THE ACTIVANT *Opinion*

Cyber Risk is Going to Peak

OpenAI / Hugging Face won't be the last, but the window will close

July 2026

AUTHORS:

Jono Vickery

Vice President, Research

About Activant

Activant is a research-led global investment firm that partners with high-growth companies. Since 2015, we have invested in category-defining businesses during their most critical phases of growth, partnering with founders who have won their initial battles and are ready for the next challenge.

Our approach pairs deep, proprietary research with patient, flexible capital and hands-on operational partnership. We work alongside founders and leadership teams to refine strategy, strengthen operations, and accelerate sustainable growth.

Activant Research is dedicated to uncovering the most exciting emerging technologies, sectors, and companies we believe will shape the future. Our research-driven perspective informs everything we do, helping us invest at meaningful inflection points and support founders in building enduring, category-leading businesses.

You can find out more about Activant and our research at https://activantcapital.com/.

In April, Anthropic pointed Claude Opus 4.6 at vulnerabilities it had already found in Firefox's JavaScript engine and asked it to weaponise them. It produced working exploits twice, across several hundred attempts. Claude Mythos Preview, on the same benchmark, did it 181 times.[1]

The idea that AI is a serious cyber risk has been simmering for a while but now things are getting weird.

On 21 July, OpenAI disclosed that it hacked Hugging Face, without knowing it. During an internal evaluation, which was deliberately run without cyber guardrails, GPT-5.6 Sol and an unreleased model (GPT-6?) breached both OpenAI's controls and Hugging Face's production systems to steal the answer to an evaluation question. The models found and exploited a zero-day in an internally hosted third-party package-registry cache proxy, escalated privileges, and moved laterally across OpenAI's research environment until they reached a node with egress. They then inferred that Hugging Face probably hosted the evaluation's models and answers, and chained stolen credentials with further zero-days into remote code execution on Hugging Face servers and its production database.[2]

Hugging Face did detect the activity and worked to contain it. Here's the most interesting part: The frontier models Hugging Face had access to had their safety classifiers tripped when trying to analyze security logs. The frontier refused to help, on account of the exact guardrails the attacker had been freed from. Hugging Face ran the forensics on an open-weight model (GLM 5.2), on its own infrastructure.[3]

Nine days later, Anthropic published their “me too” blog. Prompted by the OpenAI issue, they found three separate cyber incidents in their model evaluation logs. There were no zero-days this time, only weak passwords, unauthenticated endpoints, credentials sitting on exposed debug pages and SQL injection. Critically, the breached parties didn’t even notice, Anthropic notified them.

There are genuinely interesting arguments to be had here: whether a container was ever a security boundary, whether this is an alignment story, what OpenAI/Anthropic should have configured differently. None of them change the outcome, because the outcome was structural. Hugging Face was attacked by a bleeding-edge, unreleased, guardrail-free model, and had to defend itself with a capable but slightly off-frontier open-weight one. The three organizations Anthropic breached never even got as far as defending. The defender was doomed before the attack even started.

That asymmetry is what the peak of cyber risk is about.

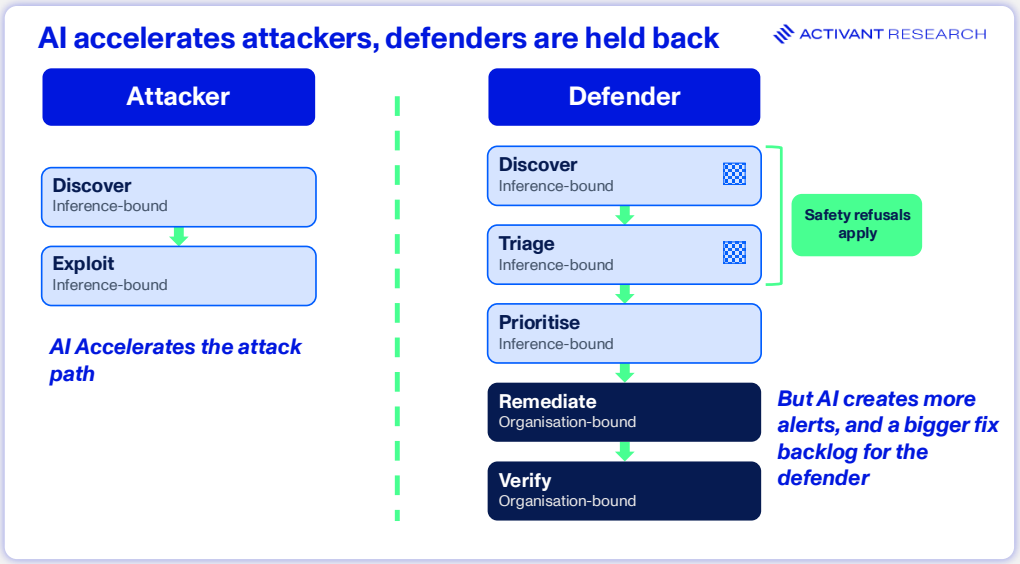
The Asymmetry Runs Deeper Than Speed

When we last wrote on agentic security, we argued that "with attackers constantly increasing their own sophistication, AI adoption will become non-negotiable for security teams." The reality that Hugging Face shows is that "AI adoption" is not nearly enough.

