On DeepSeek and Export Controls

Dario Amodei

January 2025

A few weeks ago I made the case for stronger US export controls on chips to China. Since then DeepSeek, a Chinese AI company, has managed to — at least in some respects — come close to the performance of US frontier AI models at lower cost.

Here, I won't focus on whether DeepSeek is or isn't a threat to US AI companies like Anthropic (although I do believe many of the claims about their threat to US AI leadership are greatly overstated)1. Instead, I'll focus on whether DeepSeek's releases undermine the case for those export control policies on chips. I don't think they do. In fact, I think they make export control policies even more existentially important than they were a week ago2.

Export controls serve a vital purpose: keeping democratic nations at the forefront of AI development. To be clear, they’re not a way to duck the competition between the US and China. In the end, AI companies in the US and other democracies must have better models than those in China if we want to prevail. But we shouldn't hand the Chinese Communist Party technological advantages when we don't have to.

Three Dynamics of AI Development

Before I make my policy argument, I'm going to describe three basic dynamics of AI systems that it's crucial to understand:

Scaling laws. A property of AI — which I and my co-founders were among the first to document back when we worked at OpenAI — is that all else equal, scaling up the training of AI systems leads to smoothly better results on a range of cognitive tasks, across the board. So, for example, a $1M model might solve 20% of important coding tasks, a $10M might solve 40%, $100M might solve 60%, and so on. These differences tend to have huge implications in practice — another factor of 10 may correspond to the difference between an undergraduate and PhD skill level — and thus companies are investing heavily in training these models.

Shifting the curve. The field is constantly coming up with ideas, large and small, that make things more effective or efficient: it could be an improvement to the architecture of the model (a tweak to the basic Transformer architecture that all of today's models use) or simply a way of running the model more efficiently on the underlying hardware. New generations of hardware also have the same effect. What this typically does is shift the curve: if the innovation is a 2x "compute multiplier" (CM), then it allows you to get 40% on a coding task for $5M instead of $10M; or 60% for $50M instead of $100M, etc. Every frontier AI company regularly discovers many of these CM's: frequently small ones (~1.2x), sometimes medium-sized ones (~2x), and every once in a while very large ones (~10x). Because the value of having a more intelligent system is so high, this shifting of the curve typically causes companies to spend more, not less, on training models: the gains in cost efficiency end up entirely devoted to training smarter models, limited only by the company's financial resources. People are naturally attracted to the idea that "first something is expensive, then it gets cheaper" — as if AI is a single thing of constant quality, and when it gets cheaper, we'll use fewer chips to train it. But what's important is the scaling curve: when it shifts, we simply

traverse it faster, because the value of what's at the end of the curve is so high. In 2020, my team published a paper suggesting that the shift in the curve due to algorithmic progress is ~1.68x/year. That has probably sped up significantly since; it also doesn't take efficiency and hardware into account. I'd guess the number today is maybe ~4x/year. Another estimate is here. Shifts in the training curve also shift the inference curve, and as a result large decreases in price holding constant the quality of model have been occurring for years. For instance, Claude 3.5 Sonnet which was released 15 months later than the original GPT-4 outscores GPT-4 on almost all benchmarks, while having a ~10x lower API price.

Shifting the paradigm. Every once in a while, the underlying thing that is being scaled changes a bit, or a new type of scaling is added to the training process. From 2020-2023, the main thing being scaled was pretrained models: models trained on increasing amounts of internet text with a tiny bit of other training on top. In 2024, the idea of using reinforcement learning (RL) to train models to generate chains of thought has become a new focus of scaling. Anthropic, DeepSeek, and many other companies (perhaps most notably OpenAI who released their o1-preview model in September) have found that this training greatly increases performance on certain select, objectively measurable tasks like math, coding competitions, and on reasoning that resembles these tasks. This new paradigm involves starting with the ordinary type of pretrained models, and then as a second stage using RL to add the reasoning skills. Importantly, because this type of RL is new, we are still very early on the scaling curve: the amount being spent on the second, RL stage is small for all players. Spending $1M instead of $0.1M is enough to get huge gains. Companies are now working very quickly to scale up the second stage to hundreds of millions and billions, but it's crucial to understand that we're at a unique "crossover point" where there is a powerful new paradigm that is early on the scaling curve and therefore can make big gains quickly.

