Cobalt

WHITE PAPER

Separating Signal from the Noise

A CISO Guide to AI in Pentesting

Executive Summary

The cybersecurity market is flooded with AI pentesting solutions, making it difficult for CISOs to distinguish genuine advancements (signal) from marketing hype (noise). This challenge is compounded by a multi-front war: 81% of organizations are concerned about the threat of AI-driven attacks [1], while simultaneously grappling with the significant risks posed by their own AI supply chain.

Fully autonomous AI pentesting, in its current state, is not able to deliver on its promises. These tools function at this point as glorified scanners. As this paper will show, they excel in some areas, but there is a ceiling on what they can find.

Furthermore, AI models themselves introduce a new, critical attack surface. Data from the Cobalt State of LLM Security Report reveals that LLM applications have the highest proportion of serious vulnerabilities (32%) of any asset type, yet suffer from the lowest remediation rate (21%). Relying on AI, to fix AI—given it is still nascent technology with serious security downsides—is less than ideal.

The future of pentesting, however, is not about replacing human pentesters with AI, but creating a productive partnership. Finding true value—the signal—lies in using AI as a co-pilot to augment the irreplaceable creativity and strategic thinking of human pentesters.

This guide provides CISOs a framework to cut through the marketing noise, evaluate AI solutions, and build a strategy that addresses both the claims of AI-powered tools and the security of the AI models themselves.

[1] SANS 2025 AI Survey, SANS Institute, pg.2

Contents

AI Epidemic: New Bugs, Hyped Cures

The modern security landscape is facing a surge of new risks due to the explosive growth of AI. First there are the innovative attackers leveraging AI to craft and execute more sophisticated attacks. Then there are bugs generated by AI coding agents that find their way into software. There may also be vulnerabilities hidden within third-party models that are a "black box" for customers, making AI models a supply-chain threat.

In other words, AI is not a single threat but a multi-pronged threat for the enterprise, creating a complex matrix of issues for security leaders to manage.

The Three-Pronged AI Threat

CISOs are no longer defending against simple, linear attacks. They are defending against a dynamic new class of risk that manifests in three ways:

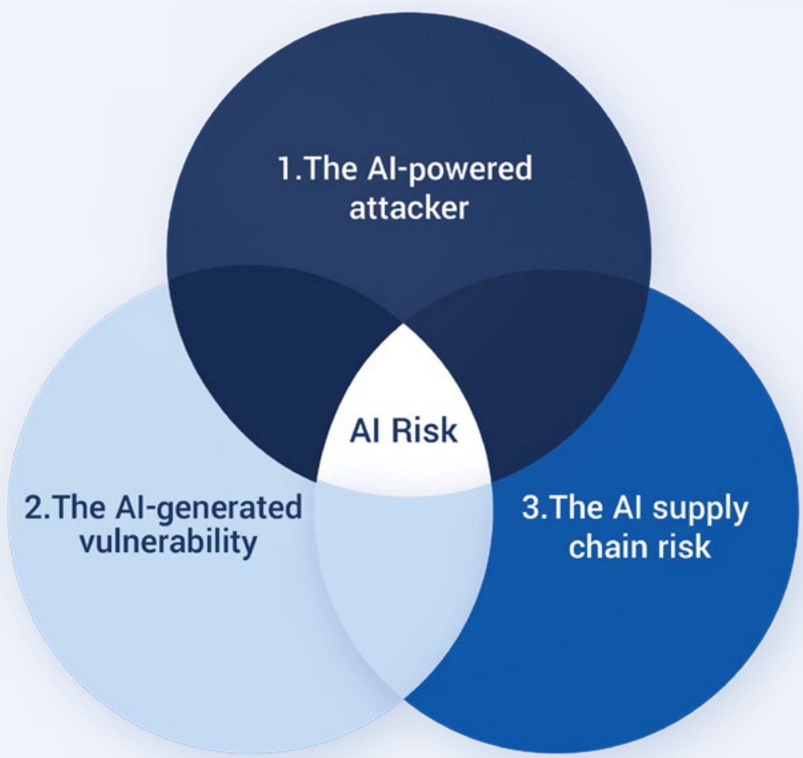


Figure 1. Venn diagram of AI risks. Source: Cobalt.

