Futurum
RESEARCH

Why AI Learned to Attack Before It Learned to Defend

A Cheap Test for Offensive Actions and an Absent Defensive One, With Consequences for Vendors, Investors, and Practitioners Alike

Analyst(s): Fernando Montenegro
Publication Date: August 24, 2026
Document #: AIOFM202608

Key Points

- The Testing Asymmetry Behind Offensive AI: A landed exploit is decidable proof, while confirming a defense holds has no equally cheap test, so AI's first security wins are offensive by structure, not by accident.
- Acceleration Only on the Provable Side: AI can more easily substitute for human effort on offense because those outcomes are easier to test, and Anthropic's own Glasswing update concedes that the bottleneck has shifted to verifying, disclosing, and patching.
- Manufactured Proof as the Durable Moat: The enduring product demonstrates both exploitability and validated remediation, rather than merely asserting coverage.

Recommendations

1. Offensive Vendors – Sell Proof, Not Volume: Lead with reproducible exploitability and reachability, and with time-to-validated-remediation, now that finding is cheap. Shipping raw findings only enlarges the customer's backlog.
2. Defensive Vendors – Drop Unverifiable Coverage Claims: Replace "our AI catches everything" with bounded, checkable claims. Buyers facing the machine-speed gap discount assertion, so unfalsifiable confidence is a liability.
3. Defensive Vendors – Build the Missing Oracle: Invest in control validation, true-positive confirmation, and reachability over more detection. It is the un-commoditized prize because it is unsolved; ceding it means competing where AI already helps every rival.
4. Both Camps – Make Proof of the Platform: Own the verification layer across the exposure lifecycle, not one step of it. As offense industrializes, tools that only find get absorbed by platforms that validate.

What You Need to Know

AI Learns to Find and Exploit

In April 2026, Anthropic launched Project Glasswing and its Claude Mythos model to find vulnerabilities in widely used software. Within a month, partners reported more than 10,000 high- or critical-severity flaws, and on an independently assessed open-source subset, 90.6% of Mythos findings were valid true positives.

Anthropic noted the constraint had shifted from finding flaws to fixing them. A typical high- or critical-severity bug took about two weeks to patch, and only 75 of 530 reported issues had been patched by then.

On August 10, 2026, OpenAI expanded its parallel effort, Daybreak, into two tiers: Daybreak Blue for defensive work such as code review and patch validation, and Daybreak Red for authorized vulnerability research and exploit validation.

Its GPT-5.6-Cyber model completed advanced offensive-security tasks at roughly 95% success, compared to 1.5% to 2% for standard models, and reported finding more than 400 kernel privilege-escalation flaws and a V8 engine flaw (CVE-2026-15903). The same class of capability now exists within commercial appsec and penetration testing tools.

Why Defense Has No Cheap Oracle

Modern AI systems get better at a task when training and evaluation can automatically check success. Reinforcement learning and benchmark loops need a reward signal, a cheap and repeatable test (referred to as an "oracle") that says whether an attempt worked, run thousands of times without a human in the loop.

Offensive security supplies that signal naturally. Did the exploit execute? Did the payload land? Was the capture-the-flag target reached? Each is decidable and scriptable.

This is the same property that made models strong at math and coding, where a unit test or a proof check can cheaply verify correctness.

Defensive security offers no equivalent. Confirming that a control blocks an attack in practice, that an alert is a true positive, or that a system is now safe generally demands knowing much more about the environment, potentially covering every path an adversary might conceivably take, whether human judgment will apply correctly, and waiting to see what happens.

A defense that is not breached today has not been proven safe; it has only not yet failed. Without a cheap success signal, the loops that AI training uses to learn offensive skills are much harder to replicate for teaching defense.

©2026 The Futurum Group, LLC. All rights reserved.

The Capability Is Diffusing Downward

The offensive edge is not staying locked in the frontier labs. In July 2026, the UK AI Security Institute and its US counterpart found that leading open-weight models trail the frontier on cyber tasks by roughly four to seven months, down from six to ten a year earlier.

The gap looks even smaller on individual benchmarks. DeepSeek's own published results for its V4-Pro model, released in August, put it near parity with a frontier model from about two months earlier on one offensive-cyber test, though such vendor-reported figures await independent validation.

GLM-5.2, the open model Hugging Face used to investigate its own breach, sits in the same tier. Unlike the access-gated frontier offensive tools, these open models are available to anyone.

A Breach, and the Industry's Response

On July 16, 2026, Hugging Face disclosed that an autonomous AI agent had breached its internal infrastructure days earlier. The agent gained code execution through two paths in the dataset-processing pipeline, escalated to node-level access, and harvested cloud and cluster credentials.

It moved laterally across internal clusters over a weekend through thousands of short-lived sandboxed actions. Hugging Face found no tampering with public models, datasets, or Spaces.

