FACTOR

A Fire Drill, Not the Fire

The United States has shown 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 – and it may already be lifted by the time you read this – the demonstration cannot be undone. That is not the disaster. It is the warning we are lucky to get while the fix is still cheap.*

By Vino Govender and Jacques Jurgens

For two years, the possibility that a single government could revoke the world's access to frontier artificial intelligence – not by geography but by nationality, and at will – lived mostly in scenario papers and thought experiments. In June 2026 it stopped being hypothetical.

On Friday 12 June 2026, at 17:21 Eastern Time, the AI company Anthropic received an export-control directive from the US government ordering it to cut off its two most capable models – Fable 5 and Mythos 5 – from every foreign national on earth. Not only abroad: inside the United States too, the company's own engineers included. Within hours Anthropic disabled both models for every customer worldwide, saying that was the only way to ensure compliance – a ban aimed at foreign nationals, which no provider can enforce perfectly by nationality, becomes in practice a shutdown for everyone. For a weekend, anyone who typed a query into the frontier of artificial intelligence – in Lagos or London or Lima, and in San Francisco besides – got an error and a quiet downgrade to an older system.

The detail that matters most is small and easy to miss. This was not a border closing. It was a nationality test – the order reached foreign nationals standing on American soil, including the very people who help build these models. Gating on nationality is not, in itself, new: US "deemed export" rules have for decades treated handing controlled technology to a foreign national inside the country as an export to their homeland. What is new is the object. That logic, built for the transfer of technology someone could carry away and rebuild, was here pointed at a foreign national's mere use of a consumer model – the act of typing a question into a chatbot reclassified as an export.

It is tempting to read this as the moment the mask came off – the day the unipolar order revealed that the country which built the frontier can hold the rest of the world's access to it hostage. That reading is half right. The half it gets wrong is the more important half.

Because this was the cheapest possible version of the lesson, and almost certainly a temporary one. The directive was disputed within hours and remains so. Anthropic says the triggering issue amounted to a model finding a handful of minor, already-known software vulnerabilities that other public models can find too, and calls the response disproportionate; the administration and the partner who reported the flaw insist it was serious – a genuine bypass of a safeguard on a system Anthropic itself had described as a kind of cyberweapon. An outsider cannot adjudicate which is right. What both sides agree on is that they want it over quickly: Anthropic says it is working to restore access, and the administration, which issued the order, says it hopes to lift the controls as soon as the flaw is fixed. The model at the centre of it had been public for three days. By the time you read this, the controls may already be gone.

But do not mistake the reversal for a reprieve. What changed in June 2026 was not that access was lost; it was that the loss became real. Before, the revocability of the frontier by nationality was a claim about what governments could in principle do. Now it is a

fact about what one government has done, with a date and a letter behind it. Lifting the ban un-bottles nothing: every government, company and builder on earth has now watched the off-switch work, and no one can go back to believing it merely theoretical. The cat is out of the bag. That is exactly why this is a fire drill worth heeding rather than an incident to file and forget – the rehearsal is over, the capability is proven, and what remains is the choice of what to do about it.

A fire drill is only useful, though, if you take the fire seriously. So before the hopeful part, the honest part.

First, the honest part

The danger here is real, and it is on the record. Access to the most capable AI is now demonstrably revocable by nationality, on national-security grounds, at a few hours' notice. And this was not an isolated spasm. Ten days earlier, the administration's executive order of 2 June, "Promoting Advanced Artificial Intelligence Innovation and Security," had already begun treating a frontier model as a strategic asset rather than an ordinary product – granting the government a pre-release look at the most capable systems, building classified evaluation benchmarks at the NSA, and letting the government help choose which "trusted partners" get early access. The order is voluntary in form and explicitly disclaims any licensing regime. But it lays the architecture, and architecture outlasts intentions.

