Dario Amodei wants two very different things to be true at once. He wants the world to believe that cheap, open-weight AI is a public good that deserves to spread freely. And he wants Washington to choke off China's access to the chips and techniques that would let it build the same thing. Put those two positions side by side, and what emerges isn't a coherent safety philosophy — it's a blueprint for a two-tier AI order, generous at home and restrictive abroad.

Clarifying What He Didn't Say

Amodei spent this past week trying to correct the record. After Anthropic's conspicuous absence from an industry letter — backed by Nvidia, Microsoft, Google and dozens of others — urging the White House not to restrict open-weight models, speculation grew that Anthropic quietly wanted the technology strangled to protect its own premium subscription business. Amodei pushed back directly, publishing a formal position statement insisting Anthropic has never called for banning open-weight models as a category, and describing those without dangerous capabilities as valuable to businesses, developers and researchers.

That should have settled the matter. Instead, it revealed something more interesting: Amodei isn't defending open models on principle so much as carving out an exception for one country.

Three Policies, One Target

Rather than a ban, Amodei proposed three narrower measures: tighter export controls on advanced chips and chipmaking equipment bound for authoritarian governments, a crackdown on what he calls "industrial-scale distillation" — the practice of training a smaller model on a larger one's outputs — and mandatory safety testing for sufficiently capable models, regardless of whether they're open or closed.

On paper, that last point is genuinely even-handed. The first two are not. Both are aimed squarely at Beijing. Anthropic has gone so far as to accuse Alibaba's Qwen lab of running a massive distillation campaign against Claude using tens of thousands of fake accounts, and Amodei has framed the entire debate around one anxiety: that without access to American chips, China cannot train models more capable than America's — and that distillation is the workaround that threatens to erase that gap entirely.

The timing is not incidental. The push gained urgency after Moonshot's Kimi K3, a Chinese open-weight model, delivered near-frontier performance at a fraction of the expected cost — a result that looked, to critics, like exactly the kind of "free rider" scenario Amodei has warned about.

Safety Rhetoric, Strategic Substance

Amodei frames all of this as a security question, not a competitive one: authoritarian governments developing more capable AI than the United States, and the misuse of powerful models for cyberattacks or biological threats, are what he calls the two risks that a blanket ban would fail to address. Fair enough, as far as it goes. But notice what his preferred remedies actually do. Chip controls and distillation crackdowns don't primarily stop misuse — they primarily slow down a competitor. A model trained through distillation is not inherently more dangerous than one trained from scratch; it is simply cheaper to build. Treating the technique itself as the threat, rather than any dangerous capability that might result from it, quietly substitutes a national-primacy argument for a safety one.

That substitution is easy to miss because the language of safety is so seductive. Everyone can agree that biological weapons uplift or infrastructure-hacking capability should be tested for and constrained. Far fewer people would sign on so readily to a policy whose plainer description is: keep the gap between American and Chinese AI capability as wide as possible,

indefinitely.

The View From Outside the US-China Binary

There is also a real irony sitting underneath the export-control push. Critics — including some inside the industry — have pointed out that cutting China off from advanced chips has, over the past several years, mostly accelerated Chinese investment in domestic chip production rather than stalled its AI progress. DeepSeek's leadership has said as much, arguing the restrictions are pushing Chinese labs toward closer cooperation with domestic chipmakers. If that's true, then the policy Amodei is asking for may not even achieve the primacy it's designed to protect — it may just guarantee that China builds its own stack, on its own terms, faster.

For countries watching from outside the US-China rivalry — including fast-growing AI ecosystems like India's — the stakes are not abstract. A world where safety testing applies evenly to all sufficiently capable models is one thing. A world where "safety" quietly becomes shorthand for "whoever isn't Chinese" is another. The latter doesn't just affect Beijing's national AI champions; it puts every researcher, student and startup operating anywhere near that boundary in the position of being pre-judged by geography rather than by what they actually build.

What a More Consistent Position Would Look Like

None of this means Chinese state misuse of AI is a fantasy, or that export-control policy is inherently illegitimate. Authoritarian governments really can and do use advanced digital tools for surveillance and coercion. But a policy that is actually about dangerous capabilities,

rather than about national rank, would look different from the one Amodei is proposing. It would define specific, testable red lines — the ability to generate actionable bioweapons guidance, or autonomously exploit critical infrastructure — and apply them uniformly to any model, wherever it was trained and by whomever. It would treat distillation as a technique to be evaluated by its output, not banned by its country of origin. And it would pursue verifiable international testing regimes rather than betting everything on unilateral chip chokepoints that history suggests are just as likely to backfire.

Amodei is not wrong that AI safety requires real constraints. He is wrong to let “safety” absorb what is, in substantial part, an argument for permanent American technological advantage. The two aims can overlap. They are not the same thing, and conflating them — deliberately or not — is exactly how policy ends up building walls where it should be building shared guardrails.