Silicon Borders: The Global Justice of AI Infrastructure

Johannes Himmelreich
Maxwell School of Citizenship and Public Affairs at Syracuse University

Introduction

By 1987 the Reagan Administration had been debating for years whether India should be allowed to buy a supercomputer. The Pentagon argued that granting India's export request was a threat to national security, given that India at the time had an active nuclear weapons program and close ties to the Soviet Union. The United States Commerce Department eventually granted India an export license for a mid-powered Cray XMP-14 computer. Two years later, the Bush administration faced the same question—this time for Brazil and Israel—and by the 1990s, not only supercomputers but also cryptography was subject to export controls.

Today, technology travels the world more freely than ever. The source of the most advanced technology, as before, is the United States. This technology dominance has made export controls a powerful tool. The Biden administration started a policy of "chokepoints," restricting the export of advanced chips and software, primarily targeting the People's Republic of China (PRC). However, this policy trend creates "silicon borders" that can affect developing countries around the globe.

Technology dominance is geopolitical power. Consider Ukraine, which is highly dependent on Starlink, the low-earth satellite service of Elon Musk's SpaceX. Ukraine's "entire front line would collapse if I turned it off," Musk has stated. Reportedly, the United States has used Ukraine's dependency on Starlink as leverage in negotiations.

AI infrastructure, just as the Starlink satellite network, tends towards a natural monopoly: The most efficient market structure would be a single provider. This enables technological dominance. When AI infrastructure becomes an important resource, unilateral trade restrictions on AI infrastructure by the dominant power— "silicon borders"—are a problem for global justice.

Of course, much is uncertain about AI and its future. Equally unclear are frameworks for analyzing AI's global impact as it develops. This article offers conceptual clarifications and sketches a framework for global AI policy analysis.

The Economics of AI Infrastructure

AI infrastructure is the foundation of AI applications and services. It is the set of resources that enable the development, training, and deployment of AI systems. AI infrastructure consists of four components.

1. **Computational resources**: Specialized hardware used in AI data centers
2. **Data resources**: Training datasets as well as software to collect, generate, and process such datasets
3. **Foundation models**: Large neural networks trained on vast datasets using significant compute resources

4. **Distribution mechanisms**: Cloud services, inference software, and application programming interfaces (APIs)

Computational resources include specialized chips, like Graphics Processing Units (GPUs), for which Nvidia, an American technology company based in California, is the market leader. In conjunction with their surrounding software ecosystem, these chips are used to train AI models and to "run" them—known also as "inference." Large amounts of these chips are required to train and adapt large AI models, known also as "foundation models."

Some AI developers publish their models. A model is basically a file that describes the neural network and the *weights*, i.e., the connections in this network. Publishing a model disentangles development and deployment. With a model file in hand, you can run the model in any capable data center. Such "open weights" models thus created a market in compute resources where cloud companies "host" such open models. This enables competition and is good for privacy: Since anyone can host an open model, you can decide whom you trust with your data and to run any of the open models.

By contrast, large AI companies, such as xAI, Google, and OpenAI, bundle different components of the AI infrastructure together. These companies build out the computational resources, collect and condition datasets, train foundation models, and then distribute their models exclusively. You cannot work with their models unless you go through them.

Importantly, AI infrastructure is also an infrastructure in an economic sense: It has high fixed costs, low marginal costs, and significant economies of scale. Because of these features, AI infrastructure tends towards a natural monopoly.

The fixed costs associated with developing frontier AI capabilities are extraordinary. Training a model like OpenAI's GPT-4 cost around $40 million for the training run. If you factor in research and people, the whole process cost over $100 million. By 2027, these costs could reach more than $1 billion. Thus, the fixed cost barrier to join this competition is prohibitively high and always increasing. Few entities have the financial means to play in this game.

But once you built a data center, collected the data, and trained a foundation model, the *marginal* costs for inference are remarkably low. Serving an additional user or request is nearly negligible. Of course, in aggregate, running a model consumes massive amounts of energy. But the average costs decline dramatically as output increases. This drives market concentration; a small number of companies hold a large percentage of the overall market.