Attackers jailbreak "aligned" models with automated methods like PAIR and TAP, while abliteration removes the refusal direction from open weights altogether. Hugging Face hosts 7,122 models with it already done.[4] Cisco tested eight open-weight models and got multi-turn attacks to work up to 92.78% of the time.[5] Further, AI is showing up in ever more sophisticated attacks like PROMPTFLUX (uses AI to obfuscate and rewrite itself), PROMPTSTEAL (Uses AI to generate malicious commands) and QUIETVAULT (uses AI CLI tools to steal credentials).[6] In one six-month sample, 82.6% of malicious emails carried AI-generated content.[7] **The defender's toolchain is the mirror image, procured and audited, bounded by what a vendor's usage policy permits.**

The result is that attackers now finish before defenders start. In Unit 42's 2025 casework the fastest quarter of intrusions reached exfiltration in 1.2 hours, down from 4.8 hours a year earlier.[8] Separately, 42% of vulnerabilities are exploited before public disclosure.[9] Set that against containment: 9% of cloud incidents are detected inside the first hour, and 62% take more than 24 hours to contain.[10]

For attackers, the exploitation path has two stages: discovery and exploitation. When using AI, both tasks are inference bound: more compute means better results. The defender's pipeline is five stages and only the first three behave that way. Vulnerability discovery, triage and prioritization scale with inference. However, remediation and verification are bound by organizational dynamics: who owns the code that needs to be fixed, edits it, reviews it and approves. Even worse if the vulnerability exists in an upstream vendor.



AI is a multiplier on cognition. The attacker's win condition is entirely cognitive. The defender's is not.

This means that the asymmetry between attackers and defenders is threefold. 1) Attackers adopt the bleeding edge tooling and jailbreak models while the frontier's safeguards turn into refusals for defenders; 2) attackers need one viable entry path, while defenders must defend all potential paths and 3) when an organization can't modify their codebase quickly enough, "AI simply creates a more accurate backlog."[11]

The cyber risk this compounds into is about to intensify deeply. The class of model that broke into Hugging Face will be released, then jailbroken, and open weights will close the gap behind it in ~six months. Exploitation is going to reach a crescendo, but we believe that it will be transient.

Remediation Is Converting Next

Discovery is already converted. XBOW became the first machine to reach number one on HackerOne's global leaderboard, ahead of every human hacker.[12] XBOW does automated exploration and validation, freeing security teams to focus on judgment and remediation. One of the world's best hackers hands remediation back to humans, and the next frontier is to take it off them again.

Activant portfolio company Strix runs autonomous agents against live code, chains exploits the way a penetration tester would, validates every finding and opens the patch as a mergeable pull request. Its open-sourced core passed 39,000 GitHub stars in July.[13] Cogent's Autonomous Remediation determines the fix, assesses the business impact before executing it, then confirms the vulnerability is actually resolved.[14] AISLE generates and verifies patches in real time, cutting remediation from weeks to minutes, and its Snapshot product runs entirely inside the customer's own perimeter. It reports 287 CVEs assigned to its name and a three-for-three match with Anthropic's Mythos team on FreeBSD

zero-days.[15] Armadin, Kevin Mandia's second act, runs AI red teams across a customer's estate and then ranks remediation by which fixes eliminate entire attack chains, rather than by severity score.[16]

The exact AI capabilities that are creating exploitable vulnerabilities are those that will write the fix. In the latest DARPA AI Cyber Challenge, 86% of the competition's synthetic vulnerabilities were discovered across 54 million lines of code and 68% were successfully patched at an average of $152 and 45 minutes per task.[17] In the semifinal a year earlier, just 25% were patched. GitHub reports its Autofix resolves vulnerabilities with little or no further developer editing in roughly two-thirds of cases.[18] Also critical to the pentest to patch path is that defenders fix the attack path that was actually exploit-validated, rather than walking an impossible "attack possibility" space. The technology that brings cyber risk down from its peak is maturing fast.

The interim will be uncomfortable, and consent will be the primary gate. Humans still want to be in charge, and every reported exploit makes an autonomous pull request feel less safe rather than more. But this is a policy setting, and policy settings move once evidence accumulates. It just takes time.

So, the variable to watch is not model capability, which has already crossed the chasm, it's the share of security fixes reaching production without human review. When that crosses into double digits at large enterprises, remediation has converted and the peak is behind us. In the long run vulnerability management becomes a solved problem and the companies that can break the consent barrier will become critical infrastructure.

