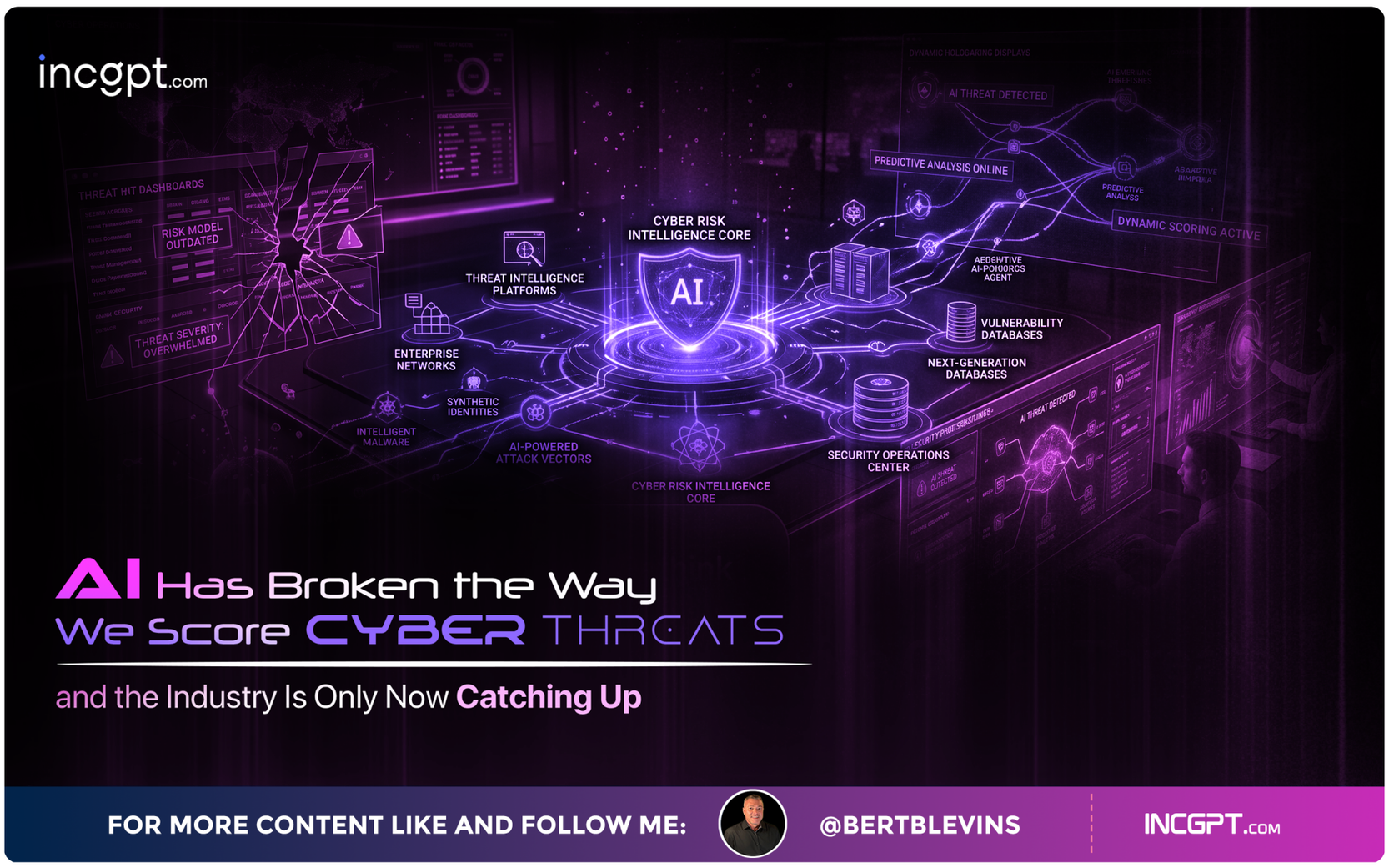
AI Has Broken the Way We Score Cyber Threats — and the Industry Is Only Now Catching Up



For decades, threat intelligence teams have leaned on a simple intuition: the more skilled the attacker, the more dangerous they are. Frameworks were built around that assumption, sorting adversaries by the breadth of techniques they could pull off and the sophistication of the tools they wielded. New research suggests that intuition no longer holds. Artificial intelligence has severed the old link between an attacker's personal skill and the damage they can actually do — and the scoring models the industry relies on are quietly failing as a result.

The findings come from Anthropic's Frontier Red Team, which analyzed 832 accounts banned for malicious cyber activity over a twelve-month window running from early 2025 into 2026. Each case was mapped against MITRE ATT&CK, the taxonomy most security teams use to catalog attacker tactics and techniques. A slice of the results also fed into Verizon's 2026 Data Breach Investigations Report. Taken together, the analysis lands on three conclusions that should make any team relying on traditional risk scoring uneasy.