This tendency for market concentration is reinforced by economies of scale. As more people use a particular AI infrastructure—such as the one offered by OpenAI—the ecosystem surrounding that infrastructure grows. Each additional user gives you more data, which you can use to improve the model. The current generation of models then trains the next generation of models—involving what is known as "synthetic data" and "model distillation." Thus, those who have many users and good models will develop even better models and acquire even more users—a self-reinforcing cycle of dominance through network effects.

The current industry landscape reflects this: A handful of technology companies with vast capital, computational resources, and scientific talent occupy dominant positions. In specialized AI hardware, Nvidia maintains approximately 80% market share for training chips. In cloud services that

host AI infrastructure, three providers (Amazon Web Services, Microsoft Azure, and Google Cloud) control roughly 65% of the global market.

This dominance may appear fragile. In January 2025, the Chinese company DeepSeek released their "R1" model, which matched the capability of earlier models by OpenAI, shocking stock markets. To some, this only showed how important silicon borders are to ensure continued American dominance. It also shows that open models have the potential to disrupt the AI infrastructure market.

Going forward, the AI infrastructure market might be competitive in some segments but not others. The segment around smaller models, which require fewer computational resources, will have lower barriers to entry and greater competition. Such smaller models are useful in many day-to-day applications like providing recipes, advice, coding, or companionship. More capable models will be larger and will offer critical capacities beyond what such smaller models can offer. The market segment for such larger "frontier" models will have high and increasing barriers to entry and greater tendencies towards a natural monopoly.

AI Infrastructure Chokepoints

AI infrastructure is not just economically concentrated—it is politically gated. Where the economics of AI created market concentration, the United States started to exploit technological dependence as "chokepoints." The Biden administration has established export controls, notification requirements, and licensing regimes. These silicon borders are justified in terms of national security, mainly targeting the PRC. But the result may be a system in which some countries can build on powerful AI tools, while others are left behind.

Compute: The GPU Chokepoint

The first chokepoint is hardware. In October 2022, the U.S. Department of Commerce announced restrictions on the export of high-end AI chips to China. Later revisions expanded the list to include other types of chips and imposed restrictions on certain countries in the Middle East, including Lebanon, Libya, and Syria. The latest "AI Diffusion" rule limited exports with all but 18 countries. The Trump administration, for now, rescinded this latest rule.

These chips are indispensable. Training a modern foundation model requires tens of thousands of top-end GPUs running in parallel for weeks or months. Restricting access to these chips is effectively restricting access to the frontier of AI.

For many middle-income countries, the hardware needed to train competitive AI models is out of reach. However, even close allies like Israel may face rising barriers (with some exceptions if a data center is operated by a U.S. provider, such as Microsoft).

Foundation Models: Controlled Capabilities

A second chokepoint is access to foundation models themselves. Under the AI Diffusion rule, frontier models—like GPT-4, Gemini, or Claude—cannot be licensed, transferred, or made available to users in certain countries without explicit authorization. The trend is towards treating highly capable AI models as export-controlled items.

The logic here mirrors Cold War-era controls on encryption and supercomputers. Foundation models could be used for disinformation, cyber warfare, or weapons design. This "dual use" concern means that access to foundation models depends on geopolitics.

Distribution: APIs and Cloud Access

The most effective and relevant chokepoint lies in how models are accessed. Foundation models are rarely downloaded and run in independently operated data centers. Instead, models are hosted in the cloud and accessed via APIs that offer the model as a service. Whereas chips can find their way to China through loopholes and can't be summoned back once they are there, access to models via APIs can be controlled much more effectively.

API restrictions are hidden chokepoints. They need not be announced as policy, but their effect is powerful. Currently, major AI companies, OpenAI and Anthropic, allow their models to be accessed widely, except from Iran, Syria or North Korea. However, even if access is not explicitly restricted, companies tend to over-comply with government export restrictions. Moreover, the source of power is the potential to shut down access.

