Written by **Paul Dragu** on June 25, 2026

Attack With AI, Defend With AI: Intel Agencies Warn of Cyber Attacks Within Months

Western intelligence agencies warn that the world's top artificial intelligence models are becoming so advanced that in a few months they'll pose serious cybersecurity risks to the United States.

da-kuk/iStock/Getty Images Plus

"(AI) is rapidly transforming cyber risk, and we must act swiftly to remain ahead," says a statement published this week by the Five Eyes intel coalition. AI "lowers barriers for malicious actors and increases the speed and complexity of attacks, shrinking the window between vulnerability discovery and exploitation ever more quickly."

The Five Eyes is an intelligence alliance made up of Australia, Canada, New Zealand, the United Kingdom, and the United States. They worked closely on worldwide censorship campaigns during Covid mania.

The warning, published on Monday, focuses on the threat to businesses. But it's widely believed that the risks apply to governments as well. AI companies are already building models especially designed for cybersecurity.

The threat should be taken seriously, we are told. From the press release:

> Cyber risk can no longer be treated as a purely technical issue. This is a core business risk and leadership responsibility. Boards and executives should ensure cyber resilience is in place and works under pressure. It is not enough to have controls. Leaders must be confident those controls will perform during a real incident.

More AI Needed

The Five Eyes urges people to increase cybersecurity by employing AI. "Adversaries are already using AI to move faster and more effectively. Defenders must do the same," says the statement. "Organizations that integrate AI tools into their security operations can detect vulnerabilities earlier, improve software quality, monitor unusual behavior, and respond faster to incidents — reducing both the cost and impact of incidents."

They also suggest that people bolster their security by limiting what's exposed to the internet, fixing system vulnerabilities much faster, getting rid of old systems, tightening access portals, and preparing for inevitable breaches.

The sophistication of AI is increasingly becoming a concern. Earlier this month, Anthropic announced that it was halting public access to its most powerful models, Fable 5 and Mythos 5. This happened after the U.S. government, citing national security concerns, ordered Anthropic to cut off access to those models to foreigners. The government claimed that it had learned of a way to get past the AI

model's safety barriers, or how to "jailbreak" the technology.

No Jailbreak?

But Anthropic isn't convinced. It said their models were robustly tested and "no testers have yet been able to find a universal jailbreak." Anthropic said that before launching Fable, it worked with the U.S. and U.K. governments as well as multiple third parties to test the model. What they found was that "Fable's safeguards are substantially more effective than those of any previously deployed model." Moreover, Anthropic claims they weren't even told about what the government specifically discovered. Anthropic:

> To date, the government has only given us verbal evidence of a potential narrow, non-universal jailbreak, which essentially consists of asking the model to read a specific codebase and fix any software flaws. Our understanding is that one potential jailbreak was shared with the government. We have reviewed a report that we believe is the basis of the government's directive and validated that the level of capability displayed there is widely available from other models (including OpenAI's GPT-5.5), and is used every day by the defenders who keep systems safe.

The prospect of AI models becoming advanced enough to pose significant hacking threats was predicted years ago. It makes sense that adversary nations could use the technology to hurt American interests. But given that the intel agencies have been at the center of many misinformation operations and other, far worse, campaigns of skullduggery, this could be as much a legitimate warning as it could be pre-programming for a coming false-flag attack.