And the way this particular switch got flipped is genuinely murky, with at least three accounts in circulation. The *Wall Street Journal* and *The Information* report that the trigger was Amazon – Anthropic's largest investor, a board presence, its cloud host, and a frontier competitor of its own – whose researchers found a way to make the public model assist with cyberattacks, after which chief executive Andy Jassy carried the finding to Treasury Secretary Scott Bessent and the order followed within days. The administration's own account came from David Sacks – its former AI czar, an aligned

voice rather than a neutral one – who named the source only as "a highly credible trusted partner," a phrase that quietly launders a commercially conflicted rival into a neutral auditor. But *Semafor* reports a different and more serious trigger: a suspicion, sourced to a single person familiar with the matter, that a China-linked group had gained access to Mythos itself. And Sacks adds that officials asked Anthropic to fix the flaw or pull the model, and that Anthropic refused. These accounts do not all fit together, and an outsider cannot rank them. But notice what the spread does to the easy story. If the trigger was a competitor's phone call, the instrument looks arbitrary and aimable; if it was suspected exfiltration to Beijing, it looks like a real and specific security response. We do not yet know which – and the one thing common to every version is the impulse to keep the capability out of foreign hands.

And look closely at the remedy itself, because it is the most revealing thing in the episode. Set aside the argument over how dangerous the jailbreak was, and ask what restricting the model to Americans actually accomplishes as a matter of safety. If the capability is genuinely dangerous – a real bypass that turns the model into a cyberweapon – then leaving it with hundreds of millions of Americans, among them whoever is malicious, careless, compromised, or working quietly for a foreign service, contains almost nothing as a general safety measure. The hazard is still loose; it is merely loose among the holders of one passport – and a state actor determined to have the capability can reach it through other models that do much the same thing. Containing a dangerous capability means patching the flaw or withholding the model from everyone until it is patched. Here the agency is contested: the administration says it asked Anthropic to fix or remove the model and was refused, after which Anthropic – unable to filter by nationality in real time – chose a worldwide blackout to comply. But strip the dispute back to the instrument the government actually reached for, and it was neither a patch nor a blanket suspension. It was a line drawn at nationality. That is not the shape of a general safety measure. It is the shape

of an export control – and export controls have never been about whether a dangerous thing exists, only about who is permitted to hold it.

Read the instrument this way and the public story strains. The action is wired as denial – keep the capability, withhold it from others – while the case made for it in public is spoken in the language of safety. The severity dispute cannot close the gap between them: a minor flaw makes export-controlling a single model over it close to theatre, a serious one makes denying it to foreigners while leaving it with Americans no containment at all, and even the most coherent trigger – *Semafor*'s report of suspected exfiltration to a China-linked group – points only to nonproliferation, not to a capability too dangerous for anyone to hold.

But there is a move the shape of the remedy cannot make on its own, and it is the one that matters. Denial to an adversary and denial to the world are not the same posture, and this was the second. Guarding against a suspected Chinese breach does not require cutting off Britain, India, Canada and the company's own foreign-national engineers. That reach – past any plausible adversary, into allied nations and Anthropic's own staff – is the surplus that breach-response would never need, and the surplus is the tell. A defensive gate is narrow by design; this one was drawn at nationality itself. What that betrays is not the prevention of theft but the treatment of the frontier as a national asset, to be kept for one's own and withheld from everyone else – a premise that needed no new doctrine, only a single Commerce Department letter to make it operative.

And this is not merely an inference from the shape of the remedy; the posture is visible in the open. Months before the Fable ban, according to the *Financial Times*, the same government was not containing Mythos's cyber capability but wielding it – the NSA had embedded around half a dozen Anthropic engineers to adapt the model for offensive operations against foreign networks, even as the Pentagon was branding the company a national-security risk in court. Whether Mythos is yet in live campaigns is unconfirmed; that

it is being readied to attack adversaries' systems is not. Hold the whole picture together: a capability its own maker called too dangerous to release, denied to the rest of the world on safety grounds, and turned by the United States into a weapon – withheld from others not so that it cannot be used, but so that only one side can use it, and so that the targets cannot harden against it. The episode was never about a capability too dangerous to exist. It was about one too valuable to share.

The NSA programme is the sharpest instance of an appetite that runs broader. The government wants the capability itself, unrestricted: for months the Pentagon pressed Anthropic for military use of Claude with no limits – "all lawful purposes," in Defence Secretary Pete Hegseth's words – and when the company drew red lines against autonomous weapons and mass surveillance, the department branded it a "supply chain risk." A federal judge blocked that as retaliatory, invoking the "Orwellian notion" that a company could be branded an adversary for disagreeing with its government; a federal appeals court declined to block the designation while the case proceeds, as the government pressed for deference during an ongoing military conflict. The administration says that fight and this one are unrelated, and maybe they are – but the structural echo holds either way: its stated fear back then was that Anthropic could switch a model off mid-operation, and the Fable ban is the mirror image, proof the government can do the same to the entire world. Both sides are now arming around one question – who can cut whom off.