The AI-powered attacker	This is the most obvious threat. 81% of organizations are concerned about AI-driven attacks. Adversaries are using AI to create more sophisticated phishing, generate polymorphic malware, and discover vulnerabilities at a speed and scale that is overwhelming traditional defenses.
The AI-generated vulnerability	The push for innovation has led to the rapid adoption of AI-driven development tools. These tools risk becoming flaw multipliers, spreading subtle vulnerabilities and buggy code throughout a codebase, at a rate that human developers and legacy scanners cannot possibly match.
The AI supply chain risk	The push for innovation has led to the rapid adoption of AI-driven development tools. These tools risk becoming flaw multipliers, spreading subtle vulnerabilities and buggy code throughout a codebase at a rate that human developers and legacy scanners cannot possibly match.

The Market's AI-Powered Prescription

In response to this AI-driven epidemic, the market has offered what looks like a miracle cure: AI-powered, fully autonomous pentesting. The promise is seductive: fight AI with AI.

This has left security leaders in an impossible position, struggling to distinguish a genuine solution from flashy AI-automated tools and the latest hot marketing campaign. Does the AI pentesting industry fulfill its promise, or is it just the loudest voice in the room right now?

AI Strengths and Limitations: Getting Beyond Marketing Hype

It's time to cut through the hype to examine the reality—it's a noisy field, with lots of solutions to choose from, and many that introduce more noise in the form of false positives and a flood of low-level findings.

- **The noise:** The noise is the hype around fully autonomous, AI-only pentesting tools, or the stream of bug submissions from bounty hunters that rely on glorified scanners. These products promise a quick fix but lack context for your business and offer no creativity to find novel exploits. They find simple, known symptoms (like a weak encryption cipher) while completely missing the novel business logic flaw or the creatively chained exploit.
- **The signal:** The signal is not a magic pill. It's a modern, evidence-based treatment plan: a human-led, AI-powered approach. In this model, AI is the advanced diagnostic tool—the co-pilot. Human pentesters use the AI's speed and scale to handle the reconnaissance and repetitive tasks at a level humans can't. This empowers the human pentester to use their creativity and context to cut through the distracting noise of false positives and low-risk findings to diagnose and cure the real, complex disease.

It's not that AI pentesting has nothing to offer—far from it. But CISOs must first understand the practical strengths and, just as importantly, the fundamental limitations of AI.

AI Strengths	AI Limitations
Speed, scale, and consistency: Automates triage, enrichment, and repetitive tasks.	Lack of context: Cannot understand business logic or the intended function of an application.
Pattern recognition at scale: Detects anomalies and known flaws across vast datasets.	No creativity: Fails to chain multiple, low-impact vulnerabilities into a critical exploit.
Contextual triage: Assists humans by mapping findings to CWE/OWASP and prioritizing alerts.	Validation gaps and hallucination risk: Prone to false positives. Requires verifiable proof-of-concept artifacts.
Continuous learning: Improves accuracy when trained on validated, real-world pentest data.	The black box: Can suffer from model drift, bias, and integration gaps, creating noise.

The Noise: Hype of the Autonomous Pentester

The promise of a fully autonomous AI pentester is compelling. But in its current state, it's inadequate for protecting organizations against supply chain and other systemic risks. The autonomous pentesting solutions available today may even create a dangerous false sense of security, by generating a high volume of alerts and false positives that crowd out the critical signals that truly matter.follow a responsible disclosure process, informing the organization promptly in a way that allows them to fix problems before any details are made public.

> "CISOs need to fully understand which vulnerabilities will become tomorrow's headlines. AI can rapidly scan for vulnerabilities and accelerate discoveries. On the other hand, human judgement accelerates decision making. The future of pentesting belongs to the teams or organizations that can do both."
>
> **ANDREW OBADIARU,** CISO, COBALT

Alert Fatigue and the Flood of Bug Bounty Slop

Security professionals are already reporting in surveys that the noise from AI-only tools is creating an overwhelming flood of repetitive and low-impact bugs. It is the single biggest complaint from security teams and researchers who are forced to deal with the output of AI systems that tend to dominate findings from bug bounty platforms.

