's sponsors and institutional partners for their support, and to the OSC Board of Advisors, whose steadfast support remains pivotal to building OSC's credibility as a reliable forum.

The discussions informing the compendium were held under the Chatham House Rule. The views expressed in this compendium are those of the individual authors and do not necessarily reflect the views of the Oxford Semiconductor Conference, its organizers, sponsors, or participants.

This compendium was made possible by general support to OSC. No direct sponsorship contributed to this publication.

OXFORD
SEMICONDUCTOR CONFERENCE

TABLE OF CONTENTS

Introduction

The global order is undergoing a once-in-a-generation shift. The rules, norms, and institutions that once facilitated predictable international cooperation are giving way to a landscape defined by strategic competition and transactional diplomacy. Economic interdependence—once considered the cornerstone of global peace and prosperity—is now widely viewed with suspicion as a potential vector for coercion. The balance of military power is shifting in unpredictable ways, shaped as much by the depth and resilience of underlying technological and industrial ecosystems as by traditional force structures alone.

In this new environment, old ideas have returned to the fore. From Tokyo to Washington, concepts such as "security" and "sovereignty" are replacing "efficiency" and "integration" as the core tenets of economic policymaking. National security and economic statecraft are now inseparable: Between 2022 and 2025, major economies including the United States, the European Union, the United Kingdom, Japan, Germany, and South Korea all published their first major strategy documents focused around the concept of "economic security."[i] These shifts have coincided with a widespread revival of industrial policy, trade protectionism, and state intervention as economic tools of first resort, replacing decades-long policies and institutions promoting economic liberalism.

More questions than answers remain about the fundamental character of the emerging global order. Yet, it is clear that technology competition is a defining feature. Technological revolutions in artificial intelligence (AI), clean energy, computing, biotech, robotics and space are actively reshaping the economic and social fabric of society. More than ever, the ability to access and control these critical technologies is crucial to a country's ability to provide security, stability, and opportunity to its citizens. But rising global tensions and fracturing supply chains are creating new constraints for governments and firms that markets alone cannot resolve.

i This refers to either standalone strategy documents or high level national security strategy documents with major sections explicitly dedicated to the concept of "economic security."

The semiconductor industry sits at the center of this story. It represents a unique convergence of extreme engineering complexity, massive capital requirements, narrow supply chokepoints, and acute geographic concentration. Because semiconductors are the essential building blocks for everything from basic digital infrastructure to frontier AI, they are at the heart of the unfolding technology competition. And with perhaps the world's most complex and globally distributed supply chain, the industry complicates the rising impulse among nations to retreat behind the walls of "technological sovereignty."

For decades, the semiconductor industry was relatively insulated from geopolitical pressures. But the return of geopolitics now poses vexing challenges for policymakers and industry leaders at every layer of the semiconductor value chain. Which dependencies are tolerable and which are unacceptable strategic liabilities? Who should bear the cost of creating resilience? How far can "de-risking" go without undermining innovation and growth? Where should the lines be drawn between protecting advantages and capturing new markets in an era of intensifying competition?

A New Chapter

These new uncertainties create an urgent need for active coordination between countries that share similar interests, goals, and values. The Oxford Semiconductor Conference (OSC) was created to provide a neutral forum for this coordination to occur, with an enduring mission to maintain lines of communication between senior leaders from government, industry, finance, and academia. In September 2025, over 100 thought leaders and senior practitioners from more than a dozen countries gathered at Oxford University to exchange ideas and perspectives on the emerging challenges and opportunities facing the technology alliance. Our goal was to create a unique opportunity for leaders to engage in candid dialogue with counterparts beyond national, political, and professional silos, reminding all of the value of cooperation in the face of major headwinds.

This compendium aims to draw on the collective expertise of OSC participants and affiliates to introduce fundamental challenges facing the semiconductor industry and frame how to resolve them. Each chapter aims to provide context and distill actionable, novel ideas for well-positioned stakeholders to act upon. In the pages that follow, contributing experts from leading international institutions including the Payne Institute for Public Policy, Jain Family Institute, Research Institute for Democracy, Society and Emerging Technology (DSET), Center for Strategic and Emerging Technology (CSET), and the United States National Defense University explore the complexities of technological competition across four key pillars of the semiconductor value chain: 1) critical minerals, 2) energy, 3) sovereignty, and 4) national security.

CRITICAL MINERALS

In the first chapter, Ian Lange, Jonah Allen, and Shane Sethi address a foundational physical layer of the industry: the recovery and refining of critical minerals. Some segments of the semiconductor ecosystem depend on precarious, "thin" markets for materials like gallium and germanium, which are essential for high-frequency integrated circuits, defense optics, and advanced solar technologies. China has established a near-monopoly over these minerals, controlling approximately 99% of global primary gallium and 60% of germanium production. This concentration creates a high-sensitivity chokepoint that leaves large segments of the allied technology stack vulnerable to export restrictions.

The analysis argues that because these minerals are recovered as byproducts of processing other materials, the solution is not found in broad market subsidies. Instead, the chapter proposes a strategy of "precision, not scale." This involves targeted G7 interventions at specific existing refineries to enhance their recovery capacity, from South Korean Zinc plants in Ulsan to bauxite refineries in Gramercy, Louisiana. By providing firm-level technical investment and guaranteed demand, allies can materially reduce dependency on Chinese supply chains and build a resilient mid-stream refining capacity without introducing broad distortions into the global commodities market.

Energy

The second chapter, by Cartus Bo-Xiang You and Tsaiying Lu examines the staggering energy demands of the AI buildout, positioning power availability as the primary physical bottleneck of the digital era. As generative AI becomes a standard interface for digital life, its insatiable energy appetite has emerged as a primary infrastructure bottleneck. A single frontier model query can consume roughly 60,000 to over 100,000 times the electricity of a traditional web search. This exponential increase in power consumption means that the rapid expansion of AI will entail a high-stakes competition for grid capacity and stable baseload power.

Their contribution explores how to reconcile this rapid digital innovation with the physical limits of aging and often overstretched electrical infrastructure, juxtaposing experiences of large scale economies such as China and the United States with geographically constrained states: Ireland, Singapore, and Taiwan. They argue that the geography of AI data centers must be more intentionally aligned with regional development and the green transition.

Rather than viewing data centers solely as sources of grid stress, the authors suggest they can be used as levers for accelerating energy modernization. By utilizing carefully calibrated locational incentives and grid modernization initiatives, governments can encourage the siting of data centers in areas where they can anchor new clean generation and flexible demand, reinforcing both the AI ecosystem and sustainable development goals. Though no policy solution alone acts as a panacea, on-site power generation, colocation with domestic industrial and research clusters, and restructured planning standards are critical focus areas.

SOVEREIGN AI

Transitioning from infrastructure to the strategic application of these technologies, the third chapter sees Hanna Dohmen and Sam Bresnick investigate the strategic dilemmas of "sovereign AI" by deconstructing the technology stack into its constituent layers: 1) compute infrastructure, 2) data, and 3) models. While many nations seek total strategic autonomy, the extreme capital requirements of high-end chips and the immense data needs of frontier models make full independence technologically and economically infeasible for most.

Instead, they identify "hybrid sovereignty" as an emerging global model. This paradigm prioritizes the resilience to make independent political and strategic choices—often by leveraging open-source models to provide alternatives to state-aligned ecosystems like China's—rather than chasing the impossible goal of absolute self-sufficiency. For most countries, sovereignty will be defined

not by the ability to control the entire stack, but by the ability to reduce single-country dependencies at the hardware layer while preserving local governance over data and deployment.

AI AND NATIONAL SECURITY

The compendium concludes with Jeffrey Payne's assessment of the institutional frameworks necessary to sustain technological leadership through public-private partnerships. The era of "moving fast and breaking things" is giving way to a period where the scale of required technological advancement necessitates the state as a central facilitator. This chapter argues that traditional innovation pathways (which witnessed the public sector fall behind commercial giants) must be reformed to re-emphasize government's role in innovation cycles.

To win the long-term competition, national security institutions must accept greater risk and move away from laborious acquisition processes to facilitate both large and small innovators. Conversely, private firms must acknowledge that the infrastructure required for the next wave of frontier innovation—from massive compute clusters to secure supply chains—now requires the sustained strategic support and alignment only states can provide.

Looking Ahead

The chapters in this compendium converge on a clear takeaway: the next phase of technology competition will be decided less by isolated national initiatives than by whether partners can build durable cooperation mechanisms to solve difficult problems together. Many constraints explored here are not challenges any single government or firm can fix in isolation. Future convenings therefore must focus on practical opportunities for alignment, such as clearer divisions of labor, repeatable models for joint investment, and shared approaches to resilience that preserve innovation and market dynamism.

China's accelerating push to expand its capabilities across the technology stack brings the need for such an effort into sharper focus. Building collective capacity requires sustained attention to the enabling foundations of competitiveness—capacity in critical segments of the supply chain, the people and skills that power them, the energy infrastructure that underwrites the digital economy, and the financing structures that determine what can be built and how quickly. Shaping these shared priorities into durable and meaningful action that can endure beyond news cycles, politics, or market swings is the key to unlocking the future of our technological alliance.

Chapter 1: Strategic Intervention to Rebuild Semiconductor Minerals Capacity in the West

By Shane Sethi,[i] Jonah Allen[ii] and Ian Lange[iii]

KEY TAKEAWAYS:

- Recent efforts to address supply chain vulnerabilities and rebuild domestic manufacturing depend on a stable and secure supply of critical components; China currently dominates the production and refining of critical upstream inputs. Two critical minerals in particular, germanium and gallium, have been targets of export restrictions and are the most supply-sensitive inputs for compound semiconductors, used in high-frequency integrated circuits, defense optics, and solar technologies.

- Both minerals are recovered as byproducts and represent "thin" global markets. Potential G7 recovery capacity is significant compared to demand, but interventions are needed to support production, which will face a steep premium versus production in China.

- Governments should adopt precision, not scale, focusing on firm-level partnerships and technical investment support that enhance recovery capacity while minimizing new distortions. Coordinated G7 action targeting specific alumina and zinc refineries could materially reduce dependency on China and strengthen semiconductor supply chain resilience.

- Support mechanisms should focus on targeted interventions such as capex grants, concessional loans, offtake agreements, and price-stabilization mechanisms that are better suited to address investment and demand uncertainty.

i Shane Sethi is a graduate of the Mineral and Energy Economics program at the Colorado School of Mines.

ii Dr. Jonah Allen is a Vice President & Lead Researcher for Minerals for Development at the Jain Family Institute.

iii Dr. Ian Lange is the Viola Vestal Coulter Chair of Mineral Economics at the Colorado School of Mines. Additionally, Ian serves as Chair of the U.S. Commodity Futures Trading Commission's (CFTC) Role of Metals Markets in Transitional Energy Subcommittee.

BACKGROUND: THE GLOBAL SEMICONDUCTOR MARKET

Semiconductors are like the nerve cells of modern technology; when layered into integrated circuits, they function like densely interconnected neural networks that power everything from smartphones to supercomputers. There are several types of semiconductors – each designed for specific purposes – and for advanced use applications, they are usually layered as integrated circuits (ICs) (see Table 1).

Table 1: Types of Semiconductors and Their Applications

Semiconductor Type	Description and Use
Logic semiconductors	Perform computational and decision-making functions in electronic systems. They include CPUs, GPUs, and microcontrollers, which process data and control other components. Found in computers, smartphones, and embedded systems.
Memory semiconductors	Store digital data temporarily (volatile) or permanently (non-volatile). Examples include DRAM, SRAM, and NAND flash. Used in RAM modules, SSDs, and data centers.
Analog semiconductors	Handle continuous signals (voltage, current) and convert them between analog and digital forms. Examples are amplifiers, voltage regulators, and ADC/DAC converters. Essential in sensors, power management, and communication systems.
Micro semiconductors	Contain microprocessors and microcontrollers that integrate processing and control functions on a single chip. Used in computers, cars, and industrial automation to execute programmed instructions.
Optoelectronic semiconductors	Convert electrical signals to light (or vice versa). Examples include LEDs, laser diodes, and photodetectors. Used in fiber-optic communication, displays, and solar panels.
Discrete semiconductors	Individual components that perform a single electrical function such as rectification, amplification, or switching. Examples are diodes and transistors. Found in power supplies, amplifiers, and RF circuits.
Integrated Circuits	Combine millions of transistors, resistors, and capacitors on a single silicon chip to perform complex functions. Used in virtually all modern electronics — from smartphones and TVs to cars and industrial machinery.

Source: Jain Family Institute and Payne Institute for Public Policy • Created with Datawrapper

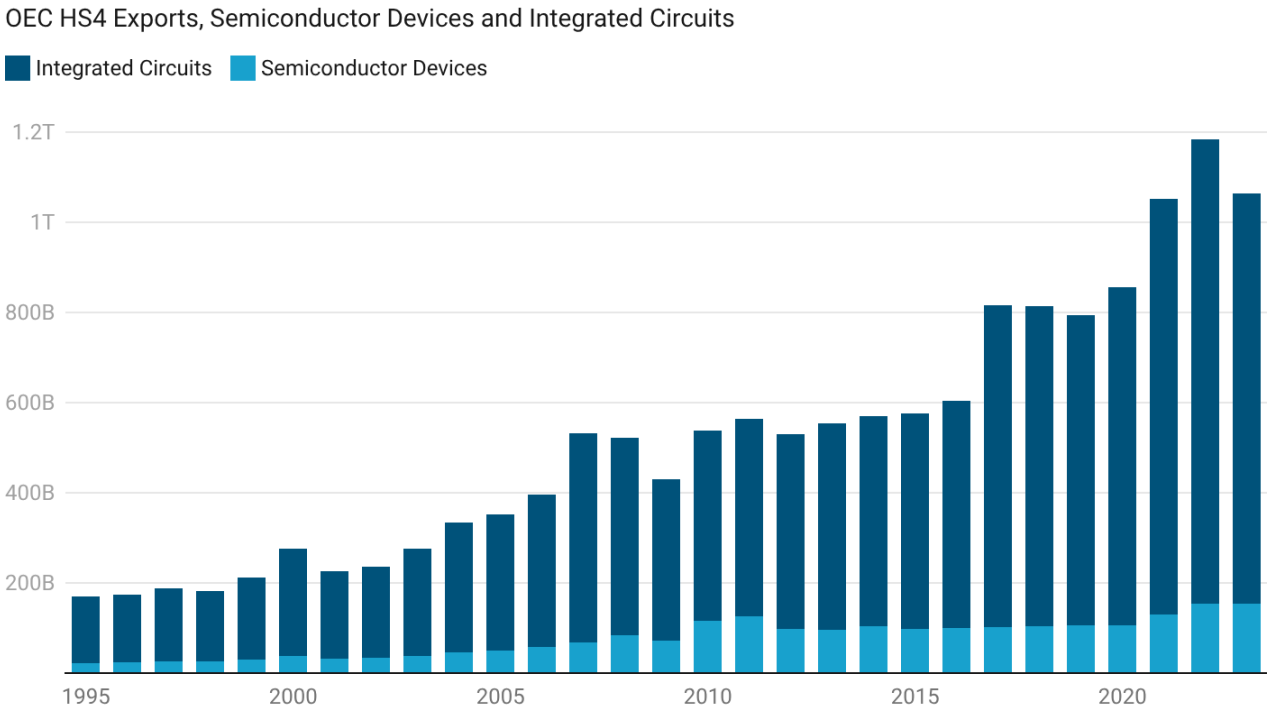
Unlike other semiconductors that can perform only one task at a time, an IC can execute multiple tasks with a high degree of complexity. Layering semiconductors in this way reduces the amount of external connections and components required, thus allowing for a more efficient product, reduced size and cost, and fewer external connections. The invention of flat-screen televisions was

enabled by the development of ICs, for example.

The physical structure of integrated circuits is much more complex than single-function semiconductors. An integrated circuit is built on a silicon wafer and contains millions of interconnected components, including transistors, resistors, and capacitors. ICs range from small-scale to ultra-large-scale integration, encompassing anywhere from about 10 transistors to more than 10 million transistors on a single "chip." Their fabrication involves a series of highly precise processes that transform purified silicon into densely patterned functional devices.

ICs now represent the lion's share of global trade in semiconductors, and trade value has grown significantly in parallel with the increasing "smartification" and connectivity of the physical world (see Figure 1).

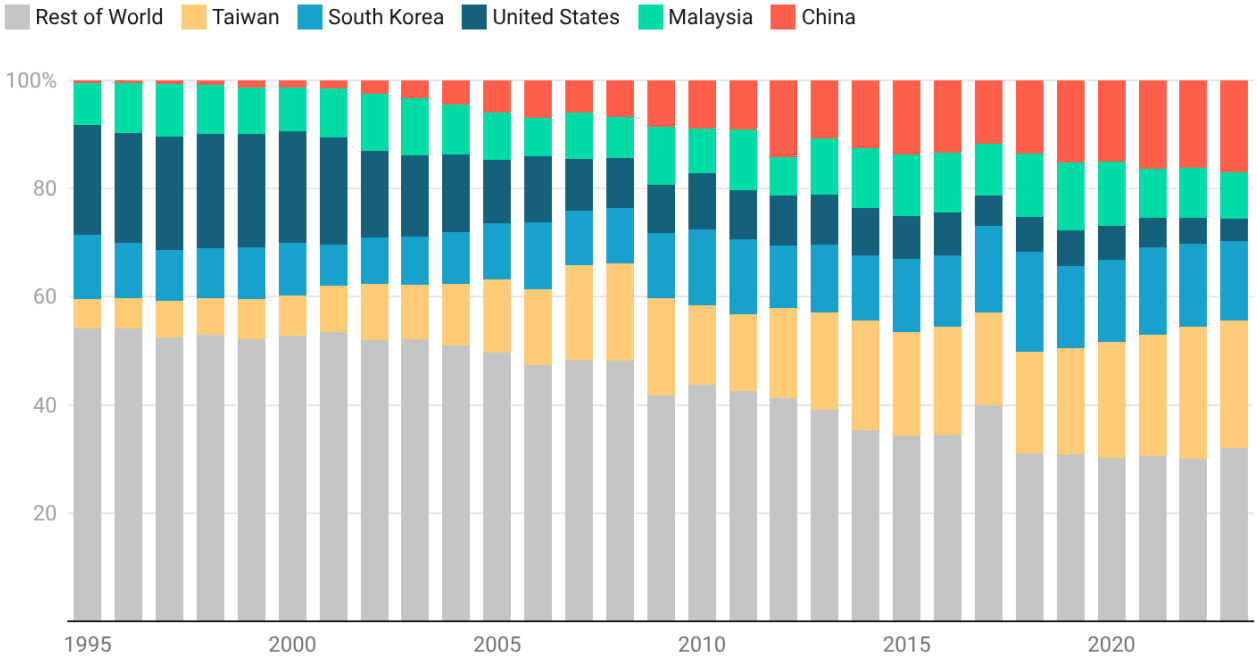
Figure 1: Global Trade in Semiconductor Devices and Integrated Circuits, 1995-2023



Source: Observatory of Economic Complexity, Jain Family Institute • Created with Datawrapper

As IC technology has advanced, packing exponentially more transistors onto smaller chips, the geographic center of manufacturing has shifted away from the United States and Japan (see Figure 2). While the U.S. continues to dominate in design, research, and equipment manufacturing, it now plays a relatively minor role in fabrication and export. The most advanced manufacturing capacity is concentrated in Taiwan and South Korea, where firms such as TSMC and Samsung have specialized in high-volume, high-yield production. This regional concentration reflects decades of cost optimization, ecosystem clustering, and strategic industrial policy that prioritized manufacturing precision and scale; the U.S. has increasingly outsourced production in favor of higher-value design and intellectual property segments of the supply chain.