That is the honest part, and it is the documented part – company statements, court filings, and the reporting around them. The question is what follows from it.

A FORK, NOT A FATE

Here is where the case turns from what has happened to what could.

The most-discussed AI forecast of the past year, *AI 2027*, is built around a single branch point. From one set of premises – the same

misalignment risks, the same concentration of power, the same arms-race pressure – it spins out two futures. In one, the race runs to its end and humanity loses control. In the other, written deliberately to begin from the identical situation, people stay in the driver's seat. The authors are explicit that the humane branch was reachable from exactly the conditions that produced the catastrophic one.

That structure is the whole point. The dystopia is not a law of physics. The zero-sum race is a coordination trap – it holds only so long as every player expects every other player to defect. Which means the real enemy of the good future is not the United States, or China, or any company. It is fatalism: the belief that the hoard-the-frontier race is inevitable, a belief that causes everyone to behave in precisely the ways that make it inevitable. Inevitability is not the weather here. It is the mechanism. And a future contingent on what decision-makers believe is a future contingent on argument.

AI WANTS TO BE ABUNDANT

The argument has physics on its side.

Oil and gas through the Strait of Hormuz – roughly a fifth of the world's seaborne supply, and a channel Iran has repeatedly shown it can throttle – is a rival, scarce good. A barrel that goes to you does not go to me. Artificial intelligence is the opposite kind of thing. It is non-rival. A model weight copied for you costs me nothing; once the system is built, the marginal cost of replicating the intelligence rounds to zero. Oil wants to be hoarded. Software wants to spread.

This is the seam where the chokepoint analogy, for all its intuitive grip, finally tears. You can close a strait because oil is physically scarce. You cannot, over any meaningful span of time, close off a non-rival good – you can only try to impose artificial scarcity on something that is naturally abundant. Nationality-gating a model is exactly that attempt. And the record on imposing artificial scarcity on software is not kind to the gatekeeper.

There is one rival exception, and honesty requires naming it: compute. The chips that train a frontier model, the fabs that print

them, the data centres and the power they draw are physical, scarce and genuinely controllable – which is why the United States' real and sustained denial campaign has always been aimed at hardware, not models, and why the Fable ban was startling for reaching up the stack to the model itself. If there is a durable chokepoint in artificial intelligence, it is here, in the silicon, and it is the firmest ground the empire model stands on. But even this moat is leakier than it looks, and in the end it cuts the other way. The compute needed for a given capability keeps falling as algorithms improve, so the frontier of one year runs on commodity hardware a few years later; and running a model, once it exists, is far less demanding and far more widely distributed than training it. Above all, compute is lumpy and capital-hungry in exactly the way that rewards pooling: a single mid-sized country cannot out-build a superpower's data centres alone, but a federation of them can. The rival layer, looked at squarely, is not a reason to abandon the federation – it is the reason to build one.

The evidence runs one way. Capabilities replicate in months, not years. Epoch AI's capability index puts the average gap between the best open-weight models and the closed frontier at around four months in early 2026 – about a year by a broader, compute-based measure – with the lead sometimes closing to nothing. The strongest open models now come out of Chinese labs – DeepSeek, Qwen, Kimi, GLM – sitting within months of the closed frontier even as US export controls bite, and *AI 2027*'s own racing scenario, with every control in force, projects the same shape: a gap counted in months, not decades. The frontier is a lead, not a moat.

And aggressive denial carries real costs on the very metric it is meant to protect. Export controls did not freeze China's programme; by most accounts they accelerated its domestic build-out of chips and models. A regime that gates by nationality, meanwhile, threatens the one input no compute cluster can replace: talent. The ban's own target was foreign-national employees – the people who build these systems. A lead can be spent defending itself.

Here is the move worth sitting with, and it turns on a distinction the easy version blurs. Two different things diffuse. Capability diffuses on its own – the weights, the methods, the know-how spread whether or not anyone wills it, and that is the part no gate durably holds. What that capability gets used for – shared abundance or concentrated harm – does not diffuse on its own. It is built, in the institutions and choices laid on top of it. So the gatekeeper and the optimist make mirror-image mistakes: one believes capability can be contained, the other believes good outcomes come free. The capability is coming either way. The outcome is the part we make – which means the dystopia is not the default that diffusion hands us, but one of the things we might build on top of it, and so is its opposite.

