FACTOR

A Fire Drill, Not the Fire

The United States just proved it can switch off two of the world's most advanced AI models by nationality, at a few hours' notice. Whatever happens to this particular ban, that demonstration cannot be undone – and it is a warning we are lucky to get while the fix is still cheap.

By Vino Govender and Jacques Jurgens

On Friday 12 June 2026, at 17:21 Eastern Time, the US government ordered Anthropic to cut its two most capable models, Fable 5 and Mythos 5, off from every foreign national on earth – not only abroad, but inside the United States, the company's own engineers included. Within hours Anthropic disabled both models worldwide; a ban it could not enforce by nationality became, in practice, a shutdown for everyone. The telling detail is small: this was not a border closing but a nationality test, and it reached foreign nationals standing on American soil.

The ban will probably be lifted soon – both sides say they want it resolved. But do not mistake the reversal for a reprieve. For two years, the idea that one government could revoke the world's access to frontier AI by nationality lived in scenario papers. Now it is a fact, with a date and a letter behind it. Lifting the ban un-bottles nothing: everyone has watched the off-switch work, and no one can go back to believing it merely theoretical. That is why this is a fire drill worth heeding – the rehearsal is over, the capability is proven, and what remains is what we do about it.

The danger is real and on the record. Ten days earlier, an executive order had already begun treating frontier models as strategic

assets – a government look at the most capable systems before release, classified benchmarks at the NSA, a hand in choosing which "trusted partners" get early access. It is voluntary in form. But it lays the architecture, and architecture outlasts intentions.

Now look at the remedy. Restricting a dangerous capability to Americans contains almost nothing as a safety measure: the hazard is still loose, merely loose among the holders of one passport. That is not the shape of a safety measure; it is the shape of an export control, which has never been about whether a dangerous thing exists, only about who may hold it. The trigger is disputed – Amazon's researchers found a jailbreak, *Semafor* reports a suspicion that a China-linked group accessed Mythos, and the former AI czar David Sacks says Anthropic was asked to fix or pull the model and refused – but across every version the impulse is to keep the capability out of foreign hands. And notice the breadth. Guarding against a Chinese breach would not require cutting off Britain, India, Canada and Anthropic's own foreign-national engineers. That surplus – denial reaching allies and the company's own staff, far past any plausible adversary – is the tell. Whatever set it off, the remedy reached far past breach-response: the frontier treated as a national asset, kept for one's own and withheld from the rest.

And the posture is visible in the open. Months earlier, the *Financial Times* reported, the same government was not containing Mythos's cyber capability but wielding it: the NSA had embedded Anthropic engineers to adapt the model for offensive operations against foreign networks, even as the Pentagon branded the company a security risk in court. A capability its maker called too dangerous to release, denied to the world on safety grounds, was reportedly being adapted for offensive use by the United States – withheld from others not so it cannot be used, but so that only one side can use it. The episode was never about a capability too dangerous to exist. It was about one too valuable to share.

None of this makes Anthropic a simple victim, either. The company spent a year describing Mythos in cyberweapon terms and

lobbying for tighter export controls; when the government reached for exactly that framing, it found the predicate already written. The words you use to sell power are the words that will be used to seize it.

Here is why that posture is unlikely to hold – and why the news is better than it looks. Oil through a strait is rival and scarce: a barrel that goes to you does not go to me. Artificial intelligence is the opposite. A model weight copied for you costs me nothing; once built, the marginal cost of replicating the intelligence rounds to zero. You can close a strait. You cannot durably close off a non-rival good – you can only try to impose artificial scarcity on something that is naturally abundant. Capabilities replicate in months, not years: by Epoch AI's capability index, the best open-weight models – now mostly Chinese – trail the closed frontier by about four months in early 2026 (about a year by a broader, compute-based measure), with the gap sometimes closing to nothing. The frontier is a lead, not a moat.

There is one rival exception, and it deserves naming: compute. The chips, the fabs, the data centres and the power behind them are physical and controllable, which is why America's real denial campaign has always been aimed at hardware, not models. That is the firmest ground the chokepoint stands on. But even it leaks – the compute needed for a given capability keeps falling, and running a model is far more distributed than training one – and its sheer expense is the strongest argument for pooling rather than hoarding. A mid-sized country cannot out-build a superpower's data centres alone. A federation of them can.

None of this means openness is safe, and it would be dishonest to pretend it does. The same diffusion that spreads benefit spreads hazard: the bioweapon capabilities that genuinely frighten people do not stay bottled either, and wide access to a pathogen-design tool is not obviously offset by wide access to its defensive uses. That is a real and unsolved problem. But nationality-gating one company's model does nothing to solve it – the capability is available elsewhere

and leaks regardless – so the foreigner-shaped hole in the defence merely decides which nationals hold a danger everyone reaches anyway. The only instruments with any purchase on a diffuse hazard are shared ones. Which means the real choice was never safety versus abundance. It is enclosure versus common governance – and enclosure does not even buy the safety it promises.

So the instinct now stirring in capitals from London to Delhi – to pull away from dependence on the United States – is sound, even if "decouple" is the wrong word. Decouple towards what? Chinese models swap one jurisdiction's politics for another's, with fewer checks. Going it alone buys sovereignty at the price of the frontier. The answer is not separation but interoperability: not an empire with one capital and one off-switch, but a federation of many centres of capability, bound by common standards, shared safety institutions no single state controls, and the open movement of talent and ideas.

None of that is easy, and the hardest case is the leader's: the United States has a real incentive to withhold the interoperability that would erode its lead, so the argument to it has to be made on its own interest. Hoarding sometimes pays, and a renewed short lead can matter. But this gate is a bad trade. It buys little time against a capability that diffuses regardless, at the cost of allies, of the talent it walls out – the ban's own targets were foreign-born engineers – and of the standing to lead the institutions that will actually govern this technology. The countries cut off this month are the ones designing around you next month.

The most-discussed AI forecast of the past year imagines two futures spun from identical starting conditions: in one the race runs to catastrophe, in the other people keep control. The difference is not technology. It is whether we believe the race is inevitable – a belief that, held widely enough, makes itself true. Enclosure is the harder thing to hold, not the easier one; it requires holding back a tide. A fire drill is a gift: the lesson delivered while the stakes are still low enough to absorb. We have had ours. What comes next is not

something that happens to us. On this, more than almost anything in front of us, it is something we choose.

Sources are linked inline where this runs online. The trigger for the ban is contested and may be multi-causal (Amazon's researchers per the Wall Street Journal; suspected Chinese access per Semafor, unconfirmed; the administration's account via David Sacks, its former AI czar and an aligned voice). The "minor" versus "serious" reading of the jailbreak is a live dispute between interested parties; the NSA's offensive use of Mythos rests on Financial Times reporting the agency has not confirmed; and the open-weight lag figures come from Epoch AI's Capabilities Index (about four months in early 2026).