THE GEOPOLITICS OF AI CHIPS WILL DEFINE THE FUTURE OF AI

Rob Toews

THE following statement is utterly ludicrous. It is also true. The world's most important advanced technology is nearly all produced in a single facility.

What's more, that facility is located in one of the most geopolitically fraught areas on earth—an area in which many analysts believe that war is inevitable within the decade.

The future of artificial intelligence hangs in the balance.

The Taiwan Semiconductor Manufacturing Company (TSMC) makes all of the world's advanced AI chips. Most importantly, this means Nvidia's GPUs; it also includes the AI chips from Google, AMD, Amazon, Microsoft, Cerebras, SambaNova, Untether, and every other credible competitor.

Modern artificial intelligence simply would not be possible without these highly specialized chips. Neural networks—the basic algorithmic architecture that has powered every important AI breakthrough over the past decade, from AlphaGo and AlphaFold to Midjourney and ChatGPT—rely on these chips. None of the breathtaking advances in AI software currently taking the world by storm would be possible without this hardware.

Little surprise, then, that *Time* magazine described TSMC as "the world's most important company that you've probably never heard of."

Rob Toews *is a partner at Radical Ventures, a venture capital fund focused on artificial intelligence and deep tech. You may follow him on X @_RobToews. The author is also an AI columnist for* Forbes*, which published the original version of this essay in May 2023.*

Source: Guliver Image

U.S. President Biden visiting the site of the new TSMC factory in Arizona, the "most important company in the world"

Nvidia CEO Jensen Huang put it more colorfully, leaving little doubt about how important TSMC is to the future of AI: "Basically, there is air—and TSMC."

TSMC's chip fabrication facilities, or "fabs"—the buildings where chips are physically built—sit on the western coast of Taiwan, a mere 110 miles from mainland China.

Today, Taiwan and China are nearer to the brink of war than they have been in decades. With tensions escalating, China has begun carrying out military exercises around Taiwan of unprecedented scale and intensity. Many policymakers in Washington predict that China will invade Taiwan by 2027 or even 2025.

A China-Taiwan conflict would be devastating for many reasons. One underappreciated consequence is that it would paralyze the global AI ecosystem. Put simply, the entire field of artificial intelligence faces an astonishingly precarious single point of failure in Taiwan. Amid all the fervor around AI today, this fact is not widely enough appreciated. If you are working on or interested in AI, you need to be paying attention.

How did we get here, and what can we do about it?

A BRIEF OVERVIEW OF THE CHIP INDUSTRY

Semiconductors—also called integrated circuits or, colloquially, chips—are the most complex object that humanity knows how to mass produce. Making semiconductors requires the world's purest metals, the world's most expensive machinery (capable of building transistors less than 100 atoms thick), legions of highly specialized engineers, and unbelievable precision: a single speck of dust can ruin an entire chip production run, wasting millions of dollars.

The semiconductor supply chain is intricate and globalized. It is also, however, shockingly concentrated in certain areas. To give one example: 100 percent of the world's supply of extreme ultraviolet lithography machines—a complex piece of equipment required to build advanced chips—comes from a single company in the Netherlands called ASML.

As Chris Miller put it in *Chip War: The Fight for the World's Most Critical Technology* (2022), his seminal book on this topic: "No other facet of the economy is so dependent on so few firms."

A China-Taiwan conflict would be devastating for many reasons. One underappreciated consequence is that it would paralyze the global AI ecosystem.

In order to understand the chip industry, including its extreme concentration, it is essential to understand the concept of "fabless" chipmakers.

In the early days of the semiconductor industry, from the 1950s through the 1970s, all chip companies were vertically integrated. Chipmakers like Fairchild Semiconductor and Texas Instruments carried out every step of the semiconductor production process in-house: they designed, manufactured and marketed their chips themselves. Every chip company owned and operated its own chip fabrication facilities.