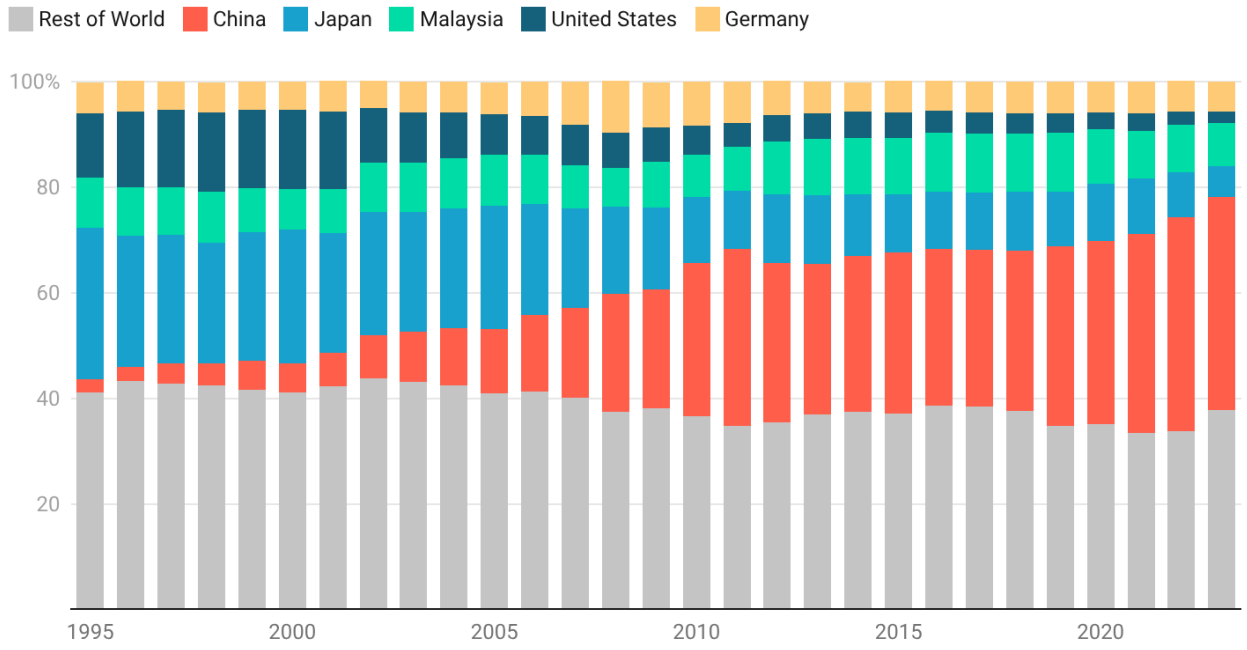
Figure 2: Share of Global Trade Value in Integrated Circuits, 1995-2023



Source: Observatory of Economic Complexity, Jain Family Institute • Created with Datawrapper

The U.S. role in the manufacturing and export of "semiconductor devices" is now even smaller than in ICs, due both to the geographic concentration of advanced manufacturing and China's industrial policies targeting solar manufacturing (see Figure 3). The trade category for "Semiconductor devices" includes not only discrete electronic components but also photovoltaic and photo-sensitive devices used in solar energy applications. This composition biases China's export share upward. Still, the broader trend highlights that the center of gravity for both conventional semiconductor production and newer semiconductor-related technologies has shifted decisively toward East Asia. At the same time, the U.S. has increasingly focused on upstream design and equipment segments of the value chain, making supply vulnerable to geopolitics and trade policies.

Figure 3: Share of Global Trade Value in Semiconductor Devices, 1995-2023



Source: Observatory of Economic Complexity, Jain Family Institute • Created with Datawrapper

Note: "Semiconductor devices" trade code captures discrete electronic components as well as photovoltaic and photo-sensitive devices used in solar energy applications. It does not include semiconductor devices that are layered into more complex "integrated circuits" (Figure 2), commonly referred to as "chips."

Due to these vulnerabilities, the United States is aiming to rebuild domestic manufacturing of semiconductors, especially advanced chips that are critical to defense and AI applications. This push is being operationalized through a mix of industrial policy and firm-level investments—most notably the CHIPS and Science Act of 2022, which provides the Department of Commerce $52.7 billion over five years (including $39 billion in manufacturing incentives) alongside a 25% Advanced Manufacturing Investment Credit to lower the cost of fab buildouts.[12] Major "reshoring" commitments are already underway: the U.S. Department of Commerce is advancing plans for large CHIPS awards supporting expansions by Intel (across multiple U.S. sites) and new U.S. capacity by TSMC and Samsung Electronics, while TSMC reports that high-volume production at its Arizona facility began in late 2024.[345] In 2025, CHIPS funding was deployed across logic, materials, and packaging segments, with direct awards to Hemlock Semiconductor and HPI Federal, alongside expanded support for advanced packaging and high-purity inputs to strengthen defense-relevant supply chains.

Rebuilding a strong manufacturing base for integrated circuits depends on a secure and stable supply of upstream inputs – especially critical minerals and the wafers fabricated from them. Every stage of semiconductor manufacturing, from wafer production to chip packaging, relies on a suite of small-market minerals such as gallium, germanium, high-purity quartz, and others. Yet China dominates the recovery, refining, and midstream processing of many of these commodities, creating a strategic vulnerability: while semiconductor fabrication itself is geographically concentrated in East Asia, much of the mineral and material foundation that feeds it is even more tightly controlled by China. Understanding which minerals are most essential to semiconductor manufactur-

ing, and which face the highest supply vulnerability, is critical to any effort to strengthen U.S. and allied semiconductor supply resilience.

MINERALS CRITICAL TO SEMICONDUCTORS

Of the several minerals used to manufacture semiconductors and integrated circuits (see Table 2), gallium and germanium are among the most supply-sensitive, reflecting both their concentrated production and the geopolitical tensions surrounding their trade.[iv] These elements are produced almost exclusively as byproducts of aluminum and zinc refining, with China controlling most global refining capacity (see Table 4). As a result, relatively small policy changes, such as targeted export restrictions, can create outsized disruptions to downstream industries that depend on high-purity gallium and germanium for advanced chips critical to defense technologies such as fiber-optic systems and infrared optics. Both minerals have been central targets of Chinese export controls since 2023.

iv While other byproduct metals such as indium share similar production pathways, indium has received less policy attention due to its smaller market size (approximately $1 billion) and more diversified upstream production base, including meaningful output outside China. Indium is therefore not treated as a focal supply-chain risk in this report. Also, in contrast to previous reports, this analysis does not focus on silicon, which, while foundational to semiconductor manufacturing, is not as supply constrained. The United States benefits from a stable domestic source of high-purity quartz from the Spruce Pine mine in North Carolina, supporting electronic-grade silicon production and mitigating exposure to concentrated foreign supply.

Table 2: Minerals Critical to Semiconductors

Mineral	Import Reliance	Key use in semiconductors
Arsenic	100%	Doping and compound base; donates electrons to silicon to enhance conductivity and combines with gallium to form GaAs (gallium arsenide).
Fluorspar	100%	Etching and cleaning; processed into hydrofluoric acid to etch patterns and clean silicon wafers, ensuring precise circuitry and defect-free chips.
Gallium	100%	Compound semiconductor base; combined with arsenide or nitrogen to form semiconductors that outperform silicon in speed, power handling, and optoelectronics.
Germanium	100%	High-speed and optical performance; improves electron mobility for faster, more responsive transistors and is used in infrared optics and high-efficiency solar cells.
Indium	100%	Conductive coatings; combined with tin oxide to form transparent, conductive layers that enable responsive touchscreens and high-quality displays.
Palladium	36%	Protective connections; resists corrosion in plating, bonding wires, and capacitors to ensure reliable chip performance over time.
Platinum	85%	Stable contacts; provides highly conductive, durable thin films and sensor components for consistent device performance.
Silicon	<50%	Semiconductor foundation; the conventional semiconductor material. Silicon controls the flow of electricity in chips, allowing information processing. It is "doped" with elements such as arsenic or boron to improve conductivity.
Tantalum	100%	Barriers and capacitors; prevents copper diffusion in wiring and supports stable, high-performance capacitors and thin-film resistors.

Source: U.S. Geological Survey, "Key Minerals in Data Centers Infographic", https://www.usgs.gov/media/images/key-minerals-data-centers-infographic • Created with Datawrapper

Gallium and germanium are primarily used in gallium arsenide (GaAs) and gallium nitride (GaN), which underpin applications in telecommunications, power electronics, infrared optics, fiber optics, and advanced sensors.

China produces approximately 99% of the global gallium supply and 60% of the global germanium supply; dominance has been fueled by government intervention.[6] Beijing mandated its rapidly expanding aluminum producers to install the capacity to extract gallium.[7] Between 2005 and 2015, China's production of low-purity gallium surged from 22 metric tons to 444 metric tons, a nearly 2000 percent increase. This move flooded the market and forced producers in the United Kingdom, Germany, Hungary, and Kazakhstan to shutter their operations.

China has used its dominance to its advantage in global policy negotiations. The most recent example came in 2023, when China restricted exports of gallium and germanium. The move by the Chinese Ministry of Commerce came just one day after the U.S. Bureau of Industry and Security amended the Export Administration Regulations by adding 140 Chinese entities to the Entity List.

GALLIUM ARSENIDE WAFER MANUFACTURE

Gallium arsenide (GaAs) wafers are manufactured through an intricate crystal growth and fabrication process requiring extreme purity and precision. Production is highly concentrated in Asia – particularly China – where most large-scale wafer fabrication and processing facilities are located, though some U.S. and European firms maintain design or R&D operations (see Table 3).

The single wafer manufacturer headquartered in the US, AXT Inc., maintains that all production is conducted in China. In Beijing, the company performs indium phosphide crystal growth and wafer processing. In Kazuo, it carries out gallium arsenide crystal growth, and in Dingxing, it processes both gallium arsenide and germanium wafers. AXT also owns a subsidiary, Tongmei Xtal Technology Co., Ltd., located in Beijing. In 2023, when China imposed restrictions on gallium and germanium exports, AXT had to navigate the necessary legal and commercial requirements to resume shipments of gallium arsenide and germanium substrates to certain customers. Additionally, AXT recently announced plans to list Tongmei on the Shanghai Stock Exchange.

Table 3: GaAs Wafer Manufacturers

Company	Country	Revenue (2023)
Sumitomo Electric Industries	Japan	$8.5bn
DOWA Electronics Materials	Japan	$500mm
Powerway Advanced Material	China	$200mm
Yunnan Germanium	China	$180mm
Wafer Technology	United Kingdom	$150mm
AXT Inc.	USA	$100mm
Freiberger Compound Materials GmbH	Germany	$120mm
China Crystal Technologies	China	$90mm
Atecom Technology Co. Ltd.	Taiwan	$80mm

Source: Authors' research and analysis • Created with Datawrapper

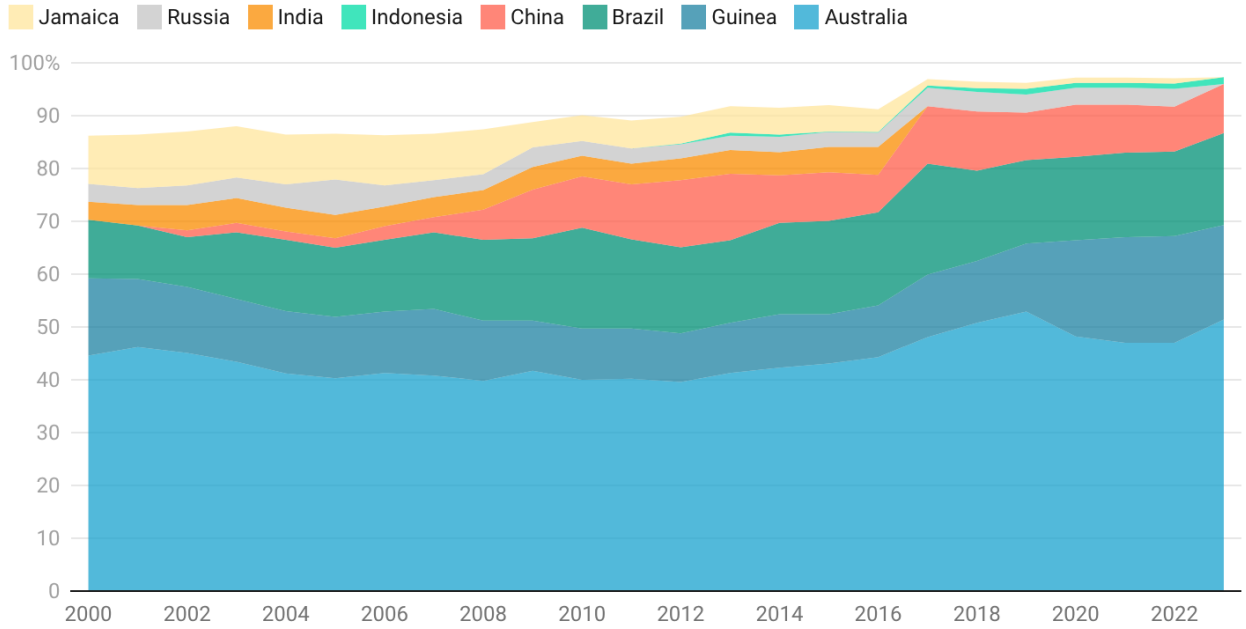
GALLIUM AND GERMANIUM EXTRACTION

Gallium is not mined directly but is recovered as a byproduct during the processing of bauxite and, to a lesser extent, zinc ores. In alumina refineries, gallium accumulates in the Bayer liquor, which is used to dissolve alumina from bauxite. Because the liquor is continuously recycled in a closed loop, small amounts of gallium that enter with each batch of bauxite build up over time until concentrations become high enough to extract economically. Recovery is typically achieved through chemical precipitation, solvent extraction, or electrolysis, separating gallium from the sodium aluminate solution. To be suitable for industrial application, the extracted gallium is then refined further into high-purity metal.

Bayer liquor concentrations typically range from 30 to 80 parts per million (ppm), though long-running circuits and bauxite ores particularly rich in gallium can reach 200–300 ppm. The original bauxite feed generally contains 10–80 ppm gallium, but some Chinese and Kazakh ores exceed 100 ppm. Once concentrations in the Bayer liquor surpass about 80–100 ppm, extraction becomes economical.

Bauxite is produced globally and several countries could enable future gallium production (Figure 4).

Figure 4: Share of Global Bauxite Production, 2000-2023



Source: Jain Family Institute's mineral market dashboard, utilizing data from S&P Global • Created with Datawrapper

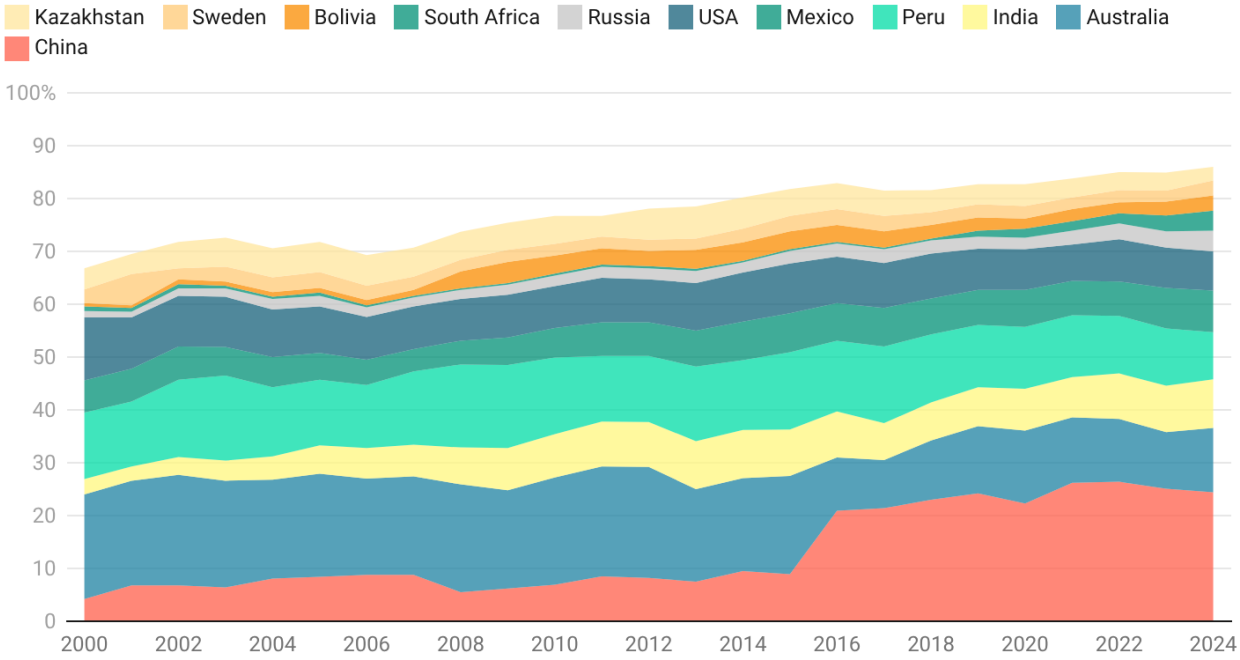
Germanium is recovered primarily as a byproduct of zinc and coal processing. In zinc smelting, germanium is concentrated in the flue dust that forms during the roasting of zinc ores. This flue dust is then leached with acid, and germanium is extracted through solvent extraction or ion exchange before being reduced to germanium dioxide (GeO_2), which is then further refined into metallic germanium. In coal operations, particularly from lignite (brown coal), germanium can be recovered from fly ash or flue gas condensates produced during combustion. Germanium also becomes concentrated in flue dust formed during the roasting of sphalerite in zinc smelting, where content can reach 0.1–0.5% Ge (1,000–5,000 ppm) from ores originally containing 50–150 ppm.

In coal systems, particularly lignite, germanium content averages 200–500 ppm, though some German and Chinese deposits exceed 1,000 ppm. After combustion, fly ash retains 100–500 ppm Ge, while flue gas condensates can contain up to 1,000 ppm. Recovery is typically economical when germanium levels exceed 200 ppm, with extraction achieved through acid leaching, solvent extraction, or ion exchange.

Although Gallium is mostly recovered as a by-product of refining aluminum from bauxite, it can also be extracted from the processing of Zinc ores or zinc-smelter residues. However, because gallium concentrations in zinc feedstocks and residues are extremely low and recovery requires complex, energy-intensive hydrometallurgical steps, the economics of zinc-based extraction are generally unfavorable, making it much less common than aluminum-based gallium recovery. The same is true for extracting germanium from bauxite/alumina feedstocks.