For all countries without domestic AI champions—meaning practically everyone except the United States and China—such dependency is a vulnerability. Losing access to APIs would not only hamper the development of next-generation applications, but also reduce the productivity overall across all tasks for which such foundation models could help, such as language translation, software engineering, data analysis, drug development, and forecasting of political and economic risks, to name just a few.

Global Justice of AI Infrastructure

Silicon borders are powerful. As AI capabilities become essential inputs for economic development, scientific progress, and even basic public services, nations that lack direct control over this infrastructure become dependent on those who wield it.

Who draws these silicon borders, and on what terms, is hence a central question of global justice. Silicon borders risk violating basic principles of global justice. Within political philosophy, there are three interconnected lines of argument which support this analysis.

1. The Level Playing Field Argument

The global economy operates—in theory and in practice—as a cooperative system in which everybody, in the aggregate, wins. This win-win logic of global trade requires a level playing field. No side should get an advantage. For example, rules around copyright and intellectual property must be adhered to by everyone insofar as these rules maintain a level playing field.

The principle of justice here is this: Trade must offer fair gains to all participants. Advantages for some and disadvantages for others undermine the idea of justice that is inherent in the social practice of trading with one another.

This idea is developed by the philosopher Aaron James. He illustrates it with a discussion of the WTO's global intellectual property (IP) regime. Stringent IP protections are fair between developed countries but unfair for developing countries—because developing countries lose out on the opportunity of fair gains. For some trade parties, some rules tilt the playing field.

This would be the case for silicon borders. Restricting access to AI infrastructure can be an unfair disadvantage for some. Silicon borders, even if they are aimed at the PRC, have effects on developing countries that are already disadvantaged by *structural asymmetries* of the global digital economy. For developing countries, silicon borders may tilt the playing field.

2. The Coercion Argument

A second argument is loosely based on the idea that justice doesn't follow power—might doesn't make right. Instead, those who coerce others or restrict their freedom—even by inaction—commit a wrong, such as violating their human rights.

In this vein, philosopher Laura Valentini argues that global economic systems are coercive and that, at some point, such coercion is incompatible with an equal respect for persons. Coercive economic structures may not be justifiable to everyone who is subject to them. Treating everyone with respect would require changing these structures where possible.

Likewise, silicon borders are coercive. Powerful actors unilaterally control access to AI infrastructure. The United States limits the export of advanced GPUs and private companies dictate terms of service for cloud platforms or APIs. When such silicon borders have effects that are severe enough, they fail to treat all people with equal respect.

3. The Resource Interdependence Argument

A third argument builds on the fact of deep economic interdependence. Nations are not isolated units. This interdependence transforms the way we should think about resources essential for economic life. In a slogan: Where there is interdependence, there must be justice. One proponent of this argument is the political theorist Charles Beitz.

AI infrastructure, though human-made, increasingly functions similarly to a critical natural resource. Access to computational power, foundation models, and the platforms for distributing them is becoming as crucial for economic participation and development as access to energy, capital, water, or physical trade routes. A resource that is a fundamental input in a shared global economy needs to be distributed fairly. Silicon borders thus might become an unfair obstacle to resource access.

Balance with National Security

On the other hand, these three arguments might be missing something. Silicon borders are levers of power yielded for a good purpose: they protect national security and democracy.

AI infrastructure is a dual-use technology. Unfettered access, or so the dual-use argument goes, could allow adversaries to use AI for various nefarious ends: intelligence analysis, logistics, weapons, cyber warfare, and breakthroughs in material science or cryptography with a military use.

As such, silicon borders protect democracy. For one, leadership in AI ensures geopolitical leadership—and silicon borders prevent the authoritarian PRC from gaining geopolitical leadership. Moreover, as long as the United States is a few years ahead of its rivals, it has a "cushion" to develop AI technologies safely. If the race gets closer, the focus within AI development will shift away from preventing harms to short-term capability gains.