But starting in the 1980s a new model developed, driven by the logic of specialization. Two distinct types of chip companies emerged: fabless chipmakers, which design but do not produce their own chips, and foundries, which manufacture chips designed by other firms.

Today, most well-known chip companies—Nvidia, Qualcomm, AMD, Broadcom—are fabless. They do not manufacture their own chips. Instead, they rely on foundries like TSMC to build their chips for them.

Transistors are getting relentlessly smaller and chips are getting relentlessly more sophisticated with every passing year (even if Moore's Law is slowing down). The process of manufacturing cutting-edge chips is thus becoming ever more elaborate and esoteric, further reinforcing the logic of and need for companies that specialize in the art of chip fabrication.

For reference, a single human hair has a width of about 100,000 nanometers.

In 1970, the smallest semiconductor transistors were about 12,000 nanometers in width.

The most important and widely used AI chip in the world today, Nvidia's A100 GPU, has transistors that are seven nanometers wide. Google's latest tensor processing unit (TPU)—the most credible alternative to Nvidia GPUs—likewise uses seven-nanometer technology.

There are only three companies in the world that can build advanced chips anywhere near the leading edge of today's most advanced semiconductor technology: TSMC, Samsung, and Intel. Of these three, only one can reliably build the most advanced chips, including chips like Nvidia's H100 GPUs that will power the next generation of artificial intelligence.

Nvidia's hotly anticipated next-generation AI chip, the H100, has 4-nanometer transistors. The H100—which will become widely available in the coming months—is poised to turbocharge the field of AI, with virtually every AI research group in the world planning to use this new chip as soon as they can get their hands on it.

There are only three companies in the world that can build advanced chips anywhere near the leading edge of today's most advanced semiconductor technology: TSMC, Samsung, and Intel. Of these three, only one can reliably build the most advanced chips, including chips like Nvidia's H100 GPUs that will power the next generation of artificial intelligence.

THE MOST IMPORTANT COMPANY IN THE WORLD

TSMC has a market capitalization of about half a trillion dollars. It is one of the 15 most valuable companies in the world, larger than JPMorgan Chase or Walmart.

TSMC sits at the center of the global semiconductor industry and thus the entire digital world. How has TSMC become such a dominant force? And why is it so difficult for any other company in the world to replicate what TSMC does?

The first key point to understand is that powerful economies of scale exist in the world of chip fabrication, leading inexorably to winner-take-all dynamics.

To quote Chris Miller's *Chip War* again: "The economics of chip manufacturing require relentless consolidation. Whichever company produces the most chips has a built-in advantage, improving its yield and spreading capital investment costs over more customers."

Powerful economies of scale exist in the world of chip fabrication, leading inexorably to winner-take-all dynamics.

Manufacturing chips requires tremendous upfront capital expenditure. In 2021, TSMC announced that it would spend a whopping $100 billion over three years to expand its fabrication capabilities. The company recently spent $20 billion to build a single fab: its legendary Fab 18, where the world's most advanced chips (including Nvidia's H100s) are built. On top of the capital expenditure, advanced chip manufacturing requires multi-billion-dollar R&D investments year in and year out.

Unlike any other company in the world, TSMC can justify these eye-popping investments because of the sheer volume of chips it produces, far more than any other company in the world.

This creates a virtuous cycle for TSMC: companies looking to get chips built, from Apple to Tesla to Nvidia, choose TSMC because it offers the most advanced chipmaking capabilities; this gives TSMC the volume required to justify and fund ongoing investments in order to maintain its lead; and this world-leading investment budget further extends the company's advantages over rivals, making it the best choice for future customers.

The durability of this moat is best illustrated with an example.

In the 2010s, a company named GlobalFoundries sought to challenge TSMC for chipmaking supremacy. GlobalFoundries was created in 2009 when semiconductor giant AMD decided to go fabless, spinning out its fabs into a separate company. The new entity was bankrolled by Mubadala, Abu Dhabi's $300 billion sovereign wealth fund.