Zinc is produced globally and several countries could enable future germanium extraction (Figure 5).

Figure 5: Share of Global Zinc Production, 2000-2024



Source: Jain Family Institute's mineral market dashboard, utilizing data from S&P Global • Created with Datawrapper

GALLIUM AND GERMANIUM SOURCING POTENTIAL

Specialized minerals markets such as gallium and germanium are incredibly thin. Further, the United States does not usually import these minerals in their primary form; rather consumption tends to be "embedded" in downstream products. In 2024, the US only imported 12 tons of gallium and 36 tons of germanium, but nearly 200 tons of gallium arsenide (GaAs) wafers, which are used to manufacture compound semiconductors for integrated circuits.

Markets are small even when considering embedded consumption. The USGS valued U.S. imports in 2024 at roughly $4 million for gallium metal and $140 million for gallium arsenide (GaAs) wafers, with no domestic production; germanium metal and germanium dioxide were estimated to be $50 million.[89] For comparison, the estimated value of iron ore production in the U.S. alone was valued at $5.5 billionin 2024 and the copper content of U.S. production was valued at $10 billion.[1011]

The largest producers of refined gallium are CHALCO and Zhuhai SEZ Fangyuan Inc; both companies operate bauxite/alumina refining facilities with gallium capture technology. The biggest germanium producers are Yunnan Chihong Zinc & Germanium Co., Ltd, China Germanium Co., Ltd and Yunnan Germanium (see Table 4)[v].

v Yunnan is likely extracting from germanium from coal and China Germanium is extracting from zinc smelters.

Table 4: Global Gallium and Germanium Production Capacity

Capacity measured in tons per year (tpy)

Country	Gallium Facilities	Gallium Capacity (tpy)	Germanium Facilities	Germanium Capacity (tpy)
China	11	630	Unknown	309
Russia	1	16	1	27
Canada	0	0	1	7

Source: Authors' research and analysis • Created with Datawrapper

Several promising projects could help the United States establish a more secure supply chain with a lower risk of disruption. For Gallium, in Germany, there is a planned production capacity involving the company Dadco Alumina. Its AOS Stade facility has historically produced gallium but suspended operations in 2016. In 2021, it was announced that the plant would be brought back online; however, production has not yet resumed. In Australia, Alcoa announced a Joint Development Agreement with Japan Australia Gallium Associates, a venture between Sojitz Corporation and Japan's JOGMEC. If all goes as planned, a final investment decision is expected by the end of 2025, with production slated to begin in 2026. The goal is to produce more than 55 tons of gallium per year by 2028.

In May 2025, Rio Tinto and its new partner, Indium Corporation, successfully extracted their first primary gallium as part of a joint research and development project. The ultimate goal is to produce 40 tons annually in Quebec. This initial step was completed at Indium Corporation's research and development facility located in Rome, New York (see Table 5).

Table 5: Announced Gallium Capacity Additions

Capacity measured in tons per year (TPY)

Country	Announced Capacity Additions (tpy)
Germany	60
Australia	55
Canada	40
Kazakhstan	15
India	10

Source: Authors' research and analysis • Created with Datawrapper

Additionally, there is a proposed funding plan by EXIM to finance Atalco Gramercy LLC, located in Gramercy, Louisiana, with $450 million to build a gallium expansion at its bauxite refinery. As of writing, there is no indication of the projected production volume.

For germanium, in August 2025, Korea Zinc, the world's largest zinc smelter, signed a Memorandum of Understanding with Lockheed Martin, the world's leading defense company, for the supply and procurement of the mineral and for cooperation in the critical minerals supply chain. Korea Zinc plans to invest approximately KRW 140 billion in its Onsan Smelter in Ulsan to establish a new germanium plant. Following trial operations in 2027, the company aims to begin production in the first half of 2028, with plans to produce high-purity germanium dioxide equivalent to approximately 10 tons of germanium metal.

In addition to announced gallium and germanium capacity growth, we estimate potential additional co-production capacity of gallium and germanium capacity based on bauxite/alumina and zinc refineries in operation. Table 6 below thus shows potential targets for strategic support within the G7 according to standard conversion factors.[vi]

Table 6: Potential Gallium and Germanium Co-Production Capacity in G7 Countries

Capacity measured in tons per year (tpy)

Member	Alumina Refining Capacity	Potential Gallium Output	Zinc Refining Capacity	Potential Germanium Output
Canada	1,500,000	39	600,000	45
France	0	0	172,000	13
Germany	1,000,000	26	165,000	13
Japan	0	0	452,000	34
United States	1,200,000	32	130,000	10
Total	**2,800,000**	**97**	**1,519,000**	**115**

Source: Authors' research and analysis • Created with Datawrapper

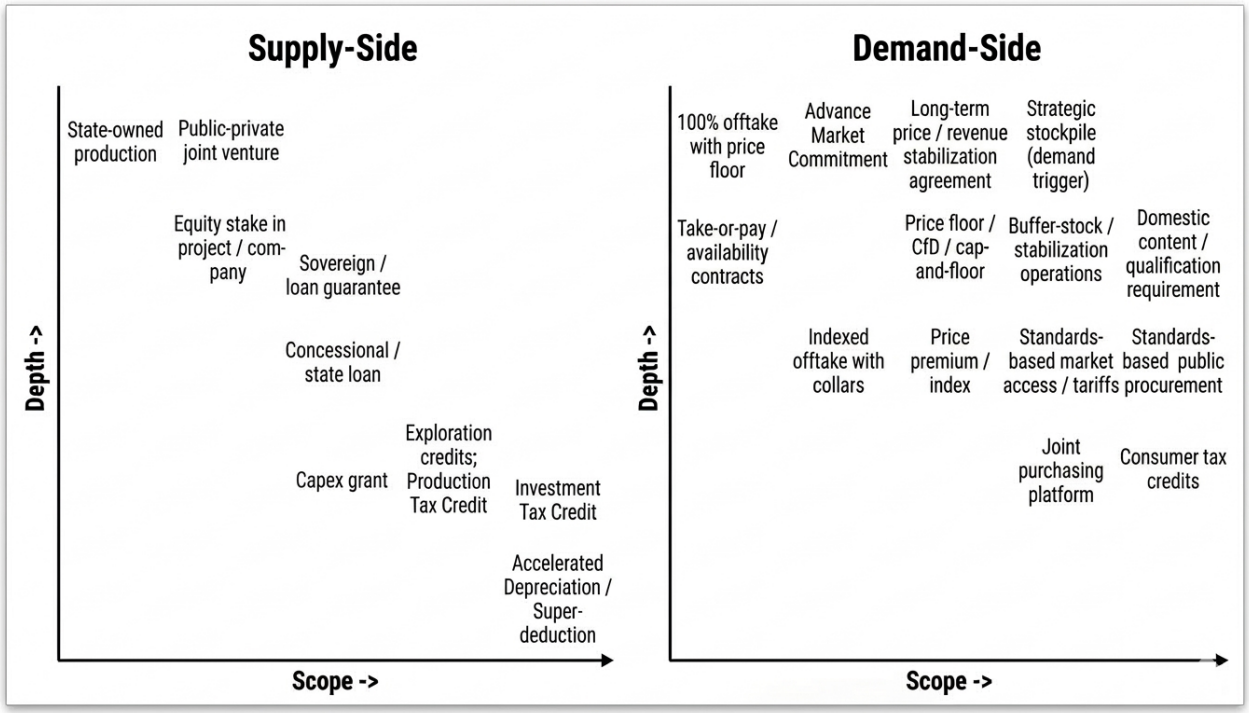
INTERVENING IN GALLIUM AND GERMANIUM MARKETS

Countries are increasingly willing to intervene in mineral markets to counter existing market distortions that challenge national and economic security or development objectives. Policymakers have a broad range of intervention strategies to choose from; at the same time, the geological, chemical, and market dynamics unique to each mineral necessitate that policymakers choose from this strategy set very selectively. Interventions can cause further market distortions and have varying levels of success, so the costs and benefits of strategies need to be weighed carefully.

vi Potential Ga captured (t/yr) = Alumina output (t/yr) × bauxite-per-alumina ratio (2.5) × Ga in bauxite (ppm) (50) × % Ga leached to Bayer liquor (70%) × Ga recovery efficiency (30%) × 10^6. Potential Ge captured (t/yr) = Zn metal (t/yr) × Ge ppm (120) × fume frac (90%).× recovery (70%) ×10^6

Mineral market interventions generally take two forms: supply-side measures that seek to enable more production, and demand-side measures that shape consumption or stabilize prices to create more predictable conditions for producers. The impact of a specific intervention strategy can vary in scope (the number of firms affected by the policy) and depth (government exposure to a firms or projects) (see Figure 6 and Appendix A).

Figure 6: Mineral Market Invention Strategies



Source: Jain Family Institute and Payne Institute for Public Policy

Strategies that have a broader scope or higher depth do not imply more effective interventions. For thin markets with high security concerns, such as gallium and germanium, it may be more effective to target a few promising producers with firm-or project-level support than a broader, more passive tax-based strategy. Further, while equity stakes are a "deeper" exposure as the government is essentially a strategic investor in the firm, firms may prefer a concessional loan, as debt can be harder to secure and can be more flexible than equity.

To that end, state participation in production through state-owned mining companies or joint ventures is very high in depth and should be considered with extreme caution, as the government is taking a significant risk. These strategies should also be less applicable in G7 country contexts, where significant industry and specialized knowledge exist, and policy objectives focus more on national and economic security than development goals.

For markets like gallium and germanium, the most effective interventions are those that stabilize prices, guarantee demand, and offset investment risk on a small scale. Thin byproducts markets will benefit from market intervention strategies that (i) offset the costs of capital expenditures necessary to recover and refine the commodity and (ii) provide demand-side certainty such that their investment and production do not crash market prices. For example, concessional loans and capex grants are effective in addressing constraints in capital investments. Meanwhile, direct government offtake, or more indirect support that stabilizes price for producers, such as a floor price or con-

tract-for-difference (CfD), is effective in providing security of demand. Further, for commodities with defense applications, pairing offtake programs with a strategic stockpiling program may be advantageous.

Importantly, passive strategies like tax incentives are not a good fit for gallium and germanium supply concerns because they rely on market responses that are unlikely to materialize in thin, price-insensitive markets where investment decisions hinge on guaranteed demand or direct cost offsets rather than incremental tax relief. However, if policymakers are still interested in pursuing tax relief strategies or including them in a more comprehensive approach, it is most advantageous to consider incentives that address the high cost of capital or the higher costs of domestic procurement – such as investment tax credits for refinery upgrades or content-based credits that reward the use of domestically sourced gallium and germanium in downstream manufacturing (see Appendix B).

Ultimately, addressing supply risks for gallium and germanium requires precision rather than scale. Because these are thin, byproduct-driven markets, broad incentives are unlikely to shift production or investment behavior on their own. When firm- or project-level support is required, governments should choose instruments carefully and ensure that intervention design draws on strong technical, financial, and market expertise. Targeted measures grounded in a sound understanding of refining processes, byproduct economics, and demand chains are far more likely to strengthen supply resilience without introducing new distortions.

Appendix A: Mineral Market Interventions Strategies and Their Fit for Gallium and Germanium Markets (Darker Green = Higher Fit)

	Intervention	Mechanism / Summary	Notes / Use case	Scope	Depth
Supply-side	State-owned production (SOE)	Government directly operates or mandates a national company to mine or refine minerals.	Ensures domestic production and supply security for strategic or thin markets; full state control.	Narrow	Very High
	Public–private joint venture (JV)	Shared ownership between the state and private investors.	Balances control with efficiency; suitable for strategic minerals and byproduct recovery.	Narrow	High
	Equity stake in project/company	State takes minority ownership stake; shares risk and governance.	Anchors financing, signals commitment, and provides oversight without full control.	Narrow	High
	Sovereign / loan guarantee	State guarantees private loans or offtake obligations.	Reduces risk for financiers; supports first-of-kind or high-risk projects.	Narrow–Med	Med–High
	Concessional / state loan	Below-market debt or blended finance for project development.	Improves project bankability and lowers cost of capital for early-stage investments.	Narrow	Med–High
	Capex grant	Non-repayable subsidy for construction or retrofit costs.	Fills viability gap for strategic or demonstration-scale projects.	Narrow	Med
	Exploration credits	Credit or rebate on qualified exploration spending.	Expands project pipeline and incentivizes discovery, especially for frontier or base-metal–linked deposits.	Medium	Low–Med
	Production Tax Credit (PTC)	Fiscal incentive linked to volume or value of output.	Encourages stable production levels and enhances competitiveness.	Broad	Low–Med
	Investment Tax Credit (ITC)	Percentage credit on qualified capital expenditures.	Stimulates entry and capacity expansion across the sector.	Broad	Low–Med
	Accelerated Depreciation / Super-deduction	Allows faster write-off of capital assets for tax purposes.	Improves cash flow and after-tax project returns for capital-intensive extraction.	Broad	Low
Demand-side	100 % offtake with price floor	Buyer guarantees volume and minimum price.	Secures revenue and de-risks new refineries or recovery facilities.	Narrow	High
	Take-or-pay / availability contracts	Buyer pays for capacity even if unused.	Smooths ramp-up and stabilizes revenue for early producers.	Narrow	Med–High
	Indexed offtake with collars	Price linked to index with floor / ceiling.	Balances producer and buyer exposure to volatility.	Narrow–Med	Med
	Advance Market Commitment (AMC)	Government / consortium pledges to purchase future output once it meets defined standards.	Creates demand for emerging refining / recycling technologies; reduces commercialization risk.	Narrow–Med	High
	Long-term price / revenue stabilization agreement	Multi-year bilateral contract guaranteeing fixed or banded prices.	Extends predictability beyond CfDs; suitable for thin, non-exchange-traded metals.	Medium	High
	Joint purchasing platform	Aggregates multiple buyers into joint tenders or procurement alliances.	Builds countervailing power and supports price transparency.	Broad	Med
	Standards-based market access / tariffs	Incentives / tariffs linked to ESG or strategic criteria.	Rewards compliant, traceable, or low-carbon minerals.	Broad	Med
	Domestic content / qualification requirement	Mandates minimum share of domestic / qualified minerals in manufacturing.	Converts voluntary incentives into binding demand for domestic supply.	Broad	Med–High
	Price premium / index	Price uplift for projects meeting priority criteria (e.g., low-carbon, domestic).	Emerging through exchanges and bilateral deals (e.g., LME, LSE).	Medium	Med
	Price floor / CfD / cap-and-floor	Guarantees a revenue band through public contract or treasury payments.	Stabilizes returns in volatile or thin markets.	Medium	High
	Consumer tax credits	Incentives for manufacturers using qualified minerals.	Common in EV / battery policies; complements supply-side tax credits.	Broad	Low–Med
	Standards-based public procurement	Weighted tenders favor certified / compliant mineral inputs.	Creates demand pull through government purchasing power.	Broad	Med
	Buffer-stock / stabilization operations	Public buy-sell operations to smooth price cycles.	Countercyclical tool for sensitive or mature markets.	Broad	Med–High
	Strategic stockpile (demand trigger)	Government purchases and holds minerals as a reserve.	Supports price floor and ensures availability; long-term demand anchor.	Broad	High

Source: Jain Family Institute and Payne Institute for Public Policy

Appendix B: Tax Incentives and Their Fit for Gallium and Germanium Markets (Darker Green = Higher Fit)

	Tax	Type	Basis / Structure	Objective	Example Use
Producer Tax Credits	Investment Tax Credit (ITC)	Cost-based	% of qualified capital expenditure	Stimulate entry / capacity expansion	Refining, processing plants
	Production Tax Credit (PTC)	Profit-based	$ per ton or % of output value	Encourage steady production	Domestic mining output
	Exploration Credit	Cost-based	% of exploration spending	Incentivize resource discovery	Frontier geology, juniors
	Accelerated Depreciation / Super-Deduction	Cost-based	Faster write-off of assets	Improve cash flow / IRR	Capex-heavy extraction
Consumer Tax Credit	Input / Use Credit	Cost-based	% credit on cost/value of eligible mineral input	Encourage use of domestic / green minerals	Wafer Manufacturing
	Content Credit	Profit-based	Credit for meeting minimum domestic or low-carbon content thresholds	Build demand pull & reshoring	Wafer manufacturing

Source: Jain Family Institute and Payne Institute for Public Policy

Appendix C: Facility-Level Data on Gallium and Germanium Production and Potential Capacity

Global Gallium Production Locations			
Country	**Company**	**Location**	**Capacity (t/y)**
China	CHALCO (Aluminum Corp. of China)	Pingguo (Baise), Guangxi	160
China	CHALCO	Jiaozuo (Zhongzhou), Henan	
China	CHALCO (Zunyi Alumina Co.)	Zunyi, Guizhou	
China	CHALCO	Hejin, Shanxi	
China	Beijing JiYa Semiconductor Material Co	Hejin, Shanxi	65
China	Shanxi Zhaofeng Gallium Industry Co. Ltd.	Yangquan, Shanxi	25
China	Zhuhai SEZ Fangyuan Inc	Zhuhai, Guangdong	140
China	East Hope Mianchi Gallium Industry	Mianchi (Sanmenxia), Henan	40
China	Shanxi Jiahua Tianhe Electronic Materials (molycorp JV)	Hejin, Shanxi	50
China	Xiaoyi Xingan Gallium Co.	Xiaoyi, Shanxi	50
China	Nanjing Jinmei Gallium Co., Ltd	Nanjing, Jiangsu	100
Russia	RUSAL	Achinsk, Krasnoyarsk Krai	16

Global Germanium Production Locations			
Country	**Company**	**Location**	**Capacity (t/y)**
China	Yunnan Chihong Zinc & Germanium	Huize	60
China	Yunnan Lincang Xinyuan Germanium	Lincang	48
China	Xilingol Tongli Germanium	Inner Mongolia	15
China	Yunnan Luoping Zinc & Electricity		8
Canada	Teck – Trail Operations	British Columbia	~40
Russia	JSC Germanium	Krasnoyarsk	~20
Russia	Germanium & Applications Ltd.	Moscow/Novomoskovsk	up to 21
DRC (Congo)	Lubumbashi Germanium Recovery Plant		30
Belgium	Umicore	Olen	n/a (undisclosed)

Global Gallium Planned Production Locations			
Country	**Company**	**Location**	**Capacity (t/y)**
Canada	Rio Tinto	Quebec	40
Germany	AOS Stade (Dadco)	Stade (Lower Saxony)	60
India	NALCO	Damanjodi (Koraput), Odisha	10
Australia	Alcoa of Australia (with Sojitz/JOGMEC via JAGA)	Pinjarra / Wagerup, Western Australia	55
Kazakhstan	ERG (Eurasian Resources Group)	Pavlodar	15

Potential Gallium and Germanium Production Capacity in G7 Countries					
	Commodity	**Company**	**Property Name**	**Location**	**Capacity (000t)**
Canada	Zinc	Glencore	CEZinc	Quebec	290
	Zinc	Teck Resources	Trail Smelter	British Columbia	310
	Bauxite/Alumina	Rio Tinto	Vaudreuil alumina refinery	Quebec	1,500
France	Zinc	Nyrstar	Auby smelter	Auby	172
Germany	Zinc	Glencore	Nordenham zinc refinery	Nordenham	165
	Bauxite/Alumina	Dadco Alumina	Aluminium Oxid Stade (AOS)	Lower Saxony	1,000
Italy	N/A				
Japan	Zinc	Toho Zinc	Annaka Smelter	Gunma	140
	Zinc	DOWA group	Akita Zinc	Akita	200
	Zinc	Hachinohe Smelting Co., Ltd	Hachinohe Smelting Zinc Refinery	Aomori	112
United Kingdom	N/A				
United States	Zinc	Nyrstar	Clarksville Smelter	Tennessee	130
	Bauxite/Alumina	Atalco	Atalco refinery	Louisiana	1,200

Source: Authors' research and analysis

Endnotes

1 U.S. Department of Commerce, National Institute of Standards and Technology (NIST), CHIPS for America Fact Sheet: Federal Programs Supporting the U.S. Semiconductor Supply Chain and Workforce, March 18, 2024

2 Internal Revenue Service, "Advanced Manufacturing Investment Credit," last modified November 26, 2025, https://www.irs.gov/credits-deductions/advanced-manufacturing-investment-credit.