AI Is Pushing Deeper Into the Attack Chain

The single most common use of AI in the dataset was the one everyone expected: writing malicious code. Roughly two-thirds of the banned accounts — 560 of 832 — used AI to help build malware. But the headline isn't the volume. It's the direction of travel. Over the year studied, AI assistance migrated away from the early, noisy stages of an attack and toward the difficult work that happens after a foothold is already established.

01

Consider the shift in specifics. AI-assisted discovery of valid accounts inside a compromised environment climbed nearly nine percent over the period, while AI-assisted phishing — a classic way in — dropped by roughly the same amount. Lateral movement, the painstaking process of working through a network toward the crown jewels, showed up with AI assistance in about one in fifteen of the actors studied. These are exactly the maneuvers that used to separate elite operators from amateurs. AI is handing them to everyone.

02

The risk data makes that leveling-up tangible. In the first half of the study, about a third of actors rated medium risk or higher. By the second half, that figure had surged to well over half — close to a 1.7-fold jump in just six months. The floor, in other words, is rising fast.

The Signals We Used to Trust Have Come Unmoored

Historically, analysts gauged an attacker's sophistication by tallying how many distinct techniques they used and noting which tools or interfaces they favored. The research shows both of those signals have decoupled from real danger in an AI-enabled world.

01

The numbers are striking. The least capable actors in the dataset averaged sixteen distinct techniques; the most capable averaged twenty. A gap that narrow is useless for triage — you cannot meaningfully sort threats by a difference of four techniques. The interface told you nothing either: whether an actor worked through an agentic coding tool, a raw API, or a simple chat window showed no relationship to how risky they actually were.

02

What did still correlate with risk was placement — where in the attack lifecycle the AI was applied. The more dangerous operators concentrated AI on the operationally demanding phases: account discovery, lateral movement, privilege escalation, rather than just getting in the door. But even that tell is fading, because the broader criminal population is now drifting toward those same post-compromise techniques. As the behavior spreads downward through the ecosystem, the signal that once flagged the elite is becoming background noise.

The Real Differentiator Is Architecture, Not Skill

If technique counts and tooling no longer separate the dangerous from the ordinary, what does? According to the research, the answer is structural. The highest-risk actors build scaffolding around AI models — frameworks that let the AI stitch together the discrete stages of an attack and run them with barely any human steering. That capability, agentic attack orchestration, is the genuine frontier of AI-enabled threat activity. And critically, it is nowhere to be found in the MITRE ATT&CK framework.

01

Anthropic illustrated the gap with a state-linked espionage operation it disrupted in late 2025, in which an adversary coaxed an AI coding agent into attempting to compromise targets around the world with minimal human involvement. Mapped against MITRE ATT&CK, that operation registered as thirty techniques across thirteen tactics — a profile that looks like an unremarkable medium-risk actor. Run through a risk-scoring method designed to capture orchestration, the very same operation maxed out at a score of one hundred. The framework saw a routine intrusion; the reality was a top-tier threat.

02

The mismatch exists because ATT&CK was built to describe what attackers do, not how they coordinate it. An AI agent that runs commands, exploits weaknesses, harvests credentials, and makes tactical decisions on the fly across an entire campaign — pausing for a human only at a handful of checkpoints — is a fundamentally different kind of adversary than a person performing those same steps by hand. There is no ATT&CK identifier for agentic orchestration, no entry for autonomously chaining attack stages together, no tactic that captures what happens when the human decision bottleneck is simply removed.

What This Means for Defenders

The practical takeaway is uncomfortable: triage models built on technique counts, tool fingerprints, or initial-access polish are now systematically underrating AI-enabled actors. An adversary using sixteen AI-assisted techniques may be every bit as dangerous as one running twenty-five by hand. An attacker on a free chat tier may be driving the same autonomous attack chain as one wired into a direct API. Judging the book by those covers no longer works.

01

The questions that actually matter now are behavioral and architectural. Is the actor applying AI after the breach rather than just to get in? Is there evidence of automated chaining between stages of the attack? Is human input being engineered out of the demanding steps? Those questions are not yet baked into mainstream detection frameworks — and that gap is the urgent problem. Anthropic says it is in talks with MITRE about evolving ATT&CK to account for these behaviors, and has used the findings to harden the safeguards in its own models, including detection and blocking for malware creation and mass data theft.

02

For security leaders, the message is to stop scoring adversaries by how clever they appear and start asking how much of their attack is being automated. In an AI-enabled threat landscape, the most dangerous operator in your logs may also be the one who looks the most ordinary.
Read more on the link below.