It stated the model behind the attack was not known, and used GLM-5.2 to analyze more than 17,000 attack events. On July 21, OpenAI attributed the activity to one of its pre-release models.

Days later, a group of large vendors formed the Open Secure AI Alliance. NVIDIA, Microsoft, CrowdStrike, IBM, and Palantir were among 37 members, and NVIDIA open-sourced a framework, NOOA. The alliance is a coordinated, standards-oriented answer to AI-enabled attacks rather than a single product.

Buyers Report a Machine-Speed Gap

Futurum's 1H 2026 Cybersecurity Decision-Makers survey found that 55.9% of respondents have AI usage policies but lack the technical ability to detect or contain a compromised AI agent operating at machine speed (N=933, unweighted).

A comparable 56.6% reported being highly concerned about the security implications of agentic AI, and 53.9% were concerned that software vendors are integrating autonomous agents without sufficient visibility or controls.

©2026 The Futurum Group, LLC. All rights reserved.

Analysis

The easy read is that offensive AI is simply the better technology today. We would offer a less comfortable one: offense wins first, not because the models attack better than they defend, but because offense offers a cheap way to check its own work, and defense does not.

That single fact, more than any benchmark, explains the pattern in the news. It sets up the questions we take up here: where AI substitutes for people, who captures the value, how defenders respond, and what all of it demands of vendors and buyers.

The Asymmetry Is Structural, Not a Head Start

The conventional read is that offensive AI is simply ahead, and defense will catch up as models improve. We argue that this is the wrong mental model.

The gap is not a maturity lag; it is a property of what can be cheaply verified, and no amount of model progress hands defense a success signal it structurally lacks.

This is why we expect more capable models to widen the gap about as often as they narrow it. Each capability jump is easy to point at offense, where a landed exploit grades the work, and hard to point at defense, where nothing does as cheaply. Time helps the side that can already measure itself.

The exception would be a different kind of model, not a bigger one. Neurosymbolic methods, or learned world models, which reason about and verify system behavior rather than pattern-match it, are a plausible route to more efficient defensive oracles; scaling alone is not.

AI Substitutes on Offense and Only Assists on Defense

We build here on a position we have long held on the SOC: AI amplifies rather than replaces. The testing asymmetry says where each happens.

Where a sound test grades output, AI can more easily replace human effort outright; where none exists, AI can only help a human work faster. That is why offense industrializes while defense, at best, accelerates.

The Hugging Face breach is the asymmetry in a single frame: the exploit was decidable, and it landed, while the attribution, which model did it, stayed unresolved even to the defender.

We should be precise, though. Offense's oracle is one-directional. A landed exploit proves exploitable; a failed one never proves safe. The advantage is real, but it is not omniscience.

©2026 The Futurum Group, LLC. All rights reserved.

Winners Provide Proof; Losers Assert Coverage

We believe durable value now accrues to vendors that manufacture proof rather than assert it. Autonomous pentest and exposure-validation players sell a checkable claim: this exploit lands, this path is reachable, this fix holds.

Horizon3.ai's NodeZero, XBOW in application security, and validation vendors such as Pentera, Cymulate, and Picus are representative of the category, not an endorsement or a ranking. Their advantage is that a buyer can check the claim, which is the un-commoditized layer precisely because the defensive oracle is unsolved.

On the other side, undifferentiated "our AI catches everything" detection pitches are the most exposed. In a market where buyers cannot verify quality at the point of purchase, only a signal costly to fake separates genuine capability from confident marketing. That is a lemons market, and unverifiable assurance is what it discounts.

Defenders Will Rebuild Around Verifiable Controls

This is not a counsel of despair. If detection at machine speed lacks a cheap test/oracle, the rational response is to lean on the parts of defense that do have one. We expect defenders to shift investment from detecting the intrusion toward shrinking what must be detected, using controls whose success is itself checkable:

- Faster, more automated patching and remediation, where a fix is verifiably applied and a specific hole verifiably closed, attacks the backlog that AI-found volume creates. Remediation-orchestration and exposure players such as Brinqa, Nucleus Security, and others sit here, as do patch-automation tools such as Automox and Tanium.
- Segmentation, which turns a blocked attack path into a reachability fact that a team can test. It runs from dedicated microsegmentation, from Illumio and Elisity, to the segmentation now inherent in SASE platforms such as Zscaler, Netskope, Palo Alto Networks, and Cisco, among many others.
- Phishing-resistant authentication, which removes an attack class rather than relying on detection. Microsoft's move to make passkeys the default in Entra ID from September 2026, retiring SMS and voice, is a clear marker, alongside the broader FIDO2 and passkey push from providers such as Okta and Yubico.

Consolidation is a parallel response. Palo Alto Networks with Cortex, CrowdStrike with Falcon, Microsoft, Cisco, Fortinet, and SentinelOne, among others, are each betting on the platform, and Sophos joined them in July with Fusion, an AI-native system that folds endpoint, SIEM, identity, and MDR into a single platform.