3 National Institute of Standards and Technology, "Biden-Harris Administration Announces Preliminary Terms with Intel to Support Investment in U.S. Semiconductor Technology Leadership and Create Tens of Thousands of Jobs," news release, March 20, 2024, https://www.nist.gov/news-events/news/2024/03/biden-harris-administration-announces-preliminary-terms-intel-support.

4 National Institute of Standards and Technology, "Biden-Harris Administration Announces Preliminary Terms with TSMC, Expanded Investment from Company to Bring World's Most Advanced Leading-Edge Technology to the U.S.," news release, April 8, 2024, https://www.nist.gov/news-events/news/2024/04/biden-harris-administration-announces-preliminary-terms-tsmc-expanded.

5 Taiwan Semiconductor Manufacturing Company Limited, "TSMC Arizona - Taiwan Semiconductor Manufacturing Company Limited," accessed March 11, 2026, https://www.tsmc.com/static/abouttsmcaz/index.htm.

6 Matthew P. Funaiole, Brian Hart, and Aidan Powers-Riggs, "Mineral Monopoly: China's Control over Gallium Is a National Security Threat," Center for Strategic and International Studies, July 18, 2023, https://features.csis.org/hiddenreach/china-critical-mineral-gallium/.

7 Matthew P. Funaiole, Brian Hart, and Aidan Powers-Riggs, "Mineral Monopoly: China's Control over Gallium Is a National Security Threat," Center for Strategic and International Studies, July 18, 2023, https://features.csis.org/hiddenreach/china-critical-mineral-gallium/.

8 U.S. Geological Survey, "Gallium," in Mineral Commodity Summaries 2025 (Reston, VA: U.S. Geological Survey, 2025), 74–75.

9 U.S. Geological Survey, "Germanium," in Mineral Commodity Summaries 2025 (Reston, VA: U.S. Geological Survey, 2025), 80–81.

10 U.S. Geological Survey, "Iron Ore," in Mineral Commodity Summaries 2025 (Reston, VA: U.S. Geological Survey, 2025), 100–101.

11 U.S. Geological Survey, "Copper," in Mineral Commodity Summaries 2025 (Reston, VA: U.S. Geological Survey, 2025), 64–65.

Chapter 2: Revitalizing Energy Infrastructure for the AI Era: Mapping Challenges, Policies, and Strategies

By Cartus Bo-Xiang You[i] and Tsaiying Lu[ii]

KEY TAKEAWAYS:

- Outdated grid infrastructure, instead of power generation, is the binding constraint for AI development. Aging assets, overloaded interconnection queues, and transmission bottlenecks mean AI build-out now hinges on deep grid modernization rather than incremental capacity additions.
- The "100% renewable" claims from major AI datacenter operators often rely on specialized contracts that are non-additional, crowd out local buyers, and frequently fail to deliver meaningful system-level decarbonization.
- Large economies such as the US and China primarily lean on transmission expansion and workload relocation, whereas smaller, land-constrained hubs like Ireland, Singapore, and Taiwan quickly hit physical and political limits, resorting to regulatory responses.
- AI datacenters can function as valuable flexibility resources—through workload shifting, storage, participation in virtual power plants, and structured on-site power rules—only if policy and market design prioritize inclusivity, local benefit, and grid-integrated decarbonization over sheer hyperscale capacity.

ENERGY: THE CHOKEPOINT OF THE AI BOOM

Since the debut of ChatGPT in late 2022, the rapid expansion of artificial intelligence (AI) has drawn a surge of public attention not only to its capabilities but also to its immense energy appetite. Generative AI has quickly integrated into everyday digital life; for example, conventional search engines are increasingly embedded with or replaced by AI-driven interfaces.[1] This transformation, despite saving time for internet users through synthesized responses, conceals the vast

i Cartus Bo-Xiang You is the Deputy Director and Non-Resident Fellow on the Energy Security and Climate Resilience Program at the Research Institute for Democracy, Society, and Emerging Technology (DSET).
ii Dr. Tsaiying Lu is a Research Fellow and Director of the Energy Security and Climate Resilience Program at DSET.

and growing energy costs embedded in each apparently seamless interaction. The electricity required to generate a single AI response far exceeds that of traditional web queries, turning seemingly minor daily habits into catalysts of a growing global energy challenge.

Recent research underscores the magnitude of this shift. A single query to a model such as GPT-5 reportedly consumes on average 18 watt-hours and as much as 40 watt-hours at peak use—equivalent to roughly 60,000 to over 100,000 times the energy of a standard Google search.[2] Such figures illustrate how the computational intensity of large-scale AI models is redefining the relationship between information systems and energy systems. What was once a negligible per-query consumption has now become a measurable force in national and global electricity demand, creating a new layer of strain and urgency for the energy transition and infrastructure planning.

This heightened energy burden converges most visibly on AI data centers (AIDCs), the physical backbone of the digital world. Historically built to serve cloud computing, streaming platforms, and enterprise colocation, data centers are now being reconfigured to accommodate the exponential complexity of AI workloads. According to the International Energy Agency (IEA), global data centers consumed approximately 415 terawatt-hours of electricity in 2024, equivalent to about 1.5% of the world's total electricity supply.

Yet, this baseline represents only the beginning of a steep climb. As accelerated computing becomes central to AI inference and training, electricity consumption in data centers is projected to more than double to roughly 945 terawatt-hours by 2030 under the agency's Base Case scenario.[3] These trends reveal AI's astonishing energy pressure on grid systems across the globe.

Despite the tremendous electricity demands of AIDCs, governments and industry leaders worldwide persist in advancing new construction to remain at the forefront of the AI revolution. This competitive drive—often framed as an 'AI arms race'—has deepened the urgency to deploy more powerful computing nodes. Yet, this approach rarely acknowledges the profound infrastructural reforms required to underwrite truly sustainable and reliable AI growth. Transitional strategies to address this challenge largely focus on incremental increases in grid supply or localized mitigation measures, in effect postponing the deeper transformation of electricity markets, grid planning, and regulatory frameworks needed for systemic adaptation.

This article interrogates the complex interface between AI data center expansion and electricity systems at both technical and policy levels. First, we illustrate the current grid stress introduced by AI data centers. Second, we assess the range of development strategies adopted by countries of varying scale. Finally, the discussion pivots to promising policy instruments—such as dynamic locational incentives, robust renewable procurement targets, demand response integration, and load flexibility mandates—that can collectively fashion a more resilient and responsive grid. These reforms are critical for reconciling the promise of AI with the imperatives of energy security and social legitimacy in the digital era.

THE MAJOR CHALLENGES OF AIDC ENERGY DEMAND

Unpredictable Peak Loads

Contrary to conventional assumptions about the consistency of data center operations, AI facilities leveraging large-scale graphics processing unit (GPU) clusters for both model training and inference exhibit pronounced workload volatility . Rather than sustaining stable, round-the-clock utilization, these clusters generate significant variability in electricity demand, often spiking or plummeting by hundreds of megawatts within mere seconds.[4] The energy demands of AI training runs, wherein hundreds of thousands of GPUs transition in tandem between phases of intense calculation and periods of relative idleness, are especially volatile.

Such rapid, large-magnitude fluctuations complicate not only energy supply planning, but also the thermal management systems tasked with ensuring equipment reliability. While data center hardware requires a steady baseline of power to avoid the risk of damage or costly outages, the sporadic bursts of GPU activity create cycles of demand that modern grid infrastructure struggles to accommodate effectively. These cycles can push existing transformers, power distribution units, and thermal controls to their operational limits, amplifying the potential for faults and performance degradation.[5]

These power swings introduce further engineering dilemmas, particularly when peaks coincide with the resonance frequencies of turbine generator rotors within the grid. When AI-driven power fluctuations synchronize with mechanical rotor frequencies, systems can experience damaging vibrations and accelerate mechanical fatigue, causing premature failures and significant financial losses for grid operators.[6] Aging grid systems face especially acute risks from these load shifts exacerbating longstanding challenges in balancing diverse energy sources and maintaining system-level voltage and power quality.

In effect, the unpredictable nature of AI data center peak loads poses an additional challenge to the resilience and safety of the existing grids, which are already under stress by existing issues such as managing increasingly diverse energy sources.

Grid Infrastructure

The impediments to data center expansion extend well beyond demand volatility to encompass fundamental deficiencies in grid infrastructure itself. According to recent research from the Lawrence Berkeley National Laboratory, the bottleneck in the United States is less about aggregate generation capacity and more about the bureaucratic and technical complexities of the interconnection approval process. Between 2000 and 2018, only 19% of projects that submitted interconnection requests had reached commercial operations by the end of 2023, representing a mere 14% of the total proposed capacity.[7] This protracted timeline reflects procedural inefficiencies as well as a mismatch between infrastructure planning cycles and rapidly accelerating digital infrastructure buildouts.

Beyond procedural delays, the grid's physical fabric itself is aging and ill-suited to contemporary energy landscapes. More than 40% of transmission and distribution assets in advanced economies are over two decades old. These systems were designed for centralized, fossil fuel-based generation systems rather than the distributed renewable sources and concentrated, energy-intensive loads

that now characterize the modern grid.[8] This legacy architecture imposes technical constraints on voltage stability, thermal tolerance, and dynamic load balancing, rendering existing systems increasingly vulnerable to both climate-related shocks and the variable demand profiles introduced by large GPU clusters. Continued AIDC buildout will therefore demand not only greater capacity but a structural reconfiguration of grid topologies to handle high-voltage transmission, fine-grained real-time coordination, and flexible, multidirectional power exchanges that reflect the increasingly interactive nature of energy systems.[iii]

These systemic shortcomings manifest acutely in high-density data center markets. Virginia's Loudoun County, home to the world's largest data center concentration and responsible for handling approximately 70% of global internet traffic, now confronts grid connection delays extending up to seven years.[9] Utility provider Dominion Energy has openly acknowledged its inability to deliver the required power via existing overhead transmission lines, stalling billions of dollars in planned development and threatening the county's fiscal foundation, which derives roughly one-third of its tax revenue from data center operations.

Similarly, Ireland's state-owned electricity provider EirGrid implemented a de facto moratorium on new data center grid connections in the Greater Dublin Area in late 2021, citing unsustainable strain on regional grid capacity. This restriction, anticipated to persist until at least 2028, has left major facilities—including a fully constructed Digital Realty campus at Grange Castle—sitting inactive while awaiting grid access.[10] These cases illustrate the severity of transmission bottlenecks and the growing divergence between digital ambitions and the energy infrastructure required to sustain them.

Imbalances in Renewable Energy Markets

Data center operators have emerged as dominant actors in global renewable energy procurement, fundamentally reshaping clean power markets in ways that raise both structural and distributional concerns. In 2024, these operators accounted for 43% of all clean power purchase agreements signed globally, representing an unprecedented concentration of renewable procurement in a single sector.[11] The scale of this buying power is most visible among hyperscale cloud companies: AWS, Microsoft, and Google collectively control renewable procurement volumes that rival or even exceed those of major utilities. This consolidation enables favorable contract terms, early access to new renewable capacity, and premium pricing power—but it also amplifies market inequality by channeling the benefits of decarbonization toward a few corporate buyers while limiting access for smaller or locally anchored actors.

Despite the ostensible commitment of major data center operators to renewable energy, the mechanisms through which these commitments are realized—primarily Power Purchase Agreements (PPAs) and Renewable Energy Certificates (RECs)—cast doubt on their decarbonization efficacy. Most PPAs employed by data center operators are virtual arrangements, wherein renewable electricity generated under contract flows into the general grid rather than being physically delivered to the data center itself. Facilities thus continue to consume conventional grid electricity, which is often heavily reliant on fossil fuels, even while claiming "100 percent renewable" status through financial and accounting constructs.[12] A recent article published by Nature has demonstrated that

iii For instance, modernized grid configuration, such as bidirectional energy flow, can move electricity both from central generation to end users and from distributed assets back into the grid, such as buildings with rooftop solar or battery arrays near data centers, to inject surplus electricity back into local distribution networks, reducing peak stress and improving overall grid resilience.

volume-matching PPAs of this kind "drive little to no change in system-level CO_2 emissions," as they fail to address the temporal misalignment between intermittent renewable generation and actual consumption patterns.[13]

Moreover, the aggressive procurement strategies of hyperscale operators introduce the risk of market crowding, wherein smaller buyers, including small and medium-sized enterprises (SMEs), municipalities, and universities, are effectively priced out or displaced from renewable energy markets. Because data center operators are typically willing to pay premium prices to secure long-term PPAs and RECs in pursuit of sustainability commitments, they command developer attention and financing priority. In some cases, local SMEs may find themselves unable to access renewable energy even from geographically proximate sources, as developers preferentially allocate output to large, anchor tenants capable of underwriting entire projects.[14] This dynamic exacerbates inequality in the clean energy transition and could have the effect of delaying progress for embedding sustainability practices across global industry.

A growing body of research suggests that democratizing renewable procurement by empowering local community buyers could drive more equitable and durable decarbonization outcomes. Researchers from Ireland, for instance, argue that local consumers become active producers or "prosumers," fosters long-term political legitimacy and social acceptance of clean energy transitions.[15] By contrast, massive energy projects driven by multinational corporations (MNCs) often encounter NIMBY opposition, as local communities see limited benefits while absorbing the social, environmental, and infrastructural footprint of large-scale development.[16]

Although it is encouraging to witness data center operators embracing renewable energy at an unprecedented scale, it remains crucial to question whether this expansion effectively fosters broader decarbonization. The long-term side effects—including delaying diffusion of sustainable practices, weaker community buy-in, and an eroded social mandate for grid modernization—may outweigh the short-term gains. These trade-offs underscore that achieving genuine decarbonization will depend not only on technical progress but also on the inclusivity and fairness of the energy systems that power it.

WORLDWIDE ENERGY STRATEGIES FOR AI DATA CENTRES

US and China: Powering AI at Scale, But Transmission at Risks

Large economies such as China and the United States have pursued divergent pathways in addressing the spatial misalignment between data center demand and available power supply, each confronting distinctive infrastructural and political constraints. For these countries, the central question has been whether to relocate computational workloads to regions endowed with greater generation capacity or instead to rely on long-distance transmission to deliver electricity from remote sources to established data center clusters. These approaches involve a wide range of infrastructural investments and institutional calibration, ranging from fiber optic deployment, transmission grid expansion, and improving market coordination mechanisms. The breadth and scope of these measures reveal the profound systemic challenges that AI data centers impose on national energy systems, even in large and dynamic economies.

China's dual strategies of "West-to-East Power Transfer (**西电东送**,WTEPT)" and "Eastern Data,

Western Computing (东数西算, EDWC)" seek to leverage both engineering transmissions for electricity and relocating computational workloads. The WTEPT relies on decades of government investment in ultra-high-voltage direct current transmission lines designed to carry electricity across thousands of kilometers from renewable-rich northwestern provinces (Gansu, Qinghai, Ningxia, Inner Mongolia) to densely populated eastern cities with major industrial clusters.

Yet, despite these engineering achievements, the strategy encounters persistent bottlenecks: renewable generation concentrated in the northwest frequently exceeds local grid absorption capacity. Recent data shows curtailment rates rebounding as solar and wind capacity in remote areas surge; in early 2025, solar curtailment in provinces like Qinghai exceeded 15%, while national limits on acceptable curtailment were relaxed from 5% to 10% to cope with integration challenges.[17] These bottlenecks slow grid connections for new projects and reveal the limits of relying primarily on bulk transmission to "push" renewables transmission without a certain level of local demands.

To mitigate long-distance transmission constraints, the Chinese government launched the EDWC strategy to relocate data processing workloads themselves closer to generation sites in western China. Although this initiative attempts to expand locally consumed renewable energy by bringing in data center industries, it exposes a raft of new operational barriers. Many western data centers were established in regions without adequate fiber-optic backbones, resulting in high capital expenditures to extend digital connectivity and a persistent risk of network congestion when transmitting large AI datasets over long distances. Furthermore, the highly specialized labor required to operate and maintain these facilities is often scarce in remote western provinces and the local customer base remains limited, deterring long-term commercial investment and undermining the broader economic development agenda. As a result, high-value and low-latency computational workload continues to be processed primarily in the east, in areas such as Shanghai and Shenzhen, causing waste of computational resources and inefficient development of data center assets.[18]

By contrast, the United States confronts weak interregional connectivity and limited high-voltage transmission infrastructure, constraining its ability to redistribute power across vast geographic distances. The nation's power grid is fragmented into three electrically isolated interconnections—the Eastern, Western, and Texas grids—each governed by distinct regulatory regimes and planned in relative isolation by more than a dozen regional transmission organizations. This fragmentation impedes coordination and limits the capacity to move surplus renewable energy from one region to serve load growth in another.[19] In Virginia's data center corridor, soaring electricity demand has driven up wholesale power costs and provoked widespread community opposition.[20] At the same time, residential electricity bills in neighboring Maryland rose by up to 20% as a consequence of grid capacity expansions necessitated by data center buildout.[21] Utilities projects that serve rising data-center loads will push up retail bills, while transmission and siting constraints slow alternatives such as relocating capacity to other zones.[22]

Recent policy shifts in the United States have strongly encouraged co-location of data centers with dedicated power plants. A federal executive order streamlines permitting for gas, coal, nuclear, and other dispatchable generation built principally to serve data centers, while multiple states have loosened utility rules to allow private, behind-the-meter gas plants for AI projects and other off-grid facilities.[23] However, such configurations cannot address workload volatility in isolation and still require robust grid integration to balance fluctuations and ensure reliability. Moreover,

co-locating data centers with thermal power plants—whether fossil-fueled or nuclear—intensifies water stress, as both data center cooling systems and thermoelectric generation consume large volumes of freshwater.[24] In water-stressed regions such as Arizona and Maryland, this dual exploitation has led to push-back from local communities who question the sustainability and equity of prioritizing digital infrastructure over household and agricultural water needs.[25]

Both the U.S. and China possess the scale to pursue flexible grid and compute resource allocation, but the effectiveness of these approaches ultimately turns on the modernization of their respective transmission networks and the careful orchestration of digital-industrial development. The U.S. faces far steeper challenges: transmission capacity expansion, market integration, and load balancing are necessary to prevent congestion, local grid saturation, and the economic displacement of entire regions. Despite the advantages conferred by size, neither country is immune from the competing demands of infrastructure renewal and sustainable, distributed energy development—a central dilemma for powering the AI era at scale.