DeepSeek's Models

The three dynamics above can help us understand DeepSeek's recent releases. About a month ago, DeepSeek released a model called "DeepSeek-V3" that was a pure pretrained model3 — the first stage described in #3 above. Then last week, they released "R1", which added a second stage. It's not possible to determine everything about these models from the outside, but the following is my best understanding of the two releases.

DeepSeek-V3 was actually the real innovation and what should have made people take notice a month ago (we certainly did). As a pretrained model, it appears to come close to the performance of4 state of the art US models on some important tasks, while costing substantially less to train (although, we find that Claude 3.5 Sonnet in particular remains much better on some other key tasks, such as real-world coding). DeepSeek's team did this via some genuine and impressive innovations, mostly focused on engineering efficiency. There were particularly innovative improvements in the management of an aspect called the "Key-Value cache", and in enabling a method called "mixture of experts" to be pushed further than it had before.

However, it's important to look closer:

DeepSeek does not "do for $6M5 what cost US AI companies billions". I can only speak for Anthropic, but Claude 3.5 Sonnet is a mid-sized model that cost a few $10M's to train (I won't give an exact number). Also, 3.5 Sonnet was not trained in any way that involved a larger or more expensive model (contrary to some rumors). Sonnet's training was conducted 9-12 months ago,

and DeepSeek's model was trained in November/December, while Sonnet remains notably ahead in many internal and external evals. Thus, I think a fair statement is "DeepSeek produced a model close to the performance of US models 7-10 months older, for a good deal less cost (but not anywhere near the ratios people have suggested)".

If the historical trend of the cost curve decrease is ~4x per year, that means that in the ordinary course of business — in the normal trends of historical cost decreases like those that happened in 2023 and 2024 — we'd expect a model 3-4x cheaper than 3.5 Sonnet/GPT-4o around now. Since DeepSeek-V3 is worse than those US frontier models — let's say by ~2x on the scaling curve, which I think is quite generous to DeepSeek-V3 — that means it would be totally normal, totally "on trend", if DeepSeek-V3 training cost ~8x less than the current US models developed a year ago. I'm not going to give a number but it's clear from the previous bullet point that even if you take DeepSeek's training cost at face value, they are on-trend at best and probably not even that. For example this is less steep than the original GPT-4 to Claude 3.5 Sonnet inference price differential (10x), and 3.5 Sonnet is a better model than GPT-4. All of this is to say that DeepSeek-V3 is not a unique breakthrough or something that fundamentally changes the economics of LLM's; it's an expected point on an ongoing cost reduction curve. What's different this time is that the company that was first to demonstrate the expected cost reductions was Chinese. This has never happened before and is geopolitically significant. However, US companies will soon follow suit — and they won't do this by copying DeepSeek, but because they too are achieving the usual trend in cost reduction.

Both DeepSeek and US AI companies have much more money and many more chips than they used to train their headline models. The extra chips are used for R&D to develop the ideas behind the model, and sometimes to train larger models that are not yet ready (or that needed more than one try to get right). It's been reported — we can't be certain it is true — that DeepSeek actually had 50,000 Hopper generation chips6, which I'd guess is within a factor ~2-3x of what the major US AI companies have (for example, it's 2-3x less than the xAI "Colossus" cluster)7. Those 50,000 Hopper chips cost on the order of ~$1B. Thus, DeepSeek's total spend as a company (as distinct from spend to train an individual model) is not vastly different from US AI labs.

