Compute Access and Research Diversity

Flyxion
Independent Researcher

November 2026

Abstract

Debate over the sale of advanced accelerators to Chinese buyers is usually conducted as a contest over who holds the lead. This opinion piece proposes a different mechanism as the one that matters most: export policy changes where expertise accumulates. A chip gives a laboratory hardware for a few years, while the kernels, debugging habits, and teaching practices built around that hardware last much longer, and a restriction can slow a laboratory today while giving it a reason to build tools that make the restricted platform less necessary tomorrow. Because the field is still unsettled at the level of basic method, the essay argues for wide access to compute together with portability away from any single platform, so that common infrastructure widens participation without narrowing the range of experiments that are easy to run. It also argues that concern about misuse is best assessed instrument by instrument, by asking what a given restriction prevents and what research it forecloses.

Keywords: export controls, compute access, accelerators, CUDA, toolchain portability, technological lock-in, epistemic diversity, research policy

1 Where Expertise Accumulates

The hardware is only part of what a supplier provides. A researcher who learns on a particular accelerator also learns its programming model, its performance quirks, and the folklore that surrounds both. Kernels are written, debugging habits form, and the whole body of practice is passed to the next cohort of students and colleagues. This accumulated knowledge belongs to the community of users and not to the vendor, yet it is the principal reason that the vendor remains the default choice. It is the familiar pattern of increasing returns, in which an early lead in adoption can lock in a technology that is not the best one available [3, 1]. Nvidia's CUDA platform is a clear present example, and recent reporting describes a developer base of several million people that rivals have found difficult to displace [6].

The consequence for export policy is a cost that the ledger of lost sales does not capture. A chip sale gives a laboratory capacity for the few years in which the chip is useful, whereas the expertise built around the chip can last far longer, and so policy toward sales is also policy toward where a very large body of practical knowledge will accumulate. When a population of engineers is denied access to the incumbent platform, it does not stop working. It builds its tools elsewhere. Reporting during 2025 and 2026 describes exactly this response: after the leading American processors were prohibited, Nvidia's chief executive stated that its share of the Chinese market for advanced accelerators had collapsed [2], a domestic vendor expanded its own software stack, and open-source programming tools for that vendor's chips were released by a prominent Chinese model developer in September 2026 [5]. The significance of the release is that it gives researchers a shared body of code they can inspect, teach, and extend. Whether that work can travel beyond the hardware it was written for is a separate question. The domestic ecosystem is still small beside the incumbent by developer count [7], and the figures that circulate about market share vary considerably between sources, so any precise claim about the present balance deserves caution. The direction of travel is nonetheless what the lock-in argument would predict. A restriction intended to preserve an advantage may be teaching a large community to be fluent in an alternative, and fluency, once acquired, is as durable as any patent.

The restriction is also not only a Washington policy. Beijing has itself discouraged purchases of chips designed to comply with American rules, which means that a discount offered to buyers would run into a second government with its own reasons for preferring domestic dependence, and a steeper discount would make the offer more attractive to buyers while making it more threatening to the policy behind the refusal. This is why the idea summed up as getting users hooked describes something real that cuts in a direction opponents of sales do not always notice. Familiarity is acquired where access exists. Engineers who learn on American hardware will, in the ordinary course, build tools for it and prefer it, while engineers who are cut off will build the same capital around a different platform. Whether one welcomes that depends on one's goals, but it should not be mistaken for a neutral side effect. Readers may reasonably disagree about the policy conclusion and still evaluate the mechanism on its own terms.

2 A Contest Declared Too Early

The public conversation about accelerators has also settled into the vocabulary of an arms race. Chips are counted, shipments are licensed, and the lead is measured in generations of hardware. This vocabulary assumes that the destination is known and that the only open question is who arrives first. That assumption is hard to defend. The field cannot yet say which architectures will matter in a decade, which training regimes will prove to be dead ends, or whether the present dominant approach will be recognizable in the systems that eventually do the most useful scientific work. Some engineering practices, such as distributed training and mixed-precision arithmetic, have become fairly standard, so the claim is not that everything is open. It is that the questions of basic method on which the largest gains would depend remain unsettled. A race presupposes a course, and here the course is still being surveyed.