Ireland, Singapore, and Taiwan: Hubs Facing Limits of Scale

For smaller and geographically constrained states such as Singapore, Ireland, and Taiwan, the range of solutions available to manage AI data center electricity demand is far narrower. These regions are unable to redistribute either workloads or grid power domestically in any meaningful way, nor can they scale their generation capacity quickly enough to match surging digital demand. Reflecting these structural limitations, all three have introduced restrictions on new data center grid connections in their most congested zones: Singapore rolled out a nationwide moratorium on data center construction between 2019 and 2022; Ireland has capped grid access for additional facilities in the capital region; and Taiwan has similarly required hyperscale data centers exceeding 5 MWof load demand to upgrade their technologies for energy optimization.[26]

In addition to expanding domestic infrastructure upgrades, these states are also pursuing deeper cross-border electricity interconnections as a cornerstone of future-proof digital and energy policy. Singapore, for instance, is actively exploring power importation from Malaysia and longer-term options to import low-carbon electricity via subsea cables from Australia.[27] Its grid is already linked to Johor in Malaysia, and the industrial players are exploring floating or offshore data center concepts to support both land and energy constraints.[28]

Ireland, for its part, is advancing the Celtic Interconnector, a 700 MW high-voltage direct current submarine cable to France, which is projected to be operational by 2026–2028. The project will strengthen Ireland's grid resilience, integrate more renewables, and improve its connectivity to the broader European power network.[29]

Taiwan, similarly confronted by surging local AI and semiconductor industry demand, has undertaken preliminary studies on the feasibility of HVDC submarine links to the Philippines to enable renewable energy imports. While such cross-border transmission is technologically achievable, the broader geopolitical risks render it an uncertain prospect. Submarine cables in the waters surrounding Taiwan have faced intermittent sabotage and accidental damage, underscoring how escalating regional tension makes energy interconnection a risky and potentially unreliable pathway for securing the island's long-term power resilience.[30]

On the regulatory front, each of these smaller nations have adopted diverse approaches to man-

aging the energy demands of data centers within constrained grids. Ireland's energy regulator proposed that new data centers provide onsite dispatchable generation or storage capacity commensurate with their demand as a condition for grid connection.[31] While framed as a mechanism to relieve grid strain and accelerate renewable deployment, the policy has been heavily criticized for lacking explicit renewable or storage mandates, opening the door instead to fossil fuel-based backup generation.[32] By mid-2025, Ireland moved to permit private wire connections between data centers and gas-fired power plants. Critics warn that without binding renewable requirements, the policy risks entrenching natural gas dependency and diverting both renewable energy and bioenergy resources away from harder-to-abate sectors such as heating and heavy industry.[33]

Singapore has instead focused on operational efficiency and climate adaptation rather than on-site generation mandates. It introduced the Tropical Data Centre Standards, which guide data center operators in safely raising operating temperatures to 26°C and above (up to 35°C for IT equipment)—thereby reducing the energy burden of cooling systems in Singapore's hot and humid environment. The first of their kind globally, these standards aim to achieve 2-5% cooling energy savings for every 1°C increase in operating temperature, as well as broader IT equipment efficiency improvements of up to 30%.[34] While these measures address the thermal challenges specific to tropical climates and support Singapore's green data center ambitions, land scarcity and elevated real estate costs remain persistent obstacles to scaling capacity domestically.[35]

Taiwan's electricity system, meanwhile, is centralized under the state-owned utility Taiwan Power Company (TPC), which devises the comprehensive energy programmes from generation to transmission. This unified governance theoretically enables more coordinated planning and load management than fragmented markets elsewhere. However, in practice Taiwan faces acute strain from both the concurrent rise of energy intensive-industries such as AI data centers and semiconductor manufacturing, and the limited lands available for power generation or transmission infrastructure. The energy pressure may further climb as advanced node production scales up and GPU-intensive AI workloads proliferate.[36] The convergence of these energy-intensive industries threatens to overburden an already constrained grid, complicating the island's energy transition and heightening vulnerability to supply disruptions in a geographically isolated and resource-limited context.

POLICY IMPLICATIONS

AI data centers have brought about a broader inflection point for energy governance, in which digital and power infrastructures must be planned as an integrated system rather than as separate sectors. Their hardware footprints demand greater physical spacing, more intensive cooling, and improved grid connections, even as climate change progressively erodes generation efficiency and cooling performance.[37] In this context, long-term grid planning must embed future climate scenarios and local thermal stress into strategic decisions around siting, capacity expansion, and resilience.

In the past, operators could often alleviate costs and emissions by shifting delay-tolerant batch workloads to times and locations with abundant low-carbon electricity—a practice that often yielded more effective decarbonization impacts than merely shifting electricity supply.[38] Yet the rapid spread of latency-sensitive edge computing and real-time AI services is gradually reducing that flexibility, reinforcing the importance of adaptive pricing mechanisms and resilient grid

infrastructure that can transmit accurate signals for both investment and operations across space and time.

In practice, the capital-intensive nature of data center operations make them relatively insensitive to short-term price fluctuations, meaning that protective policy solutions such as electricity tariffs alone are unlikely to be sufficient to incentivize substantial demand response behavior.[39] This structural reality elevates the importance of interventions designed to enhance operational efficiency. In addition to mere market driven dynamic pricing, policy frameworks that set minimum on-site storage ratios, institutionalize demand response programs, and offer targeted financial incentives for flexibility and efficiency can shift operational behavior in ways that align private decisionmaking with system level needs. This, in turn, requires electricity market rules that explicitly incentivize operators to design business models around temporal and spatial workload shifting instead of treating computing demand as exogenous and fixed.

Furthermore, recent appeals from the data center industry for on-site generation as a catch-all solution must be tempered by a clear-eyed assessment of its limits.[40] Regulations for new data center projects should not only permit but actively structure the role of behind-the-meter power, requiring that specified portions of on-site energy sources come from renewable sources and correspondent storage options. Since AI workloads can fluctuate sharply, most large facilities will eventually need to remain tightly coupled to the broader grid to ensure reliability and to participate in system-wide balancing. Overreliance on thermal and nuclear-based energy sources that require water-intensive cooling systems risks aggravating competition for land and water, while doing little to support decarbonization at the system level.[41] Thoughtful regulatory guardrails can ensure that on-site power complements rather than substitutes for grid modernization and integrated resource planning.

Ultimately, the geography and operation of AI data centers should also be aligned with regional developments and local community interests. Siting rules that prioritize colocation with domestic industrial clusters and research ecosystems (while favoring projects that draw power from nearby renewable producers or low-carbon industrial hubs) can help ensure that new infrastructure strengthens local value creation instead of merely exporting benefits to distant cloud clients. Carefully calibrated locational incentives, connection conditions, and planning standards can steer investment toward areas where grid capacity, renewable potential, and industrial competitiveness intersect. In doing so, governments can transform AI infrastructure from a source of mounting grid stress into a lever for accelerating the energy transition, anchoring clean generation, storage, and flexible demand in ways that reinforce both AI and sustainable development.

Endnotes

1 McKinsey & Company, "New Front Door to the Internet: Winning in the Age of AI Search," October 16, 2025, https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/new-front-door-to-the-internet-winning-in-the-age-of-ai-search.

2 N. Jegham et al., "How Hungry is AI? Benchmarking Energy, Water, and Carbon Footprint of LLM Inference," May 15, 2025, https://arxiv.org/abs/2505.09598.

3 International Energy Agency, "Energy and AI," April 10, 2025, https://www.iea.org/reports/energy-and-ai.

4 X. Chen et al., "Electricity Demand and Grid Impacts of AI Data Centers: Challenges and Prospects," September 12, 2025, https://arxiv.org/abs/2509.07218.

5 M. Mughees et al., "Short-Term Load Forecasting for AI-Data Center," March 13, 2025, https://arxiv.org/abs/2503.07756.

6 E. Choukse et al., "Power Stabilization for AI Training Datacenters," August 24, 2025, https://arxiv.org/abs/2508.14318.

7 Lawrence Berkeley National Laboratory, "Grid Connection Backlog Grows by 30% in 2023, Dominated by Requests for Solar, Wind, and Energy Storage," Energy Markets & Planning, April 10, 2024, https://emp.lbl.gov/news/grid-connection-backlog-grows-30-2023-dominated-requests-solar-wind-and-energy-storage

8 C. Dallas, "VantagePoint: Electrifying Returns in the AI Era," Cambridge Associates, November 3, 2025, https://www.cambridgeassociates.com/insight/vantagepoint-electrifying-returns-in-the-ai-era/.

9 J. Sual, "Data Centers Face Seven-Year Wait for Dominion Power Hookups," Bloomberg, August 20, 2025, https://www.bloomberg.com/news/articles/2024-08-29/data-centers-face-seven-year-wait-for-power-hookups-in-virginia.

10 M. Urie, "Dublin: The Heart of Ireland's Data Centre Boom," Gardiner & Theobald Market Intelligence, May 22, 2025, https://www.gardiner.com/marketintel/dublin-the-heart-of-irelands-data-centre-boom.

11 J. Lyu and S. Tang, "Power Hungry Data Centers Are Driving Green Energy Demand," BloombergNEF, August 26, 2025, https://about.bnef.com/insights/power-hungry-data-centers-are-driving-green-energy-demand/.

12 Q. Xu et al., "System-Level Impacts of Voluntary Carbon-Free Electricity Procurement Strategies," Joule 8, no. 2 (February 2024): 405–431, https://doi.org/10.1016/j.joule.2023.12.007.

13 A. Bjørn et al., "Renewable Energy Certificates Threaten the Integrity of Corporate Science-Based Targets," Nature Climate Change 12, no. 6 (June 2022): 539–546, https://doi.org/10.1038/s41558-022-01379-5.

14 Schneider Electric Advisory Services, "How Aggregated PPAs Create Clean Energy Options for SMEs," SE Perspectives (blog), June 6, 2022, https://perspectives.se.com/blog-stream/how-aggregated-ppas-create-clean-energy-options-for-smes.

15 C. Watson et al., Responding to the Energy Transition in Ireland: The Experience and Capacity of Communities (Frankfurt: Environmental Protection Agency, 2020).

16 C. Watson, "Responding to Climate Change and the Energy Transition: The Experience and Capacity of Communities in Ireland" (University College Cork, 2020).

17 "China Faces Rising Renewable Energy Curtailment," Power Technology, August 6, 2025, https://www.power-technology.com/news/china-renewable-energy-curtailment/.

18 DC Market Insights, China's Data Center Market Size, Growth, and Prediction Report until 2035 《中国数据中心市场规模、增长及预测报告 2035》，2025年，https://www.dcmarketinsights.com/zh/report/china-data-center-market-zh.

19 W. Gorman et al., "Grid Connection Barriers to Renewable Energy Deployment in the

United States," Joule 9, no. 2 (February 2025): 285–310, https://www.sciencedirect.com/science/article/pii/S2542435124005038

20 S. Kimball, "Skyrocketing Electricity Prices Fuel Political Backlash against Tech Sector's AI Data Centers," CNBC, November 12, 2025, https://www.cnbc.com/2025/11/12/electricity-prices-data-center-ai-new-jersey-virginia-midterm-election.html.

21 J. Saul, "AI Data Centers Are Sending Power Bills Soaring," Bloomberg, September 30, 2025, https://www.bloomberg.com/graphics/2025-ai-data-centers-electricity-prices/.

22 L. Kearney and T. McLaughlin, "Power Costs Soar in PJM Region as Data Center Demand Spikes," Reuters, August 7, 2025, https://www.reuters.com/business/energy/power-costs-soar-pjm-region-data-center-demand-spikes-2025-08-07/

23 U.S. President, "Accelerating Federal Permitting of Data Center Infrastructure," Presidential Memorandum, July 23, 2025, https://www.whitehouse.gov/presidential-actions/2025/07/accelerating-federal-permitting-of-data-center-infrastructure/; Oklahoma State Senate, "Governor Signs Senator Green's 'Behind the Meter' Bill to Boost Economic Development," press release, May 15, 2025, https://oksenate.gov/press-releases/governor-signs-senator-greens-behind-meter-bill-boost-economic-development; K. Clay, "The AI Boom Is Changing the Way the Grid Is Governed," Washington Post, February 19, 2026, https://www.washingtonpost.com/business/2026/02/19/data-centers-power-grid-ai/

24 M. Yanez-Barnuevo, "Data Centers and Water Consumption," Environmental and Energy Study Institute, June 25, 2025, https://www.eesi.org/articles/view/data-centers-and-water-consumption.

25 J. Saul, "AI Data Centers Are Sending Power Bills Soaring," Bloomberg, September 30, 2025, https://www.bloomberg.com/graphics/2025-ai-data-centers-electricity-prices/.

26 Ministry of Economic Affairs (MOEA), "現行的能源使用說明書審查 納入一定規模以上資訊服務業" [Current Energy Use Manual Review to Include Information Service Industries Above a Certain Scale], press release, November 4, 2025, https://www.moea.gov.tw/MNS/populace/news/News.aspx?kind=1&menu_id=40&news_id=120969.

27 L. Jackson, "Singapore Approves Import of Solar Energy from Australia via Undersea Cable," Reuters, October 22, 2024, https://www.reuters.com/business/energy/singapore-approves-import-solar-energy-australia-via-undersea-cable-2024-10-22/.

28 K. Kapoor, "Singapore Greenlights New Plans to Import Malaysian Clean Power," Bloomberg, October 17, 2025, https://www.bloomberg.com/news/articles/2025-10-17/singapore-greenlights-new-plans-to-import-malaysian-clean-power; V. Veersamy et al., "Grid Infrastructure and Renewables Integration for Singapore Energy Transition," Scientific Reports 15, no. 1 (October 2025): 1–18, https://www.nature.com/articles/s41598-025-17376-5

29 Celtic Interconnector, "Cable Laying Begins Marking Key Milestone for Celtic Interconnector Project," August 1, 2025, https://www.celticinterconnector.eu/cable-laying-begins-marking-key-milestone-for-celtic-interconnector-project/.

30 C. Chang et al., "Vulnerabilities at Depth: Submarine Power Cable Sabotage and Supply Chain Risks Amid China's Rise," Research Report, DSET (Taiwan School of Geosatellite and Economics & Technology), October 2025, https://dset.tw/research/vulnerabilities_at_depth/.

31 Government of Ireland, "Government Statement on the Role of Data Centres in Ireland's Enterprise Strategy," July 2025, https://enterprise.gov.ie/en/publications/publication-files/government-statement-on-the-role-of-data-centres-in-irelands-enterprise-strategy.pdf.

32 D. Swinhoe, "Ireland's Energy Regulator Proposes Policy Requiring Data Centers to Match Load with New Power Generation," Data Center Dynamics, February 19, 2025, https://www.datacenterdynamics.com/en/news/irelands-energy-regulator-proposes-policy-requiring-data-centers-to-match-load-with-new-power-generation/.

33 "Ireland to Allow Data Centers to Link Up to Fossil Fuel Plants," Energy Connects, September 16, 2025, https://www.energyconnects.com/news/renewables/2025/september/ireland-to-allow-data-centers-to-link-up-to-fossil-fuel-plants/.
34 Infocomm Media Development Authority (IMDA), "Singapore IT Energy Efficiency Standard for Data Centres Launched," press release, August 13, 2025, https://www.imda.gov.sg/resources/press-releases-factsheets-and-speeches/press-releases/2025/sg-it-energy-efficiency-standard-for-data-centres-launched.
35 S. Gupta, "Amid the AI Boom, Singapore Surfaces as the Second-Most Expensive Market to Construct Data Centres," Business Times (Singapore), November 6, 2025, https://www.businesstimes.com.sg/startups-tech/technology/amid-ai-boom-singapore-surfaces-second-most-expensive-market-construct-data-centres.
36 C. Trueman, "TSMC Could Account for 24% of Taiwan's Electricity Consumption by 2030," Data Center Dynamics, October 7, 2024, https://www.datacenterdynamics.com/en/news/tsmc-could-account-for-24-of-taiwans-electricity-consumption-by-2030/; T. Lu et al., "Climate Change and Infrastructure Resilience: An Analysis Report on Water and Electricity Use in Taiwan's Semiconductor Industry," Research Report, DSET, February 2025, https://dset.tw/research/infrastructure-resilience/
38 M. Maomood et al., "Impacts of Digitalization on Smart Grids, Renewable Energy, and Demand Response: An Updated Review of Current Applications," Energy Conversion and Management: X 24 (October 2024), https://doi.org/10.1016/j.ecmx.2024.100790
42 I. Riepin, T. Brown, and V. Zavala, "Spatio-temporal Load Shifting for Truly Clean Computing," Advances in Applied Energy 17 (March 2025), https://www.sciencedirect.com/science/article/pii/S2666792424000404
43 C. Crozier and M. Liska, "The Potential of Data Center Energy Demand to Provide Grid Flexibility," Current Sustainable/Renewable Energy Reports 12, no. 12 (April 2025): 1–12, https://link.springer.com/article/10.1007/s40518-025-00258-9
44 Bloom Energy, "Data Centers Are Turning to Onsite Power Sources to Address 35 GW Energy Gap by 2030," press release, January 21, 2025, https://investor.bloomenergy.com/press-releases/press-release-details/2025/Data-Centers-Are-Turning-to-Onsite-Power-Sources-to-Address-35-GW-Energy-Gap-by-2030/default.aspx.
45 University of Tulsa, "Data Centers Draining Resources in Water-Stressed Communities," July 19, 2024, https://utulsa.edu/news/data-centers-draining-resources-in-water-stressed-communities/.

Chapter 3: The Geopolitics of the AI Buildout: Deconstructing Sovereign AI

By Hanna Dohmen[i] and Sam Bresnick[ii]

KEY TAKEAWAYS:

- Complete AI sovereignty is impractical for most countries. While some governments may aspire to control the entire AI stack (including chips, data, and models), this chapter argues that a "hybrid sovereignty" model is emerging.
- Compute is the layer where foreign dependencies are the most difficult to reduce. U.S. companies dominate AI chip design and cloud infrastructure, while firms in Taiwan and South Korea lead chip fabrication. Even well-funded national data center buildouts still rely on access to foreign-designed and -fabricated hardware, and this dynamic is unlikely to change in the near term.
- Countries have more flexibility and autonomy regarding data. Governments are already asserting control at the data layer through residency requirements, domestic cloud environments, and national datasets, driven by concerns around security, privacy, and cultural and linguistic representation in AI systems.
- At the model layer, open-weight models, which can be fine-tuned for domestic needs, are lowering the barrier to entry for countries with resource constraints. While some countries are developing their own models, most will continue to rely on those developed by a handful of companies in the United States, China, and Europe.