Here the essay owes its hardest concession, because the same physics cuts the other way too. If diffusion defeats the gate, it defeats containment – and that is bad news for safety, not only for monopoly. The cyber and bioweapon capabilities that genuinely frighten people do not stay bottled either, and the comfortable answer – that spreading the defensive uses of these models offsets spreading the offensive ones – assumes a symmetry that may not hold. For cyber it is at least arguable: defenders patch, defensive tooling scales, the exchange is roughly two-sided. For biology it is much weaker. Wide access to a pathogen-design capability is not obviously cancelled by wide access to the same model's defensive uses, because biological attack and defence are not symmetric – the attacker chooses the agent and the moment, and a single release can do catastrophic harm before any defence responds. Anyone who genuinely fears these capabilities should not be told that abundance takes care of it. It may not.

What this does not do is rescue the gate. Nationality-gating one company's model fails the bio problem exactly as it fails the cyber one: the capability is available from other models, leaks, and diffuses regardless, so a foreigner-shaped hole in the defence does nothing but decide which nationals hold a danger everyone will

eventually reach. The honest conclusion is narrower, and less comforting, than "openness is safe." It is that containment by one country does not work; that the genuinely dangerous capabilities – bio above all – are an unsolved problem no unilateral gate addresses; and that the only instruments with any purchase on a diffuse hazard are shared ones – common safety standards, monitoring that crosses borders, and restraint on the most dangerous capabilities agreed among the powers that hold them. None of that is guaranteed to work. But it is the only approach matched to the shape of the problem, which is also, not by coincidence, the argument for federation. You build shared safety institutions not because they are sure to hold, but because the alternative – each power hoarding and hoping – demonstrably will not.

Federation, not empire

The instinct to pull away from dependence on the United States is already being voiced – though more cautiously, and from more specific quarters, than a sweeping revolt. A British minister for AI, Kanishka Narayan, responded to the ban by arguing it should drive deeper investment in Britain's own AI industry. In India – lately added to Anthropic's trusted-partner programme, then abruptly cut off with everyone else, and the second-largest market for these tools – the episode reignited a sovereignty debate already running through the country's national AI mission. Across the EU and Canada, the same conversation stirred. The sceptics have a point worth keeping in view: talking about weaning a country off foreign AI is far easier than doing it, and most of these states are nowhere near the frontier. But the instinct is sound, even if the word for it isn't quite "decouple." Decouple towards *what*?

Towards Chinese models? That trades a jurisdiction whose courts can call a government action "Orwellian" out loud – as a US federal judge just did, ruling against her own government – for one with fewer such checks. Towards pure open-weights and full self-reliance? That buys real sovereignty at the price of the frontier.

Towards simply spreading your bets across vendors? That diversifies the risk without escaping the underlying exposure to a single jurisdiction's politics.

The answer is not separation. It is interoperability. Not an empire with one capital and one off-switch, but a federation: many centres of capability, bound by common standards, shared governance, and the open movement of ideas and people. In practice that means sovereign-but-interoperable compute and models, built to open standards so they connect rather than isolate; multilateral institutions for safety and standards that no single state controls; benefit-sharing so the global majority that supplied the data and the demand also shares in the returns; and – pointedly, against the grain of the ban – keeping talent borders open precisely where chokepoint politics wants them shut. None of this is easy, and it would be dishonest to pretend otherwise. Every item on that list is a hard problem of collective action, and the hardest case is the leader's: the United States has a real and rational incentive to withhold the interoperability that would erode its lead, which is exactly why the case to it has to be made on its own interest rather than on goodwill. A federation is not the path of least resistance. It is a thing built against the pull of the very advantages it asks the strongest player to share.

Star Trek's Federation, worth remembering, is not a hegemon. It is cooperative multipolarity with shared institutions and shared abundance. Not a fantasy of a world without rivalry, but a structure for managing rivalry without letting it harden into a single point of control. It is something you build, not something you wish for.

What to do about it

What that looks like in practice depends on who you are.

If you govern a country outside the United States: the lesson is not to panic-buy from the nearest alternative hegemon. The building blocks already exist, scattered and unfinished – India's national AI mission is committing more than a billion dollars over five years

(₹10,000-plus crore) to public compute and home-grown foundation models; Canada has Cohere, Germany has Aleph Alpha – and the honest fact is that none of them yet sits at the frontier. That gap is precisely the argument. A lone sovereign stack a tier behind is a weak hedge; the same investments pooled – shared compute, models built to open and interoperable standards, common safety and evaluation regimes, open talent flows – are how a scatter of dependent clients becomes a federation of peers with enough collective weight to matter. And keep your doors open to the researchers the chokepoint wants to wall out: they are the ones who close the gap.