GlobalFoundries' original ambition was to take on TSMC directly. It invested billions of dollars in order to develop leading-edge node technology and build the world's most advanced chips. But in 2018, after less than a decade, GlobalFoundries leadership concluded that, given its scale, it would never make financial sense to make the multi-billion-dollar investments needed year after year to keep up with Moore's Law and stay on the leading edge of chip production.

GlobalFoundries gave up on developing leading-edge node technology, slashed its R&D costs and stopped competing with TSMC to build the most advanced chips. The company now focuses instead on producing lagging-edge semiconductors.

A related dynamic that helps explain TSMC's unassailable position is what has become known as TSMC's "Grand Alliance."

A combination of economies of scale, network effects and unrivaled specialization has made TSMC irreplaceable—and has made the entire world deeply, precariously dependent on it.

TSMC has invested heavily over decades to build deep partnerships with dozens of companies across the semiconductor supply chain, from EDA software providers like Cadence to equipment manufacturers like ASML to chip designers like Nvidia.

TSMC has established detailed standards for how these companies' technologies and processes interact. This ecosystem of companies has in turn developed their products in accordance with TSMC's standards; because compatibility with TSMC's processes is vital to these companies' existence, they have no other choice.

In the words of Morris Chang, TSMC's founder and longtime CEO: "TSMC knows it is important to use everyone's innovation—ours, the equipment makers, our customers, the IP providers. That's the power of the Grand Alliance. The combined R&D spending of TSMC and its ten biggest customers exceeds that of Samsung and Intel combined."

The bottom line: a combination of economies of scale, network effects and unrivaled specialization has made TSMC irreplaceable—and has made the entire world deeply, precariously dependent on it.

America Drops the Hammer

Tensions have escalated between the U.S. and China in recent years, drawing the world further and further into a new cold war. One of the most important axes of competition in this global power struggle is advanced technology. And no advanced technologies matter more than semiconductors and artificial intelligence.

In October 2022, the Biden administration ratcheted up this competition in dramatic fashion, announcing an extraordinary set of measures with one unmistakable purpose: to kneecap China's progress in AI by cutting off its access to AI chips.

The Biden administration banned the export of all high-end AI chips to any entity operating in China. Given that 95 percent of all AI chips used in China today are Nvidia GPUs, and most of the rest are AMD chips, this ban will likely be devastating to China's AI industry.

But the U.S. government didn't stop there. Taking a comprehensive view of the semiconductor supply chain, it identified a number of other strategic "chokepoints" without which AI chip production cannot be sustained—and cut off China's access to these as well.

While the United States has moved decisively to eliminate China's access to AI hardware, it is also taking steps to reduce its own reliance on chip fabrication facilities located in East Asia.

This includes the software needed to design chips' layouts, known as electronic design automation (EDA). It includes the manufacturing equipment needed to build chips. It even includes the components that go into that manufacturing equipment.

All three of the world's leading EDA companies—Mentor Graphics, Cadence Design Systems and Synopsys—are American. Most important semiconductor manufacturing equipment comes from the United States or its allies. Any entity operating in China is now barred from accessing these products.

"In weaponizing its dominant chokepoint positions in the global semiconductor value chain, the United States is exercising technological and geopolitical power on an incredible scale," said semiconductor policy expert Greg Allen. "These actions demonstrate an unprecedented degree of U.S. government intervention to not only preserve chokepoint control but also begin a new U.S. policy of actively strangling large segments of the Chinese technology industry—strangling with an intent to kill."

While the United States has moved decisively to eliminate China's access to AI hardware, it is also taking steps to reduce its own reliance on chip fabrication facilities located in East Asia.

In late 2022, TSMC announced that it would invest $40 billion to build two state-of-the-art fabs in the United States, in the state of Arizona. The first of these two fabs will begin production in 2024 and will be equipped to manufacture 4-nanometer chips. (Today's leading-edge AI chips, including Nvidia's H100, use 4-nanometer technology). The second Arizona fab is slated to come online in 2026; it will be capable of producing 3-nanometer chips, the next generation of leading-edge semiconductor technology.