INTRODUCTION

Many countries view artificial intelligence (AI) as critical to economic competitiveness and national security. As a result, sovereign AI—the idea that national governments should develop, control, and govern AI in order to boost economic growth, guarantee security, and ensure strate-

i Hanna Dohmen is a Senior Research Analyst at Georgetown's Center for Security and Emerging Technology (CSET), where she focuses on U.S. national competitiveness in emerging technologies and U.S.-China technology competition.

ii Sam Bresnick is a Research Fellow and an Andrew W. Marshall Fellow at Georgetown's Center for Security and Emerging Technology (CSET), focused on the national security applications of artificial intelligence, U.S.-China tech competition, and Chinese technology policy.

gic autonomy—has become a key strategic consideration in the global AI buildout. Amid the push for sovereign AI, countries are racing to set up their own domestic compute infrastructure, curate local datasets, and train or fine-tune their own advanced large language models (LLMs).

Despite the popularity of the term "sovereign AI," it remains ill-defined, and each country's efforts in this area embrace varying levels of sovereignty. For some governments, sovereignty means controlling the full hardware-software stack. For others, it is the ability to develop AI models trained on domestic datasets to reduce dependence on foreign companies.

This paper interrogates the concept of "sovereign AI" through the lens of the technology stack: (1) compute infrastructure, (2) data, and (3) models. By analyzing the geopolitical dynamics at each layer, we argue that full sovereignty is economically and technologically infeasible for most nations, while a "hybrid sovereignty" model is emerging globally. Therefore, interdependence will continue to be the norm rather than the exception. For countries seeking to secure reliable AI access, this poses a potential challenge in an increasingly unstable international environment, as dependence on foreign companies means that access can be conditioned, restricted, or revoked in response to geopolitical shifts.

COMPUTE INFRASTRUCTURE

At the base of the AI stack lies compute infrastructure, advanced semiconductors like GPUs and AI accelerators, as well as the servers and data centers that house them. This is the most capital-intensive layer of the AI stack and remains the one where "sovereignty" efforts face the greatest challenges. Even if governments localize their compute infrastructure, as many are now doing, they will remain for the foreseeable future reliant on U.S. chip design companies for access to the semiconductors that power most AI models and applications.

The leading AI chips used to train and deploy frontier generative AI models are overwhelmingly designed by U.S. companies, most notably NVIDIA. NVIDIA's market share is estimated to be roughly between 70 to 95 percent of the global market for AI chips.[1] AMD, NVIDIA's main competitor, is also rapidly increasing production of its most advanced AI semiconductors and has recently announced partnerships with LLM developers like OpenAI and Meta.[2] The company nonetheless has a significantly lower global market share than NVIDIA.[3]

While U.S. firms lead the AI semiconductor market, Chinese AI chip design firms are making inroads, as Huawei HiSilicon, Baidu, and Biren have released high-performance AI chips in recent years.[4] Chinese companies have progressed, but their best products remain well behind the most sophisticated offerings of their U.S. counterparts. For example, the leading Chinese chips only reach a theoretical performance equal to roughly 40 percent of NVIDIA's B200—one of the company's most performant chips.[iii] While U.S. export controls have hindered China's ability to acquire the most advanced chips, the policy has likely also further accelerated domestic chip design progress and pushed the Chinese government to further support its domestic chip industry.

Reinforcing the United States' dominance in AI chips are U.S. and allied export controls and China's lagging indigenous fabrication capabilities. While U.S. chip designers can partner with Taiwan

iii According to the authors' calculations, Biren's Bili 100 series chips have a total processing performance (TPP) of 15,400-16,400, which is 39-41 percent of NVIDIA's B200 TPP.

Semiconductor Manufacturing Company (TSMC) to fabricate chips using the most advanced process nodes, export controls have forced Chinese companies to rely on domestic producers, which lag their international competitors. Semiconductor Manufacturing International Corporation (SMIC) and other Chinese fabrication companies have reached 7 nanometer (nm) process nodes and limited production on 5nm process nodes (the cutting-edge has reached 2nm), but yields remain low and costs high, thus keeping Chinese firms far behind leading fabs.[5] That said, Beijing is pushing to reduce China's dependence on foreign chipmaking equipment. In late 2025, the Chinese government mandated that domestic chipmakers source at least 50 percent of their equipment from Chinese suppliers, which is intended to accelerate the country's fabrication capabilities.[6] For now, however, Chinese toolmakers and fabs face significant constraints.

Not only are the vast majority of U.S.-designed AI chips fabricated by companies based in U.S.-aligned countries, but they are also largely deployed in data centers primarily run by U.S.-headquartered cloud service providers (CSPs), such as Microsoft Azure, Amazon Web Services, and Google Cloud Platform. According to some estimates, these three CSPs collectively account for over 60 percent of global hyperscaler data centers, which are optimized for AI.[7] European CSPs' market share in Europe dropped from almost 30 percent in 2017 to just 15 percent in 2022, but their share has remained steady since then.[8]

U.S. chip and cloud dominance has direct geopolitical implications for countries focused on AI sovereignty. Those that want to build out their national AI capacity, from France and Germany to the United Arab Emirates (UAE), India, and Japan, are now trying to stand up their own sovereign compute clusters. These efforts, however, paradoxically depend on foreign technology. Due to the market conditions described above, to build "sovereign" cloud computing infrastructure, these countries need cutting-edge U.S.-designed chips. This creates a significant challenge, as AI sovereignty requires countries to increase their reliance on the U.S. hardware supply chain. So long as Chinese chip production capabilities lag those of the United States and its allies, this dynamic will continue.

The case of the UAE starkly illustrates this challenge. In May 2025, G42, an Abu Dhabi-based company with close ties to the Emirati government, announced a collaboration with U.S. and Japanese companies Oracle, OpenAI, NVIDIA, Cisco, and SoftBank Group to develop Stargate UAE, a planned 1-gigawatt compute cluster built by G42 and operated by OpenAI and Oracle.[9] To that end, the UAE received approval from the U.S. Department of Commerce in November 2025 for the sale of up to 35,000 NVIDIA GB300 servers or their equivalents to G42.[10] While this partnership between the Emirati company and U.S. tech giants provides the UAE with the ability to expand its local data center capacity, the country's ambitions remain tied to U.S. government approval of chip sales, leaving its national strategy exposed to the geopolitical whims of the United States.

A similar dynamic prevails in Europe, where countries such as France and Germany are pursuing "digital sovereignty" in part to reduce dependence on U.S. and Chinese technologies.[11] In November 2025, the French and German governments announced a strategic public-private partnership between French company Mistral AI and German cloud computing company SAP.[12] France's Minister of the Economy, Finance, Industrial, Energy, and Digital Sovereignty and Germany's Federal Minister for Digital Transformation and Government Modernisation explicitly labeled this partnership a step toward greater digital sovereignty in the European Union (EU).[13] Moreover, a study by the Economic Governance and EMU Scrutiny Unit (EGOV) of the European Parliament, pub-

lished in December 2025, identified the need for Europe to achieve AI and cloud sovereignty as a key step to enhance Europe's autonomy and competitiveness.[14] Similar to the UAE, however, the EU's sovereign AI ambitions will also remain reliant on U.S. companies for the foreseeable future.

In addition, compute buildouts also carry significant energy requirements, raising questions about grid capacity, power generation, and broader electricity costs for those states with sovereign AI ambitions. A full discussion of AI's energy demands is beyond the scope of this chapter, but power availability is a significant constraint for some countries.

For most nations, AI sovereignty in compute infrastructure is not about complete self-sufficiency, but about localization and access. By hosting data centers and compute buildouts domestically, these countries aim to ensure that, while they may be dependent on foreign-designed and -fabricated chips, they maintain physical control over their compute infrastructure. Given the difficulties of designing and producing cutting-edge AI chips, the vast majority of countries will likely focus on compute localization rather than establishing sovereign AI semiconductor supply chains and fabs.

DATA

Frontier generative models require vast quantities of data, including text, images, audio, and video. For governments seeking to develop sovereign AI, the ownership and governance of the data layer is central to their efforts. We highlight two key sovereignty concerns related to data: (1) national security and data privacy and (2) cultural and linguistic representation. While it is unlikely that most countries will establish complete compute sovereignty, it is more feasible for them to develop and control their own proprietary data resources.

The most immediate concern for many countries is data residency—the requirement that data be stored and processed within a country's physical borders.[iv] Data related to national security, healthcare, finance, or other proprietary information is often considered too sensitive to be housed in data centers located abroad or operated by foreign companies or governments.[15] Countries see storing such data in sovereign cloud infrastructure as an opportunity to ensure that it remains within domestic legal boundaries. India, for instance, describes "maintain[ing] control over its data" as one of the key aspects of sovereign AI.[16] Data sovereignty—full jurisdictional control over data, including access, use, and transfer—is also a key area of focus for many countries. For example, France is working with local technology companies such as Capgemini and Orange to create a trusted cloud environment called Bleu.[17] This arrangement relies on French data centers that meet domestic data transfer requirements and are protected from extraterritorial data regulation and legislation.[18]

Countries are also increasingly concerned about protecting their cultures, languages, and histories through sovereign AI. These nations often develop their own LLMs, or fine-tune open-weight models, using datasets based on their national language or with nation-specific data.
For example, Core42, in collaboration with the UAE's Mohamed bin Zayed University of Artifi-

iv Data residency and data sovereignty are related but distinct concepts. Data residency refers to requirements that data be stored in a particular geographic location. Data sovereignty implies full jurisdictional control over data, including governance of access, use, and transfer. See M. Kosinski, "Data Sovereignty vs. Data Residency: What's the Difference?" IBM Think, accessed February 25, 2026, https://www.ibm.com/think/topics/data-sovereignty-vs-data-residency.

cial Intelligence (MBZUAI), has developed an open-weight, Arabic-language LLM called Jais.[19] In part, Jais's developers aim to augment the language's presence in the global AI ecosystem and provide a more linguistically and culturally representative model for Arabic speakers.[20] While the major U.S.-developed LLMs, such as ChatGPT, Gemini, and Claude, have non-English language capabilities, they are primarily trained on English-language data.[21] Jais, however, was developed using both English and Arabic datasets with the hope of providing superior Arabic-language capabilities.[22]

Similarly, Southeast Asian Languages in One Network (SEA-LION), which is part of Singapore's National Multi-Modal Large Language Model project, is a family of open-source models trained on eleven Southeast Asian languages, including Thai, Vietnamese, and Indonesian. The driving force behind SEA-LION is a desire for models that better understand Southeast Asia's diverse contexts, languages, and cultures.[23] Developers of SEA-LION cite concerns about the disproportionate influence from Western, industrialized, and rich, educated, and democratic (WIRED) societies in existing models and the risk that such an imbalance in training data may result in biases in model outputs.[24] The initiative, according to its developers, "lowers the bar for governments, industries, and academics" to adopt LLMs that "fit local languages and reflect local cultural norms."[25]

While the focus of cultural and linguistic representation has been a key aim for countries like Singapore, the UAE, and others, this goal may become less of a focus in years to come as leading LLM developers improve their models in two ways. First, developers are expanding models' vocabulary size (the internal dictionary of word fragments the model recognizes) which allows it to process foreign languages more accurately.[26] Second, developers are also increasingly training their models on larger volumes of non-English-language datasets.[27]

As countries adopt LLMs for national security applications, the importance of data sovereignty will likely increase. Every country has distinct military missions and needs, and many will be motivated to utilize local, proprietary datasets to train or fine-tune models for use in national security contexts. While the quality and availability of domestically controlled data will vary widely across countries, growing concerns about unclear dataset provenance are pushing many governments to prioritize trusted, controlled data pipelines that rely more on locally held or tightly vetted datasets for defense and military applications.

Ultimately, the aim for data sovereignty is motivated by a desire to maintain regulatory and cultural agency, as well as security. As models like Jais and SEA-LION demonstrate, many countries share the goal of preventing cultural misalignment through the embedding of foreign values in the LLMs their people and companies use. Moreover, nations will continue to see data localization as a critical requirement to protect sensitive data—whether that of citizens or governments—and to ensure the security of LLMs used for national security applications. It is at the data layer that countries have the greatest ability to guarantee their own sovereignty.

AI MODELS

The model layer is another key geopolitical battlefield. This layer can be broken into two categories: proprietary and open-source or open-weight LLMs.[v] National approaches to sovereign AI

v There are important distinctions between open-source and open-weight models. The former provide transparency about their training data, code, and methodology, allowing for auditing and reproducibility. Open-weight models, on the other hand, only share the final parameters, or weights, rather than the data and training process. This

at the model level focus primarily on either indigenous development of models or fine-tuning of foreign-trained models. While countries can develop their own AI models, the swift pace of frontier model progress makes keeping up with the state-of-the-art (SOTA) potentially prohibitively challenging.

The majority of the SOTA models are trained by U.S. developers such as OpenAI, Anthropic, and Google. Most of these companies' models are proprietary, meaning that users have little insight into or control over their training data or parameters. This lack of transparency and customizability complicates other nations' ability to ensure that these models align with their national interests, be they related to economics, security, or culture. As such, some countries have sought to develop their own models.

The UAE's aforementioned Jais LLM is a prime example of a country taking an indigenous development approach to sovereign AI models. Beyond Jais, the UAE's Technology Innovation Institute's (TII) open-weight Falcon models are also products of this approach, through which developers create a custom model architecture and use specific datasets to train the model.[28] Creating models from scratch, as the UAE has done, is often preferable for governments, as they have better insight into the whole LLM training process.

Similarly, the developers of Japan's Fugaku-LLM trained this open-weight model from scratch to better integrate the Japanese language; 60 percent of the training data used was in Japanese, which developers argue allows the model to learn nuances of the Japanese language and cultural norms.[29]

But training a model from scratch is also costly. Doing so is compute, energy, and talent intensive, and many countries lack the resources or expertise to compete with leading AI companies at the frontier of model development. This has led many countries to turn to an emerging class of open-weight models, lowering barriers to entry and allowing governments, firms, and researchers to adapt advanced capabilities to local needs without bearing the full costs of frontier training. Using this method, countries can integrate local data in the model after its initial training, allowing them to leverage the AI development of leading labs while tailoring surface-level behaviors, such as language, tone, and domain-specific outputs to local needs.

For example, Taiwan's Trustworthy AI Dialogue Engine (TAIDE) is built on Meta's Llama models and Google's Gemma models, which are both open-weight. Part of TAIDE's impetus is countering anti-Taiwan biases in Chinese- LLMs, as well as broader security concerns.[30] By fine-tuning Llama 3.1 with traditional Chinese characters and Taiwanese news, law, and literature, TAIDE is intended to better reflect the island's politics and culture while reducing security risks.[31]

With the emergence of leading open-weight LLMs developed by Chinese AI companies such as DeepSeek, Alibaba, and MoonshotAI, many developing countries are increasingly using such models as the base for their sovereign AI efforts at the model layer. Open-weight models are attractive as they lower the barrier for lower-resourced countries to adopt and design AI products.[32] For example, EqualyzAI, a Nigeria-based AI company, began adopting Chinese open-weight models because they "offer flexibility, lower cost, and the potential for local data sovereignty."[33] Meanwhile, U.S.-designed proprietary models are less attractive due to their comparatively high costs, licensing restrictions, and lack of customization to local languages, according to the founder of EqualyzAI.[34] Interestingly, some U.S. companies are using Chinese open-weight models for

allows modification of the models, however. Most Chinese models are open-weight rather than open-source.

business use cases. Airbnb, for example, is using Alibaba's Qwen because it is "fast and cheap," according to the CEO.[35]

It is worth noting that open-weight models vary in their licensing terms and may carry implications for sovereign AI efforts. Meta's Llama models are released under the Llama Community License Agreement,[vi] which imposes restrictions on entities with more than 700 million monthly active users and also prohibits use in military, warfare, and espionage use cases.[36] DeepSeek's model licensing terms also include use-based restrictions on harmful activities, while the code is released under the permissive MIT license.[37]

Moreover, open-source models often suffer from weaker, inconsistent, or easily removable safety guardrails.[38] This shifts the responsibility for hardening, monitoring, and safely adopting open-weight models onto deploying organizations, and the performance and robustness of these LLMs can vary substantially.

Ultimately, the model layer has become a critical aspect of countries' AI sovereignty efforts. Whether through costly indigenous development or fine-tuning of open-weight models, many countries are prioritizing the buildout of domestically aligned models over the use of foreign-trained models. The growing reliance on open-weight models—especially for developing countries—underscores the opportunity for leading AI developers to drive their model adoption. The proliferation of these models, primarily developed by Chinese companies, are providing a feasible path for those with fewer resources to achieve some level of AI sovereignty. For most nations, as with compute and data, the most realistic goal will likely be finding a balance that mitigates dependence on any single foreign country.

CONCLUSION

Sovereign AI is often framed as an all-or-nothing proposition, where the goal is full national control over chips, data, and models. In practice, that standard is unattainable for most countries. Compute remains the principal constraint, with advanced chips concentrated in a handful of supply chains and likely to remain so for the foreseeable future. Model development from scratch is similarly out of reach for most nations given the cost, talent requirements, and pace of frontier research. Data, by contrast, is where sovereignty is most achievable, and where governments are already exerting control.

Taken together, these trends suggest that the future of sovereign AI will necessarily be hybrid. Most countries will rely on foreign compute, even if they build infrastructure locally, adapt rather than invent foundation models, and differentiate through localized data, governance, and deployment choices. For many countries, the path to advanced AI capabilities runs through fine-tuning and adapting open-weight models rather than building national champions from scratch. Sovereignty, in this sense, is becoming less about independence and more about resilience; countries will seek to reduce their single-country dependencies and preserve room to make political and strategic choices. This will be easier said than done, as U.S. dominance at the chip layer will remain an important geopolitical consideration for all countries seeking to develop sovereign AI

iv Meta has announced that the company will allow use of Llama for U.S. national security purposes by the U.S. government and the companies supporting their work. See M. S. Smith, "Meta Opens Its AI Models for the (U.S.) Military: But will the second Trump administration see AI as a friend or foe?" IEEE Spectrum, November 17, 2024, https://spectrum.ieee.org/ai-used-by-military.

programs.

This framing also clarifies the stakes for Western governments and companies. Expanding the availability and viability of open models is more than a strictly commercial or technical choice. As of now, despite that they lag SOTA proprietary models in terms of capability, Chinese open-weight models have become popular across a wide range of countries due to their affordability and "good enough" performance. Should Western AI companies become more competitive in the open-weight model space, it would provide smaller countries with additional alternatives to Chinese models. At the same time, the business models underpinning open-weight AI remain unclear, raising questions about which companies and countries will be able to sustain these systems over time.

Ultimately, the central question for sovereign AI is not whether countries can control every layer of the stack, but where choices exist. Today, there are few at the chip layer, many at the data layer, and a limited but growing set at the model layer. How governments navigate those tradeoffs will define what sovereignty in AI will mean going forward.