If you build on this technology – a company, a developer, a startup: the field drew the right lesson within a day of the ban. Design for revocability. Model-agnostic architectures, open-weight fallbacks, resilience across jurisdictions. Assume any single frontier model can vanish on a Friday, and build so that your product does not vanish with it.

If you are the United States: the honest argument is not that hoarding never pays. It sometimes does, and a lead renewed every few months can be worth more than it looks – temporary monopolies have shaped whole eras. The argument is that this particular gate is a bad trade. What it buys is time, and not much of it, against a capability that diffuses regardless. What it costs is the alliance structure, the talent base – the ban's own target was foreign-national engineers – and 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, and a decade of hardware controls has mostly taught rivals to build their own. Openness here is not charity and not naivety; it is a wager that leading an interoperable order is worth more than renting a shrinking lead. That wager can be wrong. But "hoard and hope" is a wager too, and a worse-priced one.

If you are an AI lab: the lesson is double-edged, and Anthropic embodies both halves of it. It drew red lines against autonomous weapons and mass surveillance and defended them in court at

real cost – a template for the leverage frontier-builders hold over how their work is used, and too rarely exercise. But Anthropic also spent a year describing Mythos in cyberweapon terms and lobbying for tighter export controls, and when the government reached for precisely that framing, it found the predicate already written. If you describe your product as a weapon in every release, a government eventually takes you at your word. So pre-commitment cuts two ways: choose your red lines, but choose your language too – the words you use to sell power are the words that will be used to seize it.

The choosable future

It is worth being precise about what made Star Trek's future post-scarcity. It was not that someone won the race. It was that a civilisation chose to distribute the abundance its technology made possible, rather than enclose it.

That choice is the whole game, and the Fable 5 ban is the first time we have seen it posed this sharply. The chokepoint path is real; someone is willing to walk it; the off-switch works. But the technology underneath strains towards the opposite outcome – towards spreading, towards abundance, towards everyone. Enclosure is the harder thing to hold, not the easier one. It requires holding back a tide.

A fire drill is a gift. It is the catastrophe rendered survivable – the lesson delivered while the stakes are still low enough to absorb and the exits still close enough to find. We have had ours. The fire is real enough that we should be grateful for the drill. And the response – ours, our governments', our companies' – is what decides which branch of the story we are standing in.

The future is not something that happens to us. On this one, more than almost anything in front of us, it is something we choose.

Note on sourcing: this essay attributes its central facts in-text – Anthropic's own statement, the executive order of 2 June, the Wall Street Journal and The Information reporting on Amazon's role, Semafor's report of suspected Chinese access, the administration's account via David Sacks, the Pentagon litigation and the rulings by Judge Rita Lin and the DC Circuit, and the AI 2027 scenario. The trigger is contested and may be multi-causal: the Wall Street Journal and The Information point to Amazon's researchers; Semafor reports a suspicion that a China-linked group accessed Mythos, sourced to a single person familiar with the matter and explicitly unconfirmed; and Sacks – the administration's former AI czar, an aligned voice rather than a neutral one – says officials asked Anthropic to fix or remove the model and were refused. Treat the "minor" versus "serious" characterisations of the jailbreak, and all three trigger accounts, as live disputes between interested parties rather than settled fact. The argument does not turn on whether the ban remains in force when you read this; as of mid-June 2026 it was, with both sides signalling they wanted it lifted quickly – an outcome this essay treats as the likely one, and as confirmation of the "fire drill" frame rather than a refutation of it. The NSA's offensive use of Mythos rests on Financial Times reporting, with earlier corroboration from Axios, citing people familiar with the arrangement; the NSA declined to confirm or deny it and Anthropic declined to comment, so it is presented as attributed reporting, with the question of live deployment left open. That openness is not therefore safe is argued in the essay's own voice: the offence–defence asymmetry in biology in particular is an unsolved problem, not one this argument dissolves. The open-weight lag figures come from Epoch AI's Capabilities Index, which put the average gap at about four months in early 2026, with a broader compute-based measure nearer a year; that the leading open models are now Chinese reflects current open-weight rankings.