TSMC's decision to build prized leading-edge fabs in the United States—in other words, to share its crown jewels with the Americans—was the result of heavy pressure and lavish subsidies from U.S. officials. Ultimately, as much as the United States needs Taiwan, Taiwan needs the United States even more; TSMC had little choice but to play ball.

Bringing advanced chip production to U.S. soil will help alleviate the AI industry's absolute dependence on Taiwan-based fabs. Yet the Arizona fabs are not a panacea. For one thing, their production capacity will be relatively modest: in total, these two fabs are expected to produce 600,000 silicon wafers per year. To put that figure in perspective, TSMC produces over 13 million wafers per year in total, meaning that the American fabs will represent less than 5 percent of its total output.

The most advanced semiconductor production nodes will remain in Taiwan.

Moreover, the most advanced semiconductor production nodes will remain in Taiwan. By the time the 4-nanometer Arizona fab begins production in 2024, and then the 3-nanometer Arizona fab begins production in 2026, these facilities will be one generation behind the leading-edge nodes, which only Taiwan-based fabs will be able to produce. TSMC's core R&D efforts and team will likewise stay in Taiwan.

Still, diversifying AI chip production beyond Taiwan's borders is a big deal. Many in Taiwan have fiercely opposed these moves. TSMC's legendary founder Morris Chang himself has criticized America's attempts to onshore advanced chip production and has spoken out against what he warns is a "hollowing out" of Taiwan's chip sector.

This opposition is hardly surprising: Chang and others recognize that, if the world no longer relies on the island of Taiwan for advanced semiconductors, Taiwan will lose much leverage in this complicated geopolitical dance.

WHERE DO WE GO FROM HERE?

The confluence of great power competition, the trillion-dollar semiconductor supply chain, and rapid advances in AI has brought us to a critical and delicate juncture in world affairs. It is here that the worlds of bits and atoms collide. The stakes could not be higher.

Artificial intelligence plays a central role in the drama—as a source of leverage, a weapon, and a potential casualty of this great power struggle.

How might things play out from here? No one knows for sure, but let's consider some possibilities on this three-dimensional chessboard.

Let's start with the optimistic scenario. Taiwan's central role in the global semiconductor industry is often referred to as its "silicon shield." The basic theory is this: because China relies so heavily on Taiwan for the chips it needs to fuel its economy (70 percent of all chips in China are made by TSMC), it will stop short of invading Taiwan and putting TSMC's production at risk, since doing so would decimate China's own economic health.

And because the rest of the world depends so deeply on TSMC, the United States and other powers will go to great lengths to defend the island and protect its sovereignty—a fact that China understands well. Unwilling to risk a full-fledged global conflict, China will judge that it is not rational to initiate hostilities with Taiwan. Under this theory, while China may continue to build up its military and engage in cross-strait saber-rattling, it will be dissuaded from kinetic action against Taiwan.

But the silicon shield is just a theory, not a guarantee. China's military, especially its navy, has grown far stronger in recent years and as a result has begun asserting itself increasingly confidently.

And it is important to keep a broader perspective here: China's calculus on Taiwan does not depend solely on semiconductors.

To be sure, semiconductors are a critical strategic resource. But the China/Taiwan struggle runs much deeper. Ever since the Chinese Nationalist Party's "Great Retreat" to Taiwan in 1949, the Chinese Communist Party (CCP) has viewed Taiwan as a rogue province, not an independent nation, and has considered it a non-negotiable inevitability that it would one day reabsorb the island. This is a core part of the CCP's vision, identity, and understanding of its own sovereignty, irrespective of chips.

The U.S. National Security Council recently estimated that a China-Taiwan armed conflict could cost the global economy over $1 trillion annually due to disruptions in semiconductor production.