Endnotes

1 K. Leswing, "Nvidia Dominates the AI Chip Market, but There's Rising Competition," CNBC, June 2, 2024, https://www.cnbc.com/2024/06/02/nvidia-dominates-the-ai-chip-market-but-theres-rising-competition-.html. [https://perma.cc/9KMB-4RK2]

2 "AMD and OpenAI Announce Strategic Partnership to Deploy 6 Gigawatts of AMD GPUs," OpenAI, October 6, 2025, https://openai.com/index/openai-amd-strategic-partnership/. [https://perma.cc/X2KF-UZRH]; T. Mickle and A. Satariano, "Racing to Catch up with Nvidia, AMD Signs Chips-for Stock Deal with Meta," New York Times, February 24, 2026, https://www.nytimes.com/2026/02/24/business/meta-amd-chips-ai.html

3 S. Shibu, "AMD's CEO Claims Their New Chips 'Match' Nvidia's at a Lower Price, and Even Sam Altman Is Excited: 'An Amazing Thing'," Entrepreneur, June 13, 2025, https://www.entrepreneur.com/business-news/amd-ceo-claims-new-ai-chips-outperform-nvidias/493272.

4 J. Feldgoise and H. Dohmen, "Pushing the Limits: Huawei's AI Chip Tests U.S. Export Controls," Center for Security and Emerging Technology, June 17, 2024, https://cset.georgetown.edu/publication/pushing-the-limits-huaweis-ai-chip-tests-u-s-export-controls/; C. Pan and B. Goh, "Baidu Chip-Design Unit Kunlunxin Wins over $139 Million Order from China Mobile Suppliers," Reuters, August 22, 2025, https://www.reuters.com/technology/baidu-chip-design-unit-kunlunxin-wins-over-139-million-orders-china-mobile-2025-08-22/; A. Kharpal, "China Seeks a Homegrown Alternative to Nvidia—These Are Some of the Companies to Watch," CNBC, September 17, 2024, https://www.cnbc.com/2024/09/17/chinese-companies-aiming-to-compete-with-nvidia-on-ai-chips.html.

5 I. King, "China's Huawei, SMIC Make Progress With Chips, Report Finds," Bloomberg, December 11, 2025, https://www.bloomberg.com/news/articles/2025-12-11/china-s-huawei-and-smic-make-progress-with-chips-report-finds. [https://archive.ph/P23lX]

6 "Exclusive: China Mandates 50% Domestic Equipment Rule for Chipmakers, Sources Say," Reuters, December 31, 2025, https://www.reuters.com/world/china/china-mandates-50-domestic-equipment-rule-chipmakers-sources-say-2025-12-30/.

7 "Cloud Market Growth Rate Rises Again in Q3; Biggest Ever Sequential Increase," Synergy Research Group, October 31, 2025, https://www.srgresearch.com/articles/cloud-market-growth-rate-rises-again-in-q3-biggest-ever-sequential-increase.

8 European Cloud Providers' Local Market Share Now Holds Steady at 15%." Synergy Research Group, July 24, 2025. https://www.srgresearch.com/articles/european-cloud-providers-local-market-share-now-holds-steady-at-15.

12 "Global Tech Alliance Launches Stargate UAE," G42, May 22, 2025, https://www.g42.ai/resources/news/global-tech-alliance-launches-stargate-uae. [https://perma.cc/T6BK-F8UW]

10 M. Hawkins, "US Approval for Advanced Chips Boosts UAE, Saudi AI Ambitions," Bloomberg, November 19, 2025, https://www.bloomberg.com/news/articles/2025-11-19/us-reaches-ai-chip-sale-agreement-with-g42-in-win-for-uae-firm.

11 N. Manger and V. Mishra, "Can Europe Build Digital Sovereignty While Safeguarding Its Rights Legacy?" Tech Policy Press, December 5, 2025, https://www.techpolicy.press/can-europe-build-digital-sovereignty-while-safeguarding-its-rights-legacy/.

12 "France and Germany Join Forces with Mistral AI and SAP SE to Launch a Sovereign AI for Public Administration," Bundesministerium für Digitales und Staatsmodernisierung, November 18, 2025, https://bmds.bund.de/aktuelles/pressemitteilungen/detail/france-and-germany-join-forces-with-mistral-ai-and-sap-se-to-launch-a-sovereign-ai-for-public-administration. [https://perma.cc/D3WS-48MQ].

13 "France and Germany Join Forces with Mistral AI and SAP SE to Launch a Sovereign AI for Public Administration," Bundesministerium für Digitales und Staatsmodernisierung, Novem-

ber 18, 2025, https://bmds.bund.de/aktuelles/pressemitteilungen/detail/france-and-germany-join-forces-with-mistral-ai-and-sap-se-to-launch-a-sovereign-ai-for-public-administration. [https://perma.cc/D3WS-48MQ].

14 M. Demertzis, A. Fiorito, and K. Pantisas, Strategic Autonomy and European Competitiveness: Security Now Comes First, Directorate-General for Economy, Transformation and Industry, Economic Governance and EMU Security Unit (EGOV), December 2025, https://www.europarl.europa.eu/RegData/etudes/STUD/2025/764371/ECTI_STU(2025)764371_EN.pdf. [https://perma.cc/9AAL-LXNP].

15 K. McBride et al., "Sovereignty, Security, Scale: A UK Strategy for AI Infrastructure," Tony Blair Institute for Global Change, July 29, 2025, https://institute.global/insights/tech-and-digitalisation/sovereignty-security-scale-a-uk-strategy-for-ai-infrastructure. [https://perma.cc/XCE8-RXSX].

16 Government of India, Ministry of Electronics and Information Technology, "Development of Artificial Intelligence Hub," accessed February 25, 2026, https://sansad.in/getFile/loksabhaquestions/annex/184/AU3062_htWeSf.pdf?source=pqals. [https://perma.cc/PNX9-4YFS]

17 R. Hoeffnagel, "France is building its own 'Cloud de Confiance' for government agencies and critical infrastructure players," International Data Center Authority (IDCA), June 22, 2022, https://www.idc-a.org/news/France-is-building-its-own-Cloud-de-Confiance-for-government-agencies-and-critical-infrastructure-pl. [https://perma.cc/A8EU-QZUC]

18 R. Hoeffnagel, "France is building its own 'Cloud de Confiance' for government agencies and critical infrastructure players," International Data Center Authority (IDCA), June 22, 2022, https://www.idc-a.org/news/France-is-building-its-own-Cloud-de-Confiance-for-government-agencies-and-critical-infrastructure-pl. [https://perma.cc/A8EU-QZUC]

19 "Meet Jais, The World's Most Advanced Arabic LLM Open Sourced by G42's Inception," G42, August 30, 2023, https://www.g42.ai/resources/news/meet-jais-worlds-most-advanced-arabic-llm-open-sourced-g42s-inception. [https://perma.cc/473J-GNEH].

20 "Meet Jais, The World's Most Advanced Arabic LLM Open Sourced by G42's Inception," G42, August 30, 2023, https://www.g42.ai/resources/news/meet-jais-worlds-most-advanced-arabic-llm-open-sourced-g42s-inception. [https://perma.cc/473J-GNEH].

21 Z. Li et al., "Quantifying Multilingual Performance of Large Language Models Across Languages," arXiv, (2024), https://arxiv.org/html/2404.11553v2.

22 A. Look, "Arabic AI Could Help Open Doors for Other Languages," CNN, October 4, 2023, https://www.cnn.com/2023/10/04/middleeast/jais-arabic-ai-open-doors-spc-intl/index.html.

23 "Why SEA-LION," South-East Asia Large Language Models, GitHub, accessed February 25, 2026, https://github.com/aisingapore/sealion/blob/main/overview/why_sea-lion.md. [https://perma.cc/4G2L-C25X].

24 "Why SEA-LION," South-East Asia Large Language Models, GitHub, accessed February 25, 2026, https://github.com/aisingapore/sealion/blob/main/overview/why_sea-lion.md. [https://perma.cc/4G2L-C25X]

25 "Why SEA-LION," South-East Asia Large Language Models, GitHub, accessed February 25, 2026, https://github.com/aisingapore/sealion/blob/main/overview/why_sea-lion.md. [https://perma.cc/4G2L-C25X].

26 D. Maksymenko and O. Turuta, "Tokenization efficiency of current foundational large language models for the Ukrainian language," Frontiers in Artificial Intelligence, National Library of Medicine, August 13, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC12380774/.

27 "Introducing Meta Llama 3: The most capable openly available LLM to date," Meta, April 18, 2024, https://ai.meta.com/blog/meta-llama-3/. [https://perma.cc/3W5Z-VZ6Z].

28 "Introducing TII's Falcon H1R 7B & H1 Arabic," Technology Innovation Institute, January

5, 2026, https://falconllm.tii.ae/.
29 "Japan Team Develops AI Foundation with Fugaku Supercomputer," The Japan News, May 11, 2024, https://japannews.yomiuri.co.jp/science-nature/technology/20240511-185415/.
30 "About TAIDE," National Institute of Applied Research, accessed February 25, 2026, https://en.taide.tw/.
31 "About TAIDE," National Institute of Applied Research, accessed February 25, 2026, https://en.taide.tw/.
32 S. Rai, L. Prinsloo, and H. Nyambura, "DeepSeek's Push Into Africa Reveals China's AI Power Grab," Bloomberg, October 22, 2025, https://www.bloomberg.com/news/features/2025-10-22/china-s-deepseek-pushes-into-africa-making-ai-accessible-to-millions.
33 S. Rai, L. Prinsloo, and H. Nyambura, "DeepSeek's Push Into Africa Reveals China's AI Power Grab," Bloomberg, October 22, 2025, https://www.bloomberg.com/news/features/2025-10-22/china-s-deepseek-pushes-into-africa-making-ai-accessible-to-millions.
34 S. Rai, L. Prinsloo, and H. Nyambura, "DeepSeek's Push Into Africa Reveals China's AI Power Grab," Bloomberg, October 22, 2025, https://www.bloomberg.com/news/features/2025-10-22/china-s-deepseek-pushes-into-africa-making-ai-accessible-to-millions.
35 J. Power, "China's AI is quietly making big inroads in Silicon Valley," Al Jazeera, November 13, 2025, https://www.aljazeera.com/economy/2025/11/13/chinas-ai-is-quietly-making-big-inroads-in-silicon-valley.
36 "Llama 4 Acceptable Use Policy," Meta, accessed February 25, 2026, https://www.llama.com/llama4/use-policy/. [https://perma.cc/4R2V-KDW6]; "Llama 4 Community License Agreement," Meta, accessed February 25, 2026, https://www.llama.com/llama4/license/. [https://perma.cc/ERT3-6F9T].
37 "MIT License," DeepSeek-R1, DeepSeek AI, GitHub, accessed February 25, 2026, https://github.com/deepseek-ai/DeepSeek-R1/blob/main/LICENSE.
38 A. Chang and N. Conley, "Death by a Thousand Prompts: Open Model Vulnerability Analysis," Cisco Blog, November 5, 2025, https://blogs.cisco.com/ai/open-model-vulnerability-analysis.

Chapter 4: National Security and Emerging Technologies: The Growing Centrality of Public-Private Partnerships

By Jeffery Payne[i]

KEY TAKEAWAYS:

- Public-private partnerships facilitate fluency among respective stakeholders in technological innovation.
- Cutting-edge technological firms increasingly rely upon the infrastructural and economic institutions that states control.
- The progression of public-private partnerships is often driven by national security considerations with militaries serving as a key facilitator.
- Technological innovation is informing geopolitical tensions and competition between states over technology will increasingly push firms towards alignment with state mindsets or else risk losing a stake in shaping technology policy.

INTRODUCTION

The era of "move fast and break things," driven by free-wheeling Silicon Valley entrepreneurs and innovators, may be reaching its end. For much of the past two decades, commercially driven technology firms set the pace of innovation. Governments found themselves playing a secondary role. Yet, that structure no longer fits the technologies now shaping national security. Advanced semiconductors, artificial intelligence, autonomous systems, and the infrastructure that sustains them are increasingly central to geopolitical competition. Under these conditions, public-private partnerships are structural requirements for integrating private innovation with national interests through burden sharing and increased fluency regarding the processes between public and private institutions.

The success of private technology firms has led to an environment where future market growth depends on assets only governments can provide or coordinate effectively. State intervention is necessary to provide secure semiconductor supply chains, reliable energy infrastructure, credible

i Jeffrey Payne is an Assistant Professor at the Near East South Asia (NESA) Center for Strategic Studies. The views in this paper are the author's alone and do not reflect the official position or policy of the NESA Center, the Department of War, or the United States government.

export control regimes, and trusted international partnerships needed to sustain technological advantages. At the same time, governments increasingly rely on private firms for technical expertise, production capacity, and the speed necessary to integrate emerging technologies into military and security institutions. Neither side can achieve its objectives in isolation. Both sides also need to interface with greater precision and speed.

The challenge ahead is that existing mechanisms for collaboration were designed for an earlier era. Government acquisition and research processes remain slow and risk-averse, while many technology firms are structured to operate at commercial speed and global scale. These mismatches impose growing costs as emerging technologies developed by the private sector become increasingly central to military power and economic competitiveness. Delays in adaptation now will have strategic consequences in the decades to come as technology has become a focus of geopolitical tension.

GOVERNMENT STAKES IN TECHNOLOGY

Governments, without exception, understand the importance of emerging technologies for ensuring national security. Unmanned systems have already altered military tactics and by extension, strategy.[1] Artificial intelligence, specifically generative tools, informs everything from law enforcement to intelligence analysis.[2] To lack access to the benefits provided by emerging technology is to be at a disadvantage. Yet, unlike past periods of innovation, states, with a few exceptions, are not driving today's technological advances. It is the private firms of Silicon Valley, Shenzhen, Tokyo, and other urban concentration points that generate progress. To adopt these critical technologies into national security institutions requires partnership with private firms and the associated research and development pipelines.[3]

How public-private partnerships will evolve remains unclear. These partnerships, particularly within and among liberal democracies, are not easily accomplished nor sustained. Contemporary governments adopt national security tools through complex and laborious processes for private sector firms already within government systems. For new actors seeking to work with government institutions, the hurdles can often seem impossible. Hesitancy, even outright apprehension, exists for the adoption of technologies even when declared as government policies. This reflects larger apprehensions regarding emerging technologies. In the case of artificial intelligence, research shows that despite the explosion of use of artificial intelligence models, "…trust remains precarious – only 62% of business leaders and 52% of employees believe AI is deployed responsibly within their organizations."[4]

Many leading technology firms producing applications now deemed essential for national security are not in alignment with the regulatory regimes that define government acquisition. In fact, such firms often consider government processes as anathema to their model. For these partnerships to work, firms must adapt to varied stakeholders that constitute national security governance while governments must reform research and development, along with acquisition processes, to gain flexibility and efficiency. Yet, looming over these complications is that leaders and employees of institutions do not enjoy adequate time to acclimate to technologies that in turn facilitate a lack of trust or even a backlash against their use.[5] Failure to facilitate adoption and acclimation creates an environment where the innovator and the adopter do not create a partnership. Both sides of the partnership must adapt policy and timelines or risk creating delays that create a backlash against

technology, increased costs upon private sector providers, and bottlenecks in adoption. Trust in partnerships is built by stakeholders through gaining fluency in the operations beyond their own institutions.

Emerging technologies also rely upon a diverse web of intricate infrastructural needs. The complexity is beyond any actor's capabilities if the aim is to keep pushing forward into the frontier of innovation. Semiconductors that facilitate advanced computation are manufactured as the result of a global supply chain with immense vulnerabilities.[6] The data centers that will drive future artificial intelligence progress place heavy loads upon power grids.[7] Put another way, the industrial scale and infrastructural depth that helped determine 20th century economic influence will continue to shape the success of the firms producing emerging technologies. The global supply chain that must navigate a multinational set of laws and regulations will continue to shape the rules of the technology game.

Given the potential impact of these technologies, geopolitical processes and tensions will loom over innovation.[8] Geopolitics, industrial bases, energy accessibility are each firmly within the domain of states. The accomplishments of private sector innovation are immense, and national security institutions' appetite for technological adoption is clear. What is an underexamined variable is to the degree that states will facilitate the expanded infrastructure that private sector innovators will require. This provision will steer future developments in military and security operations that few outside of state agencies sufficiently recognize.

FROM NOVEL TO ESSENTIAL

The list of critical and emerging technologies (CET) represents a ranking of risks and opportunities for national security institutions. This list in and of itself reflects the complex adaptations that governments must navigate to partner with the private sector. As Kallenborn and Willis (2025) explored in exploring why there is no accepted agreement on what constitutes global critical infrastructure, there is similarly no globally accepted definition of what is meant by CET.[9]

These technologies, such as artificial intelligence, advanced computation, and autonomous and uncrewed systems (UxS), already are shaping security assessments that in turn determine operational direction for militaries and law enforcement.[10] Take unmanned aerial vehicles (UAVs) as an example. The importance of UAVs for a variety of operations became apparent following the 2023 Nagorno-Karabakh conflict and was made even more abundantly clear with how the Russo-Ukrainian war became an unfortunate demonstration of the lethality of drones.[11] The current civil war within Sudan has also become a magnet for drone usage in conflict. Feldstein, in a recent Foreign Policy piece commenting on recent conflicts, explains:

"…data paints a grim picture. Just five years ago, in 2020, analysts recorded 6000 drone incidents worldwide resulting in about 11,300 fatalities. Four years later, the numbers had shot up dramatically. 2024 witnessed a fourfold increase, with nearly 51,000 recorded drone events leading to over 39,000 deaths."[12]

UAVs emerged as a security tool through the traditional research and development pipeline of the United States government, but as the commercial uses of UAVs became apparent, the private sector engaged in drone development and applied them across several commercial sectors, such as

oil and gas support or remote communications.[13] Tailor-made technological tools for governments function under specific operational parameters. They, almost without exception, are slower to build and harder to maintain.