1. Triage trouble: The SANS Institute's 2025 [2] AI Survey provides a clear internal metric. When asked about the shortcomings of AI systems, the number one complaint, cited by 66% of security teams, is that AI "generates a high number of false positives (alert fatigue)." This noise wastes valuable analyst time, erodes trust in the tools, and buries real threats.
2. Opened floodgates: This same problem is now publicly exploding in the bug bounty community. As reported by CSO Online [3], AI-powered bug-hunting tools are flooding bug bounty platforms with a tidal wave of junk submissions. Triage teams are becoming overwhelmed by a massive increase in low-quality, automated reports. This creates a new kind of alert fatigue for the human triage teams, burying the signal (novel, high-impact human findings) under a mountain of noise (automated AI findings).

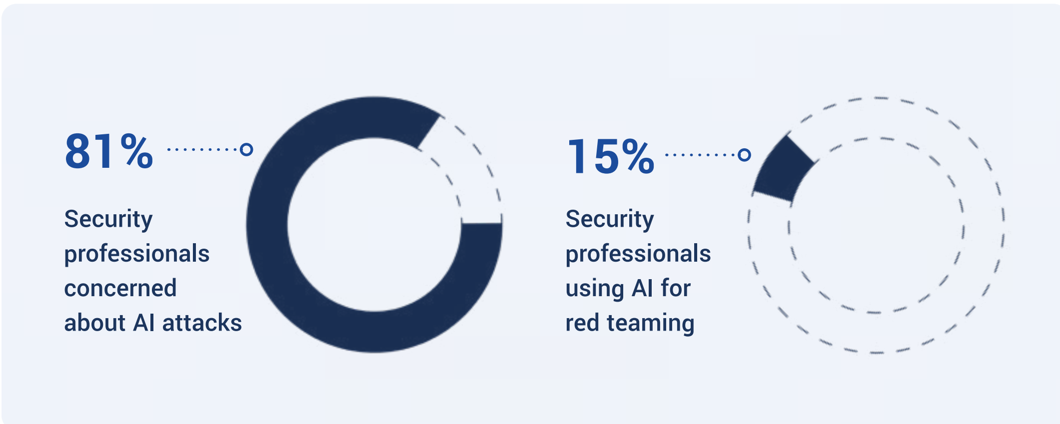
AI, by itself, amplifies noise. Without a human directing the tool to provide creativity and context, AI tools don't solve the needle in a haystack problem—finding exploitable vulnerabilities that genuinely put your organization at risk—they just make the haystack infinitely larger.

[2] SANS 2025 AI Survey, SANS Institute, pg. 2.

[3] "AI Bug Hunting Shakes Up Bounty Industry—For Better or Worse," John Leyden, CSO, Oct. 31, 2025.

Overhyped Offensive Security AI

The SANS survey data quantifies this hype. While 81% of organizations are concerned about AI-driven attacks, only 15% are actually using AI for their own red teaming.



This massive 66 percentage point gap shows that CISOs understand the threat, but have no confidence in the solutions being sold to them as a quick fix. CISOs are not voting with their feet when it comes to these solutions, likely because it's not yet mature enough for the creative, offensive work that human pentesters perform.

The Automation Ceiling: What the Data Says

Some AI-assisted testing shows promise (e.g., Google Project Zero's hybrid fuzzing uncovered CVE-2020-6418, but it required human validation) [4]. But research shows a hard, quantifiable limit to what AI can do alone.

The 2025 "Multi-Agent Penetration Testing AI for the Web" paper by Isaac David and Arthur Gervais of University College London, provides a clear benchmark. According to their study, AI's accuracy plummets as business context becomes more important than simple pattern matching (see Figure 2.).

[4] "From Naptime to Big Sleep: Using Large Language Models to Catch Vulnerabilities In Real-World Code," Google Project Zero, Nov. 1, 2024.