If the task is to reach a known finish line quickly, concentrating resources and denying them to a competitor is a coherent strategy. If the task is to discover what the finish line is, then concentration is a hazard, because a field that has committed its resources to one hypothesis has also committed its attention to it. A well known result in the philosophy of science makes the point in general terms: a scientific community performs better when its members do not all follow the same method, even when some of them follow a method that an outside observer would rate as less promising [4]. The ranking is made with the information available now, and the information that would correct the ranking can only come from people who were permitted to try the disfavored path.

3 Two Kinds of Division

The word division carries two meanings that are easily confused. One kind is a barrier: a researcher with a good idea cannot test it because the necessary compute is priced out of reach or legally withheld. This kind suppresses hypotheses before they are tried, and it deserves no defense. The other kind is independence: research groups that do not depend on one another, or on a single dominant laboratory, for permission to pursue an unfashionable approach. This kind is what allows incompatible hypotheses to coexist long enough to be compared. Cheap and widely available hardware serves independence by removing barriers.

Work on the structure of epistemic communities shows that the pattern of communication among researchers can matter as much as their individual competence. In simulations of groups that share results, densely connected networks tend to converge quickly, sometimes on a conclusion that further inquiry would have overturned, while sparser networks preserve disagreement longer and more often reach the better answer [8]. That result concerns how quickly findings propagate among researchers, and it should not be stretched to cover a shared vendor, since laboratories can all use the same accelerator and still pursue incompatible approaches, just as laboratories on different hardware can all chase the same benchmark. The relevant risk from common infrastructure is of a different kind. A good toolchain makes some computations cheap to implement, debug, and reproduce, and that advantage is real. It can also shape which ideas look practical enough to attempt, because an approach that fits the dominant platform poorly will be expensive to try even when it is promising.

Common infrastructure can therefore widen participation while narrowing the range of experiments that are easy to run.

This distinction removes an apparent contradiction. One can defend wide access to the dominant platform, which broadens participation, and also defend portability away from it, which keeps the range of easy experiments from shrinking to what the platform favors. The two commitments address different parts of the problem. Access decides who may take part, and portability decides whether taking part commits a group to a single way of working.

4 Dual Use and the Narrower Claim

The strongest objection to this view is that the same compute that allows researchers to explore also strengthens institutions that use artificial intelligence for surveillance and for the enforcement of political goals. The objection is serious. Large platforms in China already combine messaging, payments, and administrative services in ways that give the state extensive visibility into daily life, and it would be naive to expect machine learning to be left out of such systems. It would also be a mistake to claim that frontier accelerators make no difference to the scale, speed, or sophistication of every surveillance system, since greater capacity plausibly allows more data to be processed, more people to be tracked, and more capable models to be deployed.

The argument needs only a narrower claim, namely that whether an export control prevents harm is an empirical question, and that the answer must be weighed against the cost the control imposes on open research. The fact that some misuse is possible does not tell us whether a given restriction meaningfully reduces it, and it does not tell us what research the same restriction forecloses. The question to ask of any instrument is what specific harm it prevents, how much of that harm it prevents, and what it costs in forgone experimentation. A restriction on frontier chips may reduce the capability of the most demanding systems while leaving the routine classification and administrative uses largely untouched, and the same restriction can also foreclose exploratory work that cannot be done with readily available substitutes. Whether that trade is acceptable is a judgment that should be made openly and not assumed in advance.

Oversight of deployment is the obvious alternative, and it has its own weaknesses. It is difficult to enforce across jurisdictions, it depends on the cooperation of the governments whose conduct is in question, and it cannot easily see what happens inside a state apparatus. These are reasons to compare the instruments honestly, and neither is entitled to be treated as the one that works. There is a further consideration on the side of openness. The ability to anticipate and prevent abuse depends in part on understanding the technology, and understanding is produced by a diverse and well-resourced research community. A world in which only a few laboratories can afford to study these systems has few independent sources of expertise about what the systems do.