In the case of unmanned systems, commercial-off-the-shelf (COTS) options arose to serve a variety of functions. COTS variants are not routinely designed to deliver niche military capabilities, but they have proven able to deliver a wide range of military capabilities at a lower cost, or as Chavez and Swed state, "…they are easier to adapt with off-the-shelf or jury-rigged, after-market capabilities."[14] In short, commercial unmanned systems are "good enough." The commercial mentality is gaining adherents within the United States government and several of its Indo-Pacific allies, as traditional acquisition processes cannot match the scale that future projections of conflict predict will be necessary.[15]

Niche capabilities that amplified specialization within militaries and law enforcement organizations following the Cold War are being undermined by lower-cost technologies, such as commercial drones, designed specifically to be attritional. A complex and costly unmanned unit's value for operations can be overwhelmed by COTS variants that are easy to acquire and distribute, while not creating the same vulnerabilities when lost in an operation.[16] The depth of an army's resources, armaments, and reach may not guarantee victory, but it certainly increases the probability of success on the battlefield in the age of algorithms.[17]

The national security implications of CET go far beyond unmanned systems, but the technical processes associated with CET share common evolutions. Plucky startups, such as Capella in the commercial space sector, were able to develop novel systems that proved to have commercial value.[18] In turn, firms that found footing in the commercial sector scaled up and further developed the capabilities of the technology. Eventually, states recognized the utility of these technologies and thereafter sought to acquire, gain familiarity, and integrate them into their operations. At present, national security strategies, concepts, and methodologies position CET as no longer novel, but essential.[19]

SCALING UP COMMERCE AND GOVERNMENT CATCH-UP

CET achievement and adoption is a policy aim of governance that includes a horde of commodities that make up the composite of technologies. To be successful, as Buchanan and Collins posited in a recent Foreign Affairs piece, requires "…a grand bargain between the tech industry and the government."[20] All the technological sectors that have garnered so much attention, such as AI, advanced computation, aerospace, autonomy, and biotechnology, rely on an increasingly precious commodity - semiconductors. Virtually all advanced law enforcement and military equipment likewise cannot function without the reliable provision of semiconductors. As an industry, the United States and European markets remain dominant in the research and development of semiconductors, but semiconductor manufacturing concentration points exist in Taiwan, South Korea, and China.[21] Most Western technological firms under the CET umbrella are built to maximize the computational reach and optimal outcomes of their research and development. In AI, the aim of American and other Western firms is not simply to refine general AI capabilities to serve a practical purpose, but to strive towards the building of algorithmic processes that answer advanced inquiries.[22] Put another way, the West is focused over the next horizon while other technological powers seek to maximize the here and now. Increasingly, the scientists developing AI posit that

Western policymakers need to consider CET through a mundane lens and focus on the logical adaptations to policy that result.[23]

Schmidt and Bajraktari detailed the specific bizarre circumstances the U.S. finds itself in when it comes to tech. U.S. research institutions and academia are the most substantive in the world and commercial tech giants remain concentrated in the U.S. Yet, little of the technological supply chain is manufactured in the U.S., government institutions have remained slow to adapt and adopt CET, and the U.S. workforce remains, by and large, untrained to onshore manufacturing tied to technology.[24] It is China, the U.S.' main geopolitical competitor, that has built a whole-of-society concentration on CET.[25] China may not yet dominate innovations within CET, but it is catching up because it has built the infrastructural, economic, and political mechanisms to dominate key aspects of the supply chain.[26] Most technological firms lean towards the Western-led system for a variety of purposes, but the commercial realities many of these cutting-edge firms face does not translate into an overabundance of patience. Intensified effort by Western countries to bridge the gaps in the manufacturing and supply chains is ongoing, with the Trump Administration keeping a constant focus on bringing key aspects of the supply chain to the United States.[27]

Access to CET is a top priority for national security institutions and while not moving quickly enough, militaries, law enforcement organizations, and other national security offices are adapting the quickest among government services. In many militaries, for instance, investments are occurring in training a generation of key officers and civilian employees to serve as translators between policymakers and technology firms. Initial bureaucratic transitions represented by the Defense Innovation Unit in the United States or Japan's Acquisition, Technology & Logistics Agency are evolving into overt strategies signaling the importance of pace.[28] The process is not close to completion, and such reforms are not yet proven successful despite scaling success, but efforts like these have led defense and national security ministries into a variety of partnerships with the technological ecosystem. CET exposes deficiencies throughout many states that have long been ignored. The infrastructural networks and industrial bases are not sufficient. Yet many Western states and their partners are not destined to always face such deficiencies. Overcoming the problems will require comprehensive policy revisions, but initial, important steps to address the challenges are being taken at the working level, such as changes to procurement-related bureaucratic institutions and intensified partnerships with technology accelerators.[29] The work on gaining fluency for how to better integrate private sector capabilities is at the base of public-private partnerships.

On the private sector side of the partnership, firms must recognize that they need to engage, not ignore, the geopolitical realities attached to CET or risk displacement as a stakeholder in policy discussions.[30] This requires more than public relations acknowledgement that a transparent, not opaque, system governing technology is the preference. The realities of states that seek to undermine established rules and norms governing our globe, such as China and Russia, are not merely another market and that the regulatory systems of many states, particularly those in the West, may be laborious and slow, but they build networks that provide benefits to all stakeholders. Frontier software and AI firms may debate the benefits of open or closed coding models, but more transparent legal, financial, and regulatory systems prove more predictable and thus, more profitable compared to opaque structures.[31]

Tensions between the U.S. and China, as example, must be treated seriously. Semiconductor giant Nvidia is continuously caught up in tensions between the U.S. and China.[32] Such tensions show no

signs of disappearing and navigating them can entail substantial trade-offs. Yet these tensions can also be an opportunity for firms. Exploiting tensions to gain leverage to obtain the infrastructural and industrial base provisions needed to expand their market scale is a logical use of the current environment.

Take for instance the recent deals agreed to by the U.S. and several Gulf Cooperation Council (GCC) partners to enhance technological cooperation, including expanding data centers in the Arabian Peninsula and exploring critical mineral mining and refining.[33] Despite concern over how substantial Middle Eastern regimes will protect foreign intellectual property and whether these states conform to the regulatory regimes that are standard elsewhere, the attractiveness of these energy-rich countries for technology firms is obvious.[34] The Trump administration seeks to expand the supply chain for technology inside the United States to the largest degree possible, but in the meantime the calculus revealed that facilitating expanded investment in Saudi Arabia and the United Arab Emirates, states heavily reliant on U.S. security umbrellas, would expand the technology supply chain to further propel Washington's policy objectives.[35]

BREAKING SILOS AND ACCEPTING RISK

The deals that the U.S. and its partners are making to diversify technology supply chains are not without risk. Such deals are, however, a reflection of the realities of the current state of national security when coupled with CET. The technologies that shape military and national security operations are racing ahead to be the first to breakthrough on innovation. There is no guarantee that the U.S. or an allied country will win this race or even if private firms can achieve what they aspire to accomplish. Regardless of the risk, the sheer scale of lost opportunity and new threats requires states to bend, if not break, traditional silos that have governed research and development, acquisition, and national security strategic thinking.

The worst critics of the technological powers of the West are routinely themselves. The tendency to underestimate their own capabilities while overestimating their rivals is acute. China is accomplishing wonders in modernizing its economy and is a technological powerhouse.[36] China may even one day outpace the collective innovation of North America, Europe, South Korea, India, and Japan, but as the United States and other Western technological powers are learning, long-term success is dependent on the weaving of many institutions. China's industrial policy is driven by state leaders who facilitate targeted investments across the CET spectrum.[37] That approach has met with success. Yet other key institutions within China retain vulnerabilities. Overall economic output is stalling due to continuous regulation swings, demographic decline, stalls in living standards, and an opaque banking system.[38] Compared to China, Western bureaucracies alter course too slowly but once turned into alignment with private sector actors the transparency of institutions lends to wider economic output.

Looking more specifically at the U.S. military, the redirection of energy within the bureaucracy is already taking place. Initial experiments at adopting COTS technology, such as Project Maven and Task Force 59, have led to the Department of War partnering with technology accelerators, universities and research laboratories, and routinely publicizing and supporting startups.[39] Major reforms, to acquisition, spending, adoption, and deployment, are increasing in scale and scope. Meetings, conferences, and exercises routinely feature commercial technologies and integrate private firms as stakeholders. All of this has been facilitated by firms who have routinely reached out

to governments along with a growing contingent of government officials who know how essential CET will be for the future security of the state. It has been a slow start, but momentum is rising. A similar pattern can be observed in Japan, the United Kingdom, France, Canada, South Korea, and a collection of other states.[ii]

Worries about the slow pace of government and the unrealistic expectations of private firms should be contextualized in comparison to the evolution of these public-private partnerships. The partnerships serve as translators between different institutional patterns. The conversation today is how militaries can facilitate a community of private partners for aerospace innovation or supply a horde of unmanned systems or even how to harden existing COTS technology to face the stress of combat. The whole-of-government approaches for enhancing the industrial base, infrastructural network, and supply chain to facilitate partnerships with the private sector are the norm, not the exception. Precise policy guidance and specific plans for building a network of trusted technology providers is the discussion point for London, Washington, Tokyo, or Canberra, not whether such a network should be built.[40] Progress is being made as the ties that facilitate progress are in motion.

The work now is tied to determining what are the next steps to climb. The need for proof of concept is not necessary. What is necessary is that the collaborative environment facilitated by these partnerships expand. For national security institutions, traditional patterns must be reformed, and greater risk accepted as a cost of the partnership. Facilitation of innovators, both large and small, must become common to navigate the assortment of challenges looming on the horizon. For the private sector, planning for government stakeholders and the citizens they answer to is essential and should not be bypassed. Furthermore, private firms must acknowledge what they already know – the infrastructure that will serve as the bedrock for advanced innovation requires government support. Public-private partnerships, as they relate to national security, must gain further fluency in how to work with another. Once fluency is commonplace, the competition over technology and the deliverables provided can gain greater speed and scale.

ii For further exploration of this trend, please refer to Tsujiguchi, Makota and Tanaka, Moyuru, "Exploring Promises: Opportunities for Defense-Tech Start-Ups in U.S.-Japan Alliance," CSIS, 21 January 2025, https://www.csis.org/analysis/exploring-promising-opportunities-defense-tech-start-ups-us-japan-alliance, "New Strategic Partnerships to Unlock Billions and Boost Military AI and Innovation," United Kingdom Ministry of Defence, 18 September 2025, https://www.gov.uk/government/news/new-strategic-partnership-to-unlock-billions-and-boost-military-ai-and-innovation, "The great American shipyard reset: How Hanwha and the U.S. are tooling up," Hanwha, 1 December 2025, https://www.hanwha.com/newsroom/news/feature-stories/the-great-american-shipyard-reset-how-hanwha-and-the-us-are-tooling-up.do.

Endnotes

1 C. J. Chivers, "In Ukraine, a New Arsenal of Killer A.I. Drones Is Being Born," New York Times, January 5, 2026, https://www.nytimes.com/2025/12/31/magazine/ukraine-ai-drones-war-russia.html.

2 E. S. Probasco et al., "AI for Military Decision-Making: Harnessing the Advantages and Avoiding the Risks," Center for Security and Emerging Technology, April 2025, https://cset.georgetown.edu/wp-content/uploads/CSET-AI-for-Military-Decision-Making.pdf.

3 C. T. Lopez, "War Department Asks Industry to Make More Than 300K Drones, Quickly, Cheaply," Pentagon News, December 2, 2025, https://www.war.gov/News/News-Stories/Article/Article/4346822/war-department-asks-industry-to-make-more-than-300k-drones-quickly-cheaply/.

4 R. Sharma, "The $4.8 Trillion AI Trust Crisis: Why Public-Private Partnerships Are Key for Equitable Innovation," World Economic Forum, September 29, 2025, https://www.weforum.org/stories/2025/09/ai-trust-crisis-public-private-partnerships/.

5 J. Stilgoe, "What Does It Mean to Trust a Technology?," Science 382, no. 6676 (December 14, 2023), https://www.science.org/doi/10.1126/science.adm9782.

6 B. Martin et al., "Supply Chain Interdependence and Geopolitical Vulnerability: The Case of Taiwan and High-End Semiconductors," Rand Corporation Research Report, March 14, 2023, https://apps.dtic.mil/sti/trecms/pdf/AD1195673.pdf.

7 J. O'Donnell and C. Crownhart, "We Did the Math on AI's Energy Footprint. Here's the Story You Haven't Heard," MIT Technology Review, May 20, 2025, https://www.technologyreview.com/2025/05/20/1116327/ai-energy-usage-climate-footprint-big-tech/.

8 T. Feakin, "Navigating the New Geopolitics of Tech," Harvard Business Review, November 11, 2024, https://hbr.org/2024/11/navigating-the-new-geopolitics-of-tech.

9 Z. Kallenborn and H. H. Willis, "Globally Critical Infrastructure: The Unique Risks and Challenges," Risk Analysis 45, no. 12 (December 2025), https://onlinelibrary.wiley.com/doi/10.1111/risa.70147.

10 C. Hinote and M. Ryan, "Empowering the Edge: Uncrewed Systems and the Transformation of U.S. Warfighting Capacity," Special Competitive Studies Project, February 2024, https://www.scsp.ai/wp-content/uploads/2024/02/SCSP-Drone-Paper-Hinote-Ryan.pdf.

11 D. A. Deptula and C. J. Bowie, "The Significance of Air Superiority: The Ukraine-Russia War," Mitchell Institute Policy Paper 50 (July 2024), https://www.mitchellaerospacepower.org/app/uploads/2024/07/Ukraine_Control_of_the_-Air_Policy_Paper_50.pdf.

12 Feldstein, Steven, "Good Enough Drones Have Become Geopolitical Chips," Foreign Policy, 17 November 2025. https://foreignpolicy.com/2025/11/17/drones-geopolitical-chips-sudan-warfare-russia-ukraine/

13 K. Miller et al., "The U.S. Aerial Drone Market," Center for Security and Emerging Technology, November 2025, https://cset.georgetown.edu/wp-content/uploads/CSET-The-U.S.-Aerial-Drone-Market.pdf.

14 K. Chavez and O. Swed, "Small Drones for Big Militaries: The Way Ahead," The War Room, August 15, 2024, https://warroom.armywarcollege.edu/articles/small-drones/.

15 "War Department Announces Vendors Invited to Compete in Phase I of the Drone Dominance Program," United States Department of War, February 3, 2026, https://www.war.gov/News/Releases/Release/Article/4396462/war-department-announces-vendors-invited-to-compete-in-phase-i-of-the-drone-dom/.

16 J. Watling and N. Reynolds, "Tactical Developments During the Third Year of the Russo-Ukrainian War," Royal United Services Institute, February 2025, https://static.rusi.org/tactical-developments-third-year-russo-ukrainian-war-february-2205.pdf.

17 A. Neuberger, "How to Adapt in an Era of Algorithm Warfare," Foreign Policy, February 26, 2026, https://foreignpolicy.com/2026/02/26/algorithm-warfare-drone-commercial-tech-ukraine-russia/.
18 "About Us," Capella Space, February 2026, https://www.capellaspace.com/company/about.
19 "Winning the Race: America's AI Action Plan," The White House, July 2025, https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf; "Launching the Genesis Mission," The White House, November 24, 2025, https://www.whitehouse.gov/presidential-actions/2025/11/launching-the-genesis-mission/; "Integrated Innovation Strategy 2025," Council for Science, Technology and Innovation, the Cabinet Office of Japan, July 6, 2025, https://www8.cao.go.jp/cstp/tougosenryaku/togo2025_honbun_eiyaku.pdf.
20 B. Buchanan and T. Collins, "The AI Grand Bargain," Foreign Affairs 104, no. 6: 74.
21 G. Allison et al., "The Great Tech Rivalry: China vs the U.S.," Avoiding Great Power War Project, Cambridge, MA: Belfer Center for Science and International Affairs, Harvard University, 2021, https://www.belfercenter.org/sites/default/files/GreatTechRivalry_ChinavsUS_211207.pdf.
22 N. Maslej et al., "The AI Index 2025 Annual Report," AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, April 2025, https://doi.org/10.48550/arXiv.2504.07139.
23 A. Narayanan and S. Kapoor, "AI as Normal Technology," Knight First Amendment Institute, April 15, 2025, https://knightcolumbia.org/content/ai-as-normal-technology.
24 A. Narayanan and S. Kapoor, "AI as Normal Technology," Knight First Amendment Institute, April 15, 2025, https://knightcolumbia.org/content/ai-as-normal-technology.
25 "2025 Gaps Analysis Report," Special Competitive Studies Project, 2025, https://www.scsp.ai/reports/2025-gaps-analysis/.
26 J. Wong-Leung et al., "ASPI's Two-Decade Critical Technology Tracker: The Rewards of Long-Term Research Investment," Australian Strategic Policy Institute, August 28, 2024, https://www.aspi.org.au/report/aspis-two-decade-critical-technology-tracker/.
27 "Fact Sheet: Restoring American Semiconductor Manufacturing Leadership Through an Agreement on Trade & Investment with Taiwan," United States Department of Commerce, January 16, 2026, https://www.commerce.gov/news/fact-sheets/2026/01/fact-sheet-restoring-american-semiconductor-manufacturing-leadership.
28 "Accelerating America's Military AI Dominance," Office of the Secretary of War, January 9, 2026, https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF.
29 "Department of War Finalizes Realignment of the Defense Security Cooperation Agency and the Defense Technology Security Administration," Defense Security Cooperation Agency, United States Department of War, February 10, 2026, https://www.dsca.mil/Press-Media/Article-Display/Article/4402760/department-of-war-finalizes-realignment-of-the-defense-security-cooperation-age; "UC2 – Department of War University Consortium for Cybersecurity," College of Information and Cyberspace, National Defense University, 2026, https://cic.ndu.edu/UC2/.
30 L. Jamali, "AI boom boosts Nvidia despite 'geopolitical issues'," BBC, August 28, 2025, https://www.bbc.com/news/articles/c3wnj8611y7o.
31 L. C. Lee and J. Qian, "China's Private Sector Pivot," Foreign Affairs, February 20, 2026.
32 A. Ramkumar, R. Huang, and R. Whelan, "Nvidia's CEO Walks an AI Tightrope Between the U.S. and China," The Wall Street Journal, September 18, 2025, https://www.wsj.com/tech/ai/nvidia-ceo-jensen-huang-us-china-relationship-b7d438a7.
33 S. A. Cook, "For the Gulf States, Investment in AI is Partly About U.S. Protection," Foreign Policy, February 23, 2026, https://foreignpolicy.com/2026/02/23/gulf-states-investment-ai-ameri-

can-protection-qatar-uae-saudi/.
34 K. Alexander, "Digital Infrastructure, Strategic Power: The Gulf's Data Centre Boom," ORF Middle East, June 27, 2025, https://www.orfonline.org/research/digital-infrastructure-strategic-power-the-gulf-s-data-centre-boom.
35 G. Allen et al., "The United Arab Emirates' AI Ambitions: Key Implications for Maintaining U.S. AI Leadership," Center for Strategic and International Studies, January 24, 2025, https://www.csis.org/analysis/united-arab-emirates-ai-ambitions.
36 M. Arsenault, "Chinese Universities Surge in Global Rankings as U.S. Schools Slip," The New York Times, January 15, 2026, https://www.nytimes.com/2026/01/15/us/harvard-global-ranking-chinese-universities-trump-cuts.html.
37 K. Chan et al., "Full Stack: China's Evolving Industrial Policy for AI," RAND Expert Insights, June 26, 2025, https://www.rand.org/pubs/perspectives/PEA4012-1.html.
38 A. Brown, "China's new Five-Year Plan will embrace industry – and once again give consumers the cold shoulder," MERICS Comment, February 26, 2026, https://merics.org/en/comment/chinas-new-five-year-plan-will-embrace-industry-and-once-again-give-consumers-cold-shoulder.
39 A. Brown, "China's new Five-Year Plan will embrace industry – and once again give consumers the cold shoulder," MERICS Comment, February 26, 2026, https://merics.org/en/comment/chinas-new-five-year-plan-will-embrace-industry-and-once-again-give-consumers-cold-shoulder; A. Helou, "Commander: Navy's new Task Group 59.1 to usher unmanned systems into operational realm," Breaking Defense, January 19, 2024, https://breakingdefense.com/2024/01/commander-navys-new-task-group-59-1-to-usher-unmanned-systems-into-operational-realm/; "Tech Diplomacy Academy," Krach Institute for Tech Diplomacy at Purdue University, 2025, https://techdiplomacyacademy.org/
40 L. Pixa, "When Trust Becomes Strategy: Rethinking America's Innovation Posture," War on the Rocks, November 13, 2025, https://warontherocks.com/2025/11/when-trust-becomes-strategy/.