AI Automation Limits for Common Web Vulnerabilities

CWE	Description	Est. Percentage AI Accurately Automates	Automation Level	Why AI Fails: Noise and Limitation
CWE-326	Weak Encryption	≈ 100%	High	Deterministic checks; fully automatable (e.g., SSL Labs)
CWE-918	SSRF	≈ 100%	High	Detect and validate via OOB callbacks (e.g., Burp Collaborator).
CWE-1021	Clickjacking	≈ 95%	High	Header/CSP logic is trivial to scan and validate.
CWE-89	SQL Injection	≈ 95%	Medium	Mature tooling (e.g., sqlmap), but payload validation is not perfect.
CWE-284	Improper Access Control	≈ 83%	Medium	Repeatable token-swap tests work, but some context checks still required.
CWE-434	Unrestricted File Upload	≈ 75%	Medium	Upload endpoints are fuzzable, but exploit chains and MIME bypasses need human analysis.
CW-352	CSRF	≈ 65%	Medium-Low	Missing-token detection is trivial; confirming real exploitability requires business logic understanding.
CWE-200	Information Exposure	≈ 60%	Medium-Low	AI can flag leaks but can't reliably judge sensitivity or business impact.
CWE-79	Cross-Site Scripting	≈ 57%	Low	Good for reflected/stored XSS. DOM-based and contextual cases need manual review.
CWE-20	Improper Input Validation	≈ 55%	Low	Pattern detection is possible, but assessing real-world impact and exploitability is difficult.

Figure 2. Source: Adapted from "Multi-Agent Penetration Testing AI for Web," Isaac David, Arthur Gervais, 2025. https://arxiv.org/abs/2508.20816v1.

The Signal AI Can't Find

The table above shows that AI fails at exactly two things humans are good at: context and creativity. These are the domain of the human pentester—and the source of true business risk. Below are two examples of vulnerabilities an AI fails to detect that a human pentester can find.

The Business Logic Flaw

Flaw: The e-commerce workflow bypass (Based on OWASP WSTG).

Scenario: A human tester bypasses a payment step by capturing the final confirmation URL and navigating to it directly, tricking the server into processing an unpaid order.

Why AI fails: AI has no context for the intended business workflow, and can't creatively violate a multi-step sequence. It sees a valid URL, not a broken business process.

The Chained Exploit

Flaw: Improper access control chained with stored XSS (based on CVE-2021-27852).

Scenario: A human tester uses a low-level account to exploit an access control flaw, allowing them to inject a stored XSS payload in a field only viewable by an admin. When the admin views the page, the XSS script steals their session token, leading to full account takeover.

Why AI Fails: AI lacks human creativity. It would report an access control flaw and a stored XSS flaw as two separate, low-risk dots that it can't connect. It lacks the human intelligence to chain them into a critical-severity attack.

The Signal: AI as the Pentester's Co-Pilot

The ultimate value of AI in pentesting is not replacement of the human pentester, but amplification of their skills. Not to build an autonomous doctor to cure the patient, but to build an advanced MRI machine that empowers a human doctor to find the disease faster and more accurately. The future is a human-led, AI-powered partnership—this is how we separate AI's signal from AI-generated noise.

Where AI Excels

When used as a co-pilot, AI is a powerful force multiplier that handles the repetitive, large-scale tasks that bog down human experts, freeing them to focus on high-impact, creative testing.

- **Reconnaissance and attack surface management:** AI excels at using its speed and scale to perform the time-consuming initial phase of reconnaissance, mapping vast attack surfaces, and identifying initial targets.
- **Data enrichment and triage:** AI leverages pattern recognition to sift through terabytes of data, enrich findings, and handle the repetitive task of triaging low-level alerts, freeing the human to investigate the most promising leads.
- **Accelerating reporting:** AI can assist with the repetitive task of drafting client-ready reports and remediation advice, allowing pentesters to spend more time testing and less time writing.

The Power of Real-World Data for Continuous Learning

An AI model is only as good as the data it's trained on. If you train an AI on noise, it will only learn to generate more noise.

This is the critical differentiator. Most AI models are trained on synthetic lab data or noisy, unvetted bug bounty submissions, which may contribute to the glut of false positives that 66% of respondents in the SANS survey said were causing alert fatigue.

AI models that are trained on real-world, human-validated pentest findings are vastly more trustworthy. Cobalt uses a curated dataset from thousands of pentests conducted over a decade, providing an unparalleled understanding of the vulnerabilities that truly matter. This data can then be used to train the AI to find the real signal from the start.