©2026 The Futurum Group, LLC. All rights reserved.

The bet is that defenders who collapse tool sprawl can at least react closer to machine speed. None of these manufacture the cheap detection oracles, but together they change the game defense plays, from spotting the fast attacker to denying cheap paths.

The Diffusion Clock Sets the Timeline

We expect the ungated spread of offensive capability to open models to compress defender timelines faster than hiring or budgets can absorb. Access-gated frontier tools slow proliferation; open models at near-parity, available to anyone, do not.

The attacker's cost curve is declining on a schedule that defenders cannot control.

Our specific expectation for the next 12 to 18 months is that exposure programs that prioritize proof of exploitability will visibly outperform those that prioritize severity scores, and that validated remediation evidence will begin appearing as a procurement requirement.

The Open Secure AI Alliance is a reasonable move, but coordination does not manufacture an oracle; it organizes the search for one.

What Vendors Must Sell, and Buyers Must Demand

The behavior change follows directly. Offensive vendors should lead with proof and remediation validation, not finding counts.

Defensive vendors should retire unfalsifiable coverage claims and invest in the missing verification layer: control validation, true-positive confirmation, and reachability. And practitioners should re-anchor programs to the confirmed-finding rate and remediation velocity, not to alert volume or mean-time metrics that measure activity rather than outcome.

The market that prices proof above the assertion is coming, pulled forward by an offense that can already grade itself. The open question is who has built the oracle when it arrives.

What to Watch

- Will anyone manufacture cheap defensive oracles? Watch neurosymbolic AI and world-model research for a way to automatically grade whether a control stopped an attack. Whoever cracks that resets the board.
- Will the platforms absorb the validation layer? If Palo Alto Networks, CrowdStrike, or Microsoft buy their way into proof-of-exploitability, the prize commoditizes. Watch the M&A flow around pentest and BAS vendors.
- Will buyers demand proof in procurement? Watch for RFPs that ask vendors to demonstrate exploitability and validated remediation, not assert coverage.

©2026 The Futurum Group, LLC. All rights reserved.

- Will open models erase the offensive gap? If the next government evaluations land within weeks of the frontier, access controls on tools such as Daybreak Red lose their point. Watch the AISI and CAISI cadence.

Other Insights from Futurum

So This Is How AIs Attack: Observations From the OpenAI & Hugging Face Incident

The Hard(er) Challenge in Agent Governance Is Authorization

Securing Agentic AI Is the Multi-Level Challenge for Security Teams

©2026 The Futurum Group, LLC. All rights reserved.

About Us

About the Authors

Fernando Montenegro is the Vice President and Practice Lead for Cybersecurity at The Futurum Group. In this role, he leads the development and execution of the Cybersecurity research agenda, working closely with the team to drive research for clients. His background is centered on enterprise cybersecurity and includes diverse roles in industry research, customer support, security, IT operations, professional services, and sales engineering. Fernando can be reached at fmontenegro@futurumgroup.com.

About The Futurum Group

Every day, The Futurum Group's analysts, researchers, and advisors help business leaders worldwide anticipate tectonic shifts in their industries and leverage disruptive innovation. Unlike traditional analysts, The Futurum Group works not only in analysis and research but also takes that insight and knowledge even further, engaging all the way through the go-to-market process.

Futurum Research provides in-depth research and insights on global technology markets using advisory services, custom research reports, strategic consulting engagements, digital events, go-to-market planning, and message testing. It also creates, distributes, and amplifies rich media content that all stakeholders read, watch, and listen to.

See more details on The Futurum Group at futurumgroup.com.

Copyright & Use License

Copyright Notice

Copyright ©2026 by The Futurum Group, LLC. All rights, including that of translation into other languages, are specifically reserved. No part of this publication may be reproduced in any form, stored in a retrieval system, or transmitted by any method or means, electrical, mechanical, photographic, or otherwise, without the express written permission of The Futurum Group futurumgroup.com. United States copyright laws and international treaties protect this publication. Unauthorized distribution or reproduction of this publication, or any portion of it, may result in severe civil and criminal penalties and will be prosecuted to the maximum extent necessary to protect the publisher's rights.

License Notice

This document may be distributed within the licensed organization only.

The following acts are prohibited:

Transmittal to others outside your immediate organization including partners, resellers, external consultants, etc. in any media format

Posting on a website which is accessible to others outside your immediate organization

The possession or use within an unlicensed organization

Limitation of Liability Notice

The information contained herein has been obtained from sources believed to be reliable. The Futurum Group shall have no liability for errors, omissions, or inadequacies in the information contained herein or for interpretations thereof. The reader assumes sole responsibility for the selection of these materials to achieve their intended results. The opinions expressed herein are subject to change without notice.