THE WASHINGTON POST – 15 GIUGNO 2026

How not to defend against cyberthreats

Editoriale

In an attempt to strike at the danger that artificial intelligence poses to cybersecurity, the United States government late Friday took a swing at one of the country's leading frontier labs. Its punch landed, but the blow did nothing to address the threat it was aimed at.

The lab was Anthropic, the creator of some of the most advanced AI models on the market today, Claude Fable and Mythos 5. After being warned that Fable had been jailbroken to extract information useful for cyberattacks, officials met on Friday — and chose the bluntest tool available. At 5:21 p.m., the Trump administration sent the company an export-control order barring any foreign national from using the models.

Anthropic, unable to comply selectively, turned off both models worldwide. They had been publicly accessible for three days.

It's difficult to overstate how counterproductive that was. Cybersecurity has always been a race without a finish line. Attackers are constantly probing to find vulnerabilities to exploit in software, and defenders are always rushing to patch them before it's too late. AI speeds the race up considerably. But, crucially, it's also both the tool that finds the flaw and the tool for fixing it.

The problem is that while Anthropic's models are widely considered to be among the best available at this moment, its competitors are not far behind. OpenAI's GPT-5.5 is nearly on par with Anthropic's flagship in unearthing software vulnerabilities, according to the London-based AI Security Institute.

Much cheaper Chinese models, though still not on the frontier, just keep getting better. By whacking Anthropic, the government hasn't pushed back the cyberthreat. It just took an important tool out of the hands of defenders.

Worse, the mistake was predictable. Only 10 days earlier, the Trump administration signed an executive order about AI guardrails. The administration created a voluntary review process for frontier models, with an explicit promise that nothing in it would create a licensing or preclearance regime.

But at the first sign of trouble, it abandoned its own restraint and reached for the kill switch. The crackdown reportedly was prompted by conversations between government officials and Andy Jassy, CEO of Amazon, where Post owner Jeff Bezos is executive chairman.

Anthropic has done itself no favors by hyping its latest models as fundamentally game-changing rather than simply ahead of the curve. That is defensible as marketing. But now the company is acting surprised that bureaucrats took the pitch at face value.

Technology waits for no one, neither the regulators nor the labs. Anthropic's models are not the unique source of the cyberthreat. The only answer to dangerous AI is more of it in the right hands.