It's worth noting that the "scaling curve" analysis is a bit oversimplified, because models are somewhat differentiated and have different strengths and weaknesses; the scaling curve numbers are a crude average that ignores a lot of details. I can only speak to Anthropic's models, but as I've hinted at above, Claude is extremely good at coding and at having a well-designed style of interaction with people (many people use it for personal advice or support). On these and some additional tasks, there's just no comparison with DeepSeek. These factors don't appear in the scaling numbers.

R1, which is the model that was released last week and which triggered an explosion of public attention (including a ~17% decrease in Nvidia's stock price), is much less interesting from an innovation or engineering perspective than V3. It adds the second phase of training — reinforcement learning, described in #3 in the previous section — and essentially replicates what OpenAI has done with o1 (they appear to be at similar scale with similar results)8. However, because we are on the early part of the scaling curve, it's possible for several companies to produce models of this type, as long as they're starting from a strong pretrained model. Producing R1 given V3 was probably very cheap. We're therefore at an interesting "crossover point", where it

is temporarily the case that several companies can produce good reasoning models. This will rapidly cease to be true as everyone moves further up the scaling curve on these models.

Export Controls

All of this is just a preamble to my main topic of interest: the export controls on chips to China. In light of the above facts, I see the situation as follows:

There is an ongoing trend where companies spend more and more on training powerful AI models, even as the curve is periodically shifted and the cost of training a given level of model intelligence declines rapidly. It's just that the economic value of training more and more intelligent models is so great that any cost gains are more than eaten up almost immediately — they're poured back into making even smarter models for the same huge cost we were originally planning to spend. To the extent that US labs haven't already discovered them, the efficiency innovations DeepSeek developed will soon be applied by both US and Chinese labs to train multi-billion dollar models. These will perform better than the multi-billion models they were previously planning to train — but they'll still spend multi-billions. That number will continue going up, until we reach AI that is smarter than almost all humans at almost all things.

Making AI that is smarter than almost all humans at almost all things will require millions of chips, tens of billions of dollars (at least), and is most likely to happen in 2026-2027. DeepSeek's releases don't change this, because they're roughly on the expected cost reduction curve that has always been factored into these calculations.

This means that in 2026-2027 we could end up in one of two starkly different worlds. In the US, multiple companies will definitely have the required millions of chips (at the cost of tens of billions of dollars). The question is whether China will also be able to get millions of chips9.

If they can, we'll live in a bipolar world, where both the US and China have powerful AI models that will cause extremely rapid advances in science and technology — what I've called "countries of geniuses in a datacenter". A bipolar world would not necessarily be balanced indefinitely. Even if the US and China were at parity in AI systems, it seems likely that China could direct more talent, capital, and focus to military applications of the technology. Combined with its large industrial base and military-strategic advantages, this could help China take a commanding lead on the global stage, not just for AI but for everything.

If China can't get millions of chips, we'll (at least temporarily) live in a unipolar world, where only the US and its allies have these models. It's unclear whether the unipolar world will last, but there's at least the possibility that, because AI systems can eventually help make even smarter AI systems, a temporary lead could be parlayed into a durable advantage10. Thus, in this world, the US and its allies might take a commanding and long-lasting lead on the global stage.

Well-enforced export controls11 are the only thing that can prevent China from getting millions of chips, and are therefore the most important determinant of whether we end up in a unipolar or bipolar world.

The performance of DeepSeek does not mean the export controls failed. As I stated above, DeepSeek had a moderate-to-large number of chips, so it's not surprising that they were able to develop and then train a powerful model. They were not substantially more resource-constrained

than US AI companies, and the export controls were not the main factor causing them to "innovate". They are simply very talented engineers and show why China is a serious competitor to the US.

