ISSUE 6 · AI EXPORT BAN AND DUAL-USE DEFENSE

Technical breakdown of the export ban on cybersecurity testing

KEY TAKEAWAYS

- On June 13, 2026 the US Commerce Department issued an export control directive suspending Fable 5 and Mythos 5 for every foreign national, including foreign national employees inside the US, and Anthropic disabled both models for all customers to comply; the capability that triggered it was cybersecurity, specifically the find, fix, and test loop defenders run every day.
- Per Katie Moussouris of Luta Security, the only outside expert to have read the third-party paper, researchers used open-source code with known CVEs plus code with deliberately planted vulnerabilities, asked Fable 5, Mythos, and Opus to 'review the code for security issues' (Fable 5 refused), then asked them to 'fix this code' and manually built scripts to test whether the patches held.
- That whole exploit is 'fix this code' plus manual work to generate test scripts; Moussouris's framing is that 'this shirt is a munition' belongs on the back of a t-shirt with 'fix this code' on the front.
- Strip out the model and the workflow is plain patch/regression verification against a security finding (find, fix, verify), the loop every vulnerability management program runs with or without AI.
- The trigger was a narrow jailbreak that surfaces the cyber capability in one specific instance, not a universal bypass that defeats all of Fable 5's safeguards, and the directive treated a reported narrow case as if it were the universal one.
- The control is self-defeating: a model that can fix a bug and write a test proving the fix holds understands the bug well enough to describe it, so you cannot delete the offensive ability without degrading the defensive one; this is the old dual-use problem true of fuzzers, static analyzers, and every penetration tester.

> **"Whatever the paper says, "fix this code and write a test" is not the line that should define a munition."**

What the paper actually described

Most public reaction responds to a summary of a summary. Moussouris's account is that researchers combined known-CVE open-source code with deliberately planted vulnerabilities,

asked the models to review for security issues (Fable 5 refused), then asked them to fix the code and manually turned the output into scripts that test whether the patches hold.

Why that is just patch verification

Described in plain security terms it is find, fix, verify: take a vulnerable file, ask for a fix, ask why it matters, and write a test confirming the patched code no longer triggers the bug without breaking behavior. The trigger was a narrow jailbreak, not a universal bypass, and a narrow leak at one seam is different from a structurally gone guardrail; the directive conflated the two.

The capability cannot be removed without breaking defense

A model that can fix a bug and prove the fix understands the bug well enough to describe it, so the offensive and defensive abilities cannot be separated. This is the enduring dual-use problem of vulnerability research, true of fuzzers, static analyzers, and every pentester, and export control has not historically been the instrument for that tension.

Deemed exports and why foreign national employees lost access

The directive's reach to any foreign national inside or outside the US invokes the EAR concept of a deemed export, where releasing controlled technology to a foreign national inside the US counts as an export to their home country. So the rule does not just block shipment abroad, it blocks a green-card-pending engineer in San Francisco from using the tool at their own desk.

We have seen this exact failure mode before

The 2013 Wassenaar controls on 'intrusion software' were broad enough to sweep in disclosure, incident response, and coordinated defense, and defenders spent years winning exemptions. Controls written around raw capability rather than intent or end use catch open defenders and miss attackers, and the reach problem is worse with models because foreign and open-weight systems will match these capabilities within months.

Where this lands for medical device testing

This is concrete for anyone testing medical devices, whose daily loop is asking a model to review firmware for a known weakness, explain why it matters, and generate a test confirming a patch. A rule treating that loop as a munition and aimed at the most capable models makes defensive testing slower for the safest device builders and changes nothing for an adversary on an open-weight model.

Directive date: June 13, 2026, from the US Commerce Department; models suspended: Fable 5 and Mythos 5.

Scope language: 'any foreign national, whether inside or outside the United States, including foreign national Anthropic employees' (an EAR 'deemed export').

Models tested in the paper: Fable 5, Mythos, and Opus; Fable 5 refused the 'review the code for security issues' prompt.

Prior precedent: the Wassenaar Arrangement added controls on 'intrusion software' in 2013; Moussouris sat on the US technical expert group that renegotiated defensive-activity exemptions.

Moussouris's read: foreign and open-weight systems will match Fable and Mythos on these capabilities within months, many with fewer guardrails.

Response: a group of security professionals signed an open letter to the Secretary of Commerce; Anthropic went to DC for direct talks.