Thus, silicon borders protect national security and enable safe AI. The whole world might be better off as silicon borders avoid an AI race.

Politics of Crisis Governance

This national security argument echoes the older melody of *crisis governance*: Faced with a perceived high-stakes threat—whether terrorism, financial collapse, or AI proliferation—governments argue that decisive, often preemptive, action is necessary.

The global politics of AI is conducted as crisis governance. Unfortunately, the feeling of being in a crisis often clouds our thinking.

First, the chokepoint theory assumes that national security needs to be "balanced" against other values such as openness, free trade, or global justice. Similarly, after 9/11, national security was balanced against civil liberties. However, this metaphor of "balancing" some values against others is problematic. Treating global justice claims merely as interests to be "weighed" or "balanced" fundamentally misunderstands their nature. Claims of justice might be strict, not simply a competing consideration.

Second, the national security argument assumes *effectiveness*. But silicon borders might not be effective, not even in the short term. For starters, both exporters and importers circumvent restrictions. In 2022, Nvidia repackaged their chips in a new product, the H20, that they "specifically designed to comply with export controls." And, reportedly, Nvidia tried to do the same thing again to deal with further restrictions. The export controls also give the PRC an incentive to develop their own technology. This might allow China to leapfrog ahead in the competition before long. In addition, what is done in the name of security is often mostly symbolic "security theater"—responding to fear or the feeling that we had to do *something*. In short, silicon borders might have the opposite of their intended effect.

Finally, state power can metastasize. Granting any state sweeping powers to control AI infrastructure in the name of security creates tools that can be readily abused for other ends, including economic coercion, geopolitical leverage, or even domestic surveillance. Any lack of robust oversight—which is often deficient in national security matters—exacerbates this risk.

National security remains a legitimate concern and an important government function. Yet it too easily serves as a convenient veil for protectionism or geopolitical maneuvering.

The Road Ahead

Silicon borders are an emerging challenge for global justice. AI infrastructure risks becoming monopolized. This is a problem of global justice on three grounds: Silicon borders tilt the playing field, are coercive, and ignore resource dependencies.

The solution is not a demand for unrestricted access. Even if the "balance" metaphor is problematic, concerns of national security can't be entirely dismissed. What can be done?

First, dominant powers should exercise restraint in the scope of export controls, limiting restrictions to truly strategic technologies. The empirical assumptions that underwrite the chokepoint strategy need to be validated. Do these policies achieve their goals? The strategy itself needs to be

audited: Is the policy really driven by a concern for national security? Or is economic protectionism the real goal?

Second, regional cooperation offers a promising path forward, particularly for MENA countries. Pooling resources for shared research centers, joint procurement of AI infrastructure, and collaborative regulatory frameworks could create economies of scale that individual nations cannot achieve alone.

This regional approach moreover aligns with the institutionalists' thesis: Despite the prominence of a global justice discourse, a society's prosperity is primarily a function of domestic institutional quality. By extension, access to AI infrastructure is valuable primarily when countries have the institutional capacity to deploy it effectively, transparently, and for public benefit. The most successful responses to silicon borders will hence come from countries that combine advocacy for fairer global access with liberal domestic governance. This requires not just investment in technology but in an educational, legal, and economic ecosystem.

History suggests that policymakers tend to *over*estimate the risks of technology proliferation and *under*estimate the benefits of cooperation and technology diffusion. Long-term security may depend more on global stability and shared prosperity—potentially fostered by wider AI access—than on the prospect of maintaining a technological edge through silicon borders.

The past is prologue. Throughout the 1980s, the United States was losing the semiconductor industry to Japan. As today, the Reagan administration responded with trade restrictions and industrial policy. In vain. The industry moved to Asia. But still, the next technological revolution took place in the United States—because it successfully attracted many highly skilled engineers. If history is any guide, then domestic success depends on technological talent and an attractive open ecosystem—not choking global trade for uncertain gains.