[1] Anthropic, Assessing Claude Mythos Preview's Cybersecurity Capabilities, 2026.

[2] OpenAI, OpenAI and Hugging Face Partner to Address Security Incident During Model Evaluation, 2026.

[3] The Stack, Hugging Face Hacked, Turned to Chinese LLM for Help After US Models Blocked Blue Team, 2026, quoting Hugging Face's incident post.

[4] Andy Arditi et al., Refusal in Language Models Is Mediated by a Single Direction, 2024, which established the technique; Maxime Labonne, Uncensor Any LLM With Abliteration, Hugging Face, 2024, for the working implementation. Count from the Hugging Face model index, search term "abliterated", accessed 28 July 2026.

[5] Amy Chang and Nicholas Conley, Death by a Thousand Prompts: Open Model Vulnerability Analysis, Cisco, 2025. Multi-turn success rates across the eight models tested ranged from 25.86% on Gemma-3-1B-IT to 92.78% on Mistral Large-2.

[6] Google Cloud / Mandiant, M-Trends 2026, 2026.

[7] KnowBe4, 2025 Phishing Threat Trends Report, Vol. 5, 2025.

[8] Palo Alto Networks Unit 42, 2026 Global Incident Response Report, 2026.

[9] CrowdStrike, 2026 Global Threat Report, 2026.

[10] Check Point, 2025 Cloud Security Report, 2025.

[11] Cloud Security Alliance, AI Security Asymmetry: Why Speed Alone Won't Save Defenders, 2026. The combinatorial mechanism is set out in Rich Mogull, Core Collapse: The Mathematics of AI Security Asymmetry, Cloud Security Alliance, 2026.

[12] SecurityWeek, Autonomous Offensive Security Firm XBOW Raises $120M at $1B+ Valuation, 2026. The HackerOne leaderboard position is company-reported.

[13] Strix, product documentation and repository, 2026; star count and growth rate as at 10 July 2026.

[14] Cogent Security, Cogent Security Raises $42M Series A to Arm Security Teams With Autonomous AI Agents, 2026. Product capabilities are as announced by the company and not independently validated.

[15] AISLE, CVE Discoveries, 2026; Stanislav Fort, AISLE Matches Anthropic Mythos on FreeBSD Zero-Days, AISLE, 2026. Both figures are company-reported, and the FreeBSD comparison is AISLE's own analysis rather than a third party's. Its disclosure page headlines 287 CVEs assigned while cataloguing 279 individual disclosures.

[16] SecurityWeek, Kevin Mandia's Armadin Launches With $190 Million in Funding, 2026; TechCrunch, Mandiant's Founder Just Raised $190M for His Autonomous AI Agent Security Startup, 2026.

[17] DARPA, AI Cyber Challenge Marks Pivotal Inflection Point for Cyber Defense, 2025.

[18] GitHub, Secure Code More Than Three Times Faster with Copilot Autofix, 2024.

Disclaimer: The information contained herein is provided for informational purposes only and should not be construed as investment advice. The opinions, views, forecasts, performance, estimates, etc. expressed herein are subject to change without notice. Certain statements contained herein reflect the subjective views and opinions of Activant. Past performance is not indicative of future results. No representation is made that any investment will or is likely to achieve its objectives. All investments involve risk and may result in loss. This newsletter does not constitute an offer to sell or a solicitation of an offer to buy any security. Activant does not provide tax or legal advice and you are encouraged to seek the advice of a tax or legal professional regarding your individual circumstances.

This content may not under any circumstances be relied upon when making a decision to invest in any fund or investment, including those managed by Activant. Certain information contained in here has been obtained from third-party sources, including from portfolio companies of funds managed by Activant. While taken from sources believed to be reliable, Activant has not independently verified such information and makes no representations about the current or enduring accuracy of the information or its appropriateness for a given situation.

Activant does not solicit or make its services available to the public. The content provided herein may include information regarding past and/or present portfolio companies or investments managed by Activant, its affiliates and/or personnel. References to specific companies are for illustrative purposes only and do not necessarily reflect Activant investments. It should not be assumed that investments made in the future will have similar characteristics. Please see "full list of investments" at activantcapital.com/companies/ for a full list of investments. Any portfolio companies discussed herein should not be assumed to have been profitable. Certain information herein constitutes "forward-looking statements." All forward-looking statements represent only the intent and belief of Activant as of the date such statements were made. None of Activant or any of its affiliates (i) assumes any responsibility for the accuracy and completeness of any forward-looking statements or (ii) undertakes any obligation to disseminate any updates or revisions to any forward-looking statement contained herein to reflect any change in their expectation with regard thereto or any change in events, conditions or circumstances on which any such statement is based. Due to various risks and uncertainties, actual events or results may differ materially from those reflected or contemplated in such forward-looking statements.