So: what would happen if China did move decisively to retake Taiwan? Big picture, the economic impact would be catastrophic. The U.S. National Security Council recently estimated that a China-Taiwan armed conflict could cost the global economy over $1 trillion annually due to disruptions in semiconductor production.

It is beyond the scope of this essay to speculate as to whether and how the U.S. military would respond to defend the island. One thing that we can say with some confidence, though, is that TSMC's fabs would almost certainly be rendered inoperative before they fell into China's hands.

It is plausible that the Taiwanese or even the U.S. military would preemptively destroy the fabs rather than permitting the CCP to take control of this invaluable strategic resource. Even if the physical buildings were to remain undamaged after a Chinese invasion, it is unrealistic that the CCP would be able to continue operating them to produce cutting-edge chips. Keeping leading-edge fabs running requires collaboration from partners across the global semiconductor ecosystem and a constant inflow of materials, equipment, and services from suppliers, which would be denied to an invading power.

"If a totalitarian regime forcibly occupied TSMC, its kaiser would never get its partner democracies on the phone," explained Wired's Virginia Heffernan. "The relevant material suppliers, chip designers, software engineers, 5G networks, augmented-reality services, artificial-intelligence operators, and product manufacturers would block their calls. The fabs themselves would be bricked."

Former U.S. State Department undersecretary Keith Krach put it more vividly: "They call Taiwan the porcupine, right? It's like, just try to attack. You may just blow the whole island up, but it will be useless to you."

Let's continue playing this scenario out. If China were to take Taiwan by force, and TSMC's fabs thus went offline—if no more Nvidia A100s or H100s or any other AI chips could be produced—what would happen? What could the world do to fill this gaping hole in chip supply? What would it mean for the field of AI?

After TSMC, the company best positioned to step up and produce cutting-edge AI chips is Samsung. Samsung is currently the only company in the world other than TSMC that can produce 3-nanometer chips, today's leading-edge technology.

Yet Samsung's production quality is far inferior to TSMC's. "Yield" is an important industry metric that indicates the percentage of silicon wafers introduced into a fabrication process that end up as functioning chips. TSMC's yield on its 3-nanometer chips is estimated to be as high as 80 percent. Samsung's, meanwhile, was between 10 percent and 20 percent when it began 3-nanometer production in 2022 (though more recent reports suggest that it may be improving). Samsung's subpar production quality recently prompted Nvidia to move the production of all its GPUs—not just its high-end AI chips—away from Samsung to TSMC.

In a best-case scenario, it would take Samsung years to scale up to TSMC's current AI chip production levels and yields.

Moreover, from the perspective of the western world, Samsung's fabs are themselves in a vulnerable location: Korea is a tiny peninsula directly bordering China, thousands of miles away from the United States.

This brings us to America's erstwhile chip champion Intel. It was hardly a decade ago that Intel's chipmaking capabilities were the envy of the world. But in recent years, following strategic miscalculations and manufacturing setbacks, the company has fallen behind. Intel struggled mightily in its transition to both 10-nanometer and 7-nanometer node technology, with its 7-nanometer chips only entering full production this year. In order to avoid further delays, the company has even resorted to outsourcing some parts of its 7-nanometer fabrication process to rival TSMC, a humbling and previously unthinkable move for the proud chipmaker.

Under new CEO Pat Gelsinger, Intel aspires to regain its chip manufacturing supremacy. The company has set ambitious goals to begin production of 2-nanometer chips by 2024 and to deliver five new nanometer nodes over the next four years, leapfrogging TSMC.

Some observers believe that, were TSMC's fabs to be rendered inoperative, the U.S. government would move forcefully to broker a partnership between American competitors Intel and Nvidia. In broad strokes, the idea would be that Intel would aggressively ramp up its manufacturing capabilities by any means necessary in order to support production of Nvidia's GPUs as soon as practicable.

Whether a "frenemy" collaboration along these lines is realistic, though—in particular, whether Intel has the manufacturing chops to pull this off on any reasonable timeline—is far from clear.

Before we despair too much, let us note a few more encouraging points.