5 Scale as a Thought Experiment

A final consideration concerns what the compute is for. Life extension requires models that span molecular, cellular, tissue, and organismal scales and that are revised against evidence over many years. The design of very large engineered structures requires coupled models of materials, construction, climate, and maintenance whose uncertainties compound. As a thought experiment, suppose the number of accelerators in the world were multiplied a thousandfold. The essay makes no forecast about whether such a world is near and does not need to show that the increase would still be insufficient. The interesting question is how the cheaper compute would change what people decide to investigate: more competing models could be run, uncertainty could be carried across more scales, and simulations could be revisited more often as evidence arrives.

More compute would not settle which model is right. A larger run cannot tell you whether a mechanism missing from the model was important, and it cannot tell you whether a structure behaves outside a simulation as it does within one. Only experiments and field observation can supply that information. A single enormous run of one idea is a poor use of an enormous budget when the principal uncertainty concerns the idea, whereas the same budget distributed over many ideas, each pursued by a group free to disagree with the rest, yields more information about which ideas deserve further scaling. The thought experiment is useful because cheaper computation exposes bottlenecks other than permission, and the essay's own examples name two of them, obtaining experimental evidence and validating designs in the field. The question worth asking of a world with less scarce compute is therefore which bottlenecks become visible once the first one is relaxed, and which of them a diverse set of research programs is better placed to address than a single large one.

6 What Follows

The practical implication is modest in form and broad in effect. Policy toward compute should be evaluated by whether it makes it cheaper to try a different approach and easier to carry that work elsewhere, and not by whether it maximizes a rival's disadvantage in the short run. In practice this could take the form of compute credits aimed at a wide range of independent groups instead of a few large buyers, together with public funding for portability tools such as open compilers and kernel languages that run on more than one vendor's hardware. A discount can make a platform available to more groups, and portability helps keep that platform from becoming a condition of participation. A subsidy that merely finances the largest clusters would fail this test, however many chips it moved, because the number of chips sold is a poor proxy for the number of independent groups able to pursue ideas that the dominant laboratories would never try. If the research frontier is still unsettled, the aim should be to give more groups the means to discover that the rest of us are wrong.

References

[1] Arthur, W. Brian. 1989. "Competing Technologies, Increasing Returns, and Lock-In by Historical Events." *The Economic Journal* 99 (394): 116–131.

[2] Caixin Global. 2025. "US Export Ban Wipes Out Nvidia's China Market, CEO Says." October 18, 2025. https://www.caixinglobal.com/2025-10-18/us-export-ban-wipes-out-nvidias-china-market-ceo-says-102372899.html

[3] David, Paul A. 1985. "Clio and the Economics of QWERTY." *The American Economic Review* 75 (2): 332–337.

[4] Kitcher, Philip. 1990. "The Division of Cognitive Labor." *The Journal of Philosophy* 87 (1): 5–22.

[5] Let's Data Science. 2026. "DeepSeek Open-Sources Ascend Software With Huawei." https://letsdatascience.com/news/deepseek-releases-ascend-software-with-huawei-847f368b

[6] The Next Web. 2026. "US Export Controls Are Pushing China's AI Chip Industry Away From GPUs and Toward Custom Silicon." https://thenextweb.com/news/china-ai-chip-asic-gpu-nvidia-export-controls

[7] Tom's Hardware. 2026. "Huawei Expects $12 Billion in AI Chip Revenue This Year as Nvidia's China Market Share Hits Zero." https://www.tomshardware.com/tech-industry/huawei-expects-12-billion-in-ai-chip-revenue-this-year-as-nvidias-china-market-share-hits-zero

[8] Zollman, Kevin J. S. 2007. "The Communication Structure of Epistemic Communities." *Philosophy of Science* 74 (5): 574–587.