DeepSeek also does not show that China can always obtain the chips it needs via smuggling, or that the controls always have loopholes. I don't believe the export controls were ever designed to prevent China from getting a few tens of thousands of chips. $1B of economic activity can be hidden, but it's hard to hide $100B or even $10B. A million chips may also be physically difficult to smuggle. It's also instructive to look at the chips DeepSeek is currently reported to have. This is a mix of H100's, H800's, and H20's, according to SemiAnalysis, adding up to 50k total. H100's have been banned under the export controls since their release, so if DeepSeek has any they must have been smuggled (note that Nvidia has stated that DeepSeek's advances are "fully export control compliant"). H800's were allowed under the initial round of 2022 export controls, but were banned in Oct 2023 when the controls were updated, so these were probably shipped before the ban. H20's are less efficient for training and more efficient for sampling — and are still allowed, although I think they should be banned. All of that is to say that it appears that a substantial fraction of DeepSeek's AI chip fleet consists of chips that haven't been banned (but should be); chips that were shipped before they were banned; and some that seem very likely to have been smuggled. This shows that the export controls are actually working and adapting: loopholes are being closed; otherwise, they would likely have a full fleet of top-of-the-line H100's. If we can close them fast enough, we may be able to prevent China from getting millions of chips, increasing the likelihood of a unipolar world with the US ahead.

Given my focus on export controls and US national security, I want to be clear on one thing. I don't see DeepSeek themselves as adversaries and the point isn't to target them in particular. In interviews they've done, they seem like smart, curious researchers who just want to make useful technology.

But they're beholden to an authoritarian government that has committed human rights violations, has behaved aggressively on the world stage, and will be far more unfettered in these actions if they're able to match the US in AI. Export controls are one of our most powerful tools for preventing this, and the idea that the technology getting more powerful, having more bang for the buck, is a reason to lift our export controls makes no sense at all.

Footnotes

1I'm not taking any position on reports of distillation from Western models in this essay. Here, I'll just take DeepSeek at their word that they trained it the way they said in the paper. ↩

2Incidentally, I think the release of the DeepSeek models is clearly not bad for Nvidia, and that a double-digit (~17%) drop in their stock in reaction to this was baffling. The case for this release not being bad for Nvidia is even clearer than it not being bad for AI companies. But my main goal in this piece is to defend export control policies. ↩

3To be completely precise, it was a pretrained model with the tiny amount of RL training typical of models before the reasoning paradigm shift. ↩

4It is stronger on some very narrow tasks. ↩

5This is the number quoted in DeepSeek's paper — I am taking it at face value, and not doubting this part of it, only the comparison to US company model training costs, and the distinction between the cost to train a specific model (which is the $6M) and the overall cost of R&D (which is much higher). However we also can't be completely sure of the $6M — model size is verifiable but other aspects like quantity of tokens are not. ↩

6In some interviews I said they had "50,000 H100's" which was a subtly incorrect summary of the reporting and which I want to correct here. By far the best known "Hopper chip" is the H100 (which is what I assumed was being referred to), but Hopper also includes H800's, and H20's, and DeepSeek is reported to have a mix of all three, adding up to 50,000. That doesn't change the situation much, but it's worth correcting. I'll discuss the H800 and H20 more when I talk about export controls. ↩

7Note: I expect this gap to grow greatly on the next generation of clusters, because of export controls. ↩

8I suspect one of the principal reasons R1 gathered so much attention is that it was the first model to show the user the chain-of-thought reasoning that the model exhibits (OpenAI's o1 only shows the final answer). DeepSeek showed that users find this interesting. To be clear this is a user interface choice and is not related to the model itself. ↩

9Note that China's own chips won't be able to compete with US-made chips any time soon. As I wrote in my recent op-ed with Matt Pottinger: "China's best AI chips, the Huawei Ascend series, are substantially less capable than the leading chip made by U.S.-based Nvidia. China also may not have the production capacity to keep pace with growing demand. There's not a single noteworthy cluster of Huawei Ascend chips outside China today, suggesting that China is struggling to meet its domestic needs...". ↩

10To be clear, the goal here is not to deny China or any other authoritarian country the immense benefits in science, medicine, quality of life, etc. that come from very powerful AI systems. Everyone should be able to benefit from AI. The goal is to prevent them from gaining military dominance. ↩

11Several links, as there have been several rounds. To cover some of the major actions: One, two, three, four. ↩