First, keep in mind that a considerable stock of AI chips already exists in the world and could remain in use.

Most AI chips in the world today are owned by cloud providers, who make them widely available to other organizations "as a service." This includes the cloud giants Amazon Web Services, Microsoft Azure and Google Cloud Platform, as well as a handful of upstart cloud challengers specializing in AI workloads like Oracle, CoreWeave, and Lambda Labs.

The cost to access AI chips would skyrocket in this scenario. After all, even without any disruption to TSMC's operations, the world already faces a massive shortage of GPUs thanks to breathtaking recent growth in the AI market. These cloud providers' margins would balloon, though their revenue growth would slow dramatically.

Such a scenario might heavily advantage incumbents. AI giants like OpenAI (via its Microsoft relationship) and Google (via its TPU program) would continue to have access to vast AI computing resources, enabling them to continue pushing forward the frontiers of AI research. Other large companies would also be better equipped to foot the bill to use AI chips.

While the most advanced AI chips, like Nvidia's H100s and Google's TPUs, can only be manufactured in Taiwan, there are plenty of fabs around the world capable of producing lagging-edge logic chips at scale.

Resource-constrained startups, on the other hand, might find it untenable to pay to access AI hardware at scale. This could discourage them from building new AI models, challenging incumbents, and taking the field of AI in new directions. The world would be the worse for it.

A second consideration to keep in mind: while the most advanced AI chips, like Nvidia's H100s and Google's TPUs, can only be manufactured in Taiwan, there are plenty of fabs around the world—from the U.S. to Europe to Israel—capable of producing lagging-edge logic chips at scale.

Though much less powerful than today's leading AI chips, previous-generation chips could be used in a pinch to support some AI computing workloads.

This scenario would benefit chip manufacturers like GlobalFoundries and Samsung that could ramp up production of lagging-edge AI chips in response to surging demand.

This solution would not be without its challenges, though.

Lagging-edge semiconductors are less cost-efficient and energy-efficient than today's leading AI chips. As a result, any given task—say, training an AI model of a given size—would be far more expensive, time-consuming and carbon-intensive to carry out.

A large-scale transition to lagging-edge semiconductors across the global AI ecosystem would in a best-case scenario be tremendously disruptive. As COVID made clear, supply chain disruptions can wreak far-reaching economic havoc and require years of adjustment.

AI companies around the world have developed their technology stacks, vendor relationships, product roadmaps, commercial timelines and financial budgets based on the availability of leading-edge chips like Nvidia's A100 and H100 GPUs. These would all need to be reformulated.

Still, lagging-edge semiconductors would at least provide some hope for a path forward in the event that Taiwan's fabs went offline.

One final consideration that gives reason for optimism: many cutting-edge AI models already exist.

A researcher or entrepreneur that wants to build a new product with best-in-class large language models need not use a bunch of GPUs to train their own model from scratch. They can go to OpenAI, Cohere or Anthropic to access state-of-the-art LLMs via API; or they can go to Hugging Face and get the model weights for any number of high-performing open-source LLMs (e.g., Meta's LLaMA, Stanford's Alpaca, Databricks' Dolly).

Even if all AI research progress were to stop tomorrow, the technology as it exists today promises to create trillions of dollars of enterprise value in the years ahead as the world figures out how best to productize, commercialize and integrate AI across the economy.

Even if all AI research progress were to stop tomorrow, the technology as it exists today promises to create trillions of dollars of enterprise value in the years ahead as the world figures out how best to productize, commercialize and integrate AI across the economy.

Yet, having said all of this, we cannot escape the fact that it would be devastating for humanity to lose our ability to produce the advanced chips that power today's cutting-edge artificial intelligence. Progress in AI would be profoundly disrupted; the global technology ecosystem would be debilitated.

Given dangerously rising tensions between China and Taiwan, combined with the extreme global concentration in AI chip manufacturing, this is an all too real possibility today.

Let us hope that cooler heads prevail.