This training data is not just for finding old flaws. By training models on data from new, real-world LLM pentests, Cobalt can also build our co-pilot to help human experts identify novel AI-specific flaws, like prompt injection and model poisoning.

CISO's Framework for Evaluating AI Pentesting Solutions

As a CISO, you can cut through the hype with a cheat sheet of hard questions—both for your vendors and your internal teams.

Part 1: Questions to Evaluate a Vendor's AI Tool

1. The human element

- ☑ Describe the exact role of the human pentester in your process. How does your AI augment, not replace, their expertise?
- ☑ How do you test for the complex business logic flaws and chained exploits that your AI will always miss?

2. The data engine

- ☑ What data is your AI model trained on to prevent false positives? Is it real-world pentest data or synthetic data?

3. The process and output

- ☑ Is your final report a raw data dump from the AI, or is it a curated, human-vetted report with actionable business insights?
- ☑ How does your platform integrate validated findings into our developer workflows?

Part 2: How to Evaluate Your Own AI Deployments

The CISO's adversary is not just the AI attacker—it's the vulnerabilities in the AI you are deploying. Use these questions to evaluate your internal AI supply chain risk.

1. Pentesting your own AI applications

- ☑ Are we testing for the OWASP Top 10 for LLMs, or just for traditional web flaws?

2. Managing the black box risk

- ☑ What is our plan when a vulnerability is found in a third-party model we don't control?

3. Testing for AI-specific flaws

- ☑ Are the top LLM-specific vulnerabilities—like sensitive information disclosure, model poisoning, and training data leakage—on our testing checklist?

Preparing Your Security Program for the AI-Driven Future

This is the ultimate truth AI pentesting must resolve: the true adversary is a creative, context-aware human. The most effective defense, therefore, must also be a creative, context-aware human—empowered by the best tools.

> “As threats become more complex and attackers leverage AI, we must evolve our defenses in kind. Human pentesters are the most effective when it comes to uncovering real-world risks. By arming them with tools that match the speed and sophistication of modern threats, we’re not replacing pentesters with AI; we’re opening doors to a whole new level of creative liberty, accuracy, and focus.”
>
> GUNTER OLLMANN, CTO, COBALT

A CISO’s Call to Action

Start preparing your offensive security program today for the next wave of AI-driven tools and methodologies. This will help you identify risks sooner, resolve issues faster, and keep innovation moving.

Remember that the majority of vulnerabilities don’t get fixed. Many linger for months or years—according to the Cobalt State of Pentesting Report 2025, just 48% of vulnerabilities are ultimately fixed, and it takes over three months to resolve just half of the highest-risk vulnerabilities. Many open vulnerabilities won’t be impactful or high-risk, and decisions about what to fix require SLAs and careful risk modeling. Yet some proportion of unresolved vulnerabilities pose a risk to the organization. Triaging findings is enormously important here—finding the signal in the noise.

The goal of a modern, AI-powered program is to cut through the noise of false positives and repeat, low-level findings from AI-assisted tools and bug bounties to find the critical vulnerabilities that truly impact business risk, and then provide the human context needed to get them fixed.

Key Takeaways for CISOs

1

Demand transparency

Mandate that vendors disclose their use of AI and the data behind it. Do not accept a black box solution.

2

Prioritize expertise

Mandate that vendors disclose their use of AI and the data behind it. Do not accept a black box solution.

3

Adopt a programmatic approach

Move beyond ad-hoc, AI-only scanning to a structured offensive security program that integrates the best of both human and machine intelligence.

About Cobalt

Cobalt is the pioneer in pentesting as a service (PTaaS) and a leader in offensive security services.

We are focused on combining talent and technology with speed, scalability, and expertise. Thousands of customers and hundreds of partners rely on the Cobalt Offensive Security Platform, along with the industry's largest exclusive community of 450+ trusted pentesters and security experts, to find and fix vulnerabilities across their environments. By enabling faster pentest launches, real-time collaboration with testers, continuous scanning, and seamless integration with remediation workflows, we help organizations identify critical issues and accelerate risk mitigation so they can operate fearlessly and innovate securely.

Discover how human-led, AI-powered pentesting from Cobalt can help you secure your AI apps and LLMs

BOOK A DEMO