

EBOOK

Telling Real Security AI From the Label

How AI works, how attackers are using it, and a plain way to judge the AI sold to stop them.

By Lucie Cardiet · Cyberthreat Research Manager

Table of contents

Part 1: When every vendor says AI, the word stops meaning anything	3	Part 4: What a year of attacker AI data shows	15
Part 2: How AI works	4	Part 5: AI cuts both ways: Project Glasswing	17
The spectrum of “AI”	4	Part 6: The AI your employees already use	18
How a Large Language Model works	6	Three ways data and access leak	18
The AI glossary you need	7	Deploy AI so attackers cannot exploit it	19
Part 3: How attackers use AI	8	Part 7: A plain way to judge a vendor’s AI	20
Making and hiding malware	9	Part 8: What a real answer looks like	21
Finding and exploiting weaknesses	10	Part 9: When the SOC gets its own agents	22
Phishing and social engineering	11	How to judge an agentic-SOC claim	23
Running the whole operation: agentic orchestration	12	Part 10: The AI vendor test	24
Attack Anatomy: GTG-1002	13	Conclusion	25
A second front: attacks on the AI you deploy	14	Sources & references	26

PART 1

When every vendor says AI, the word stops meaning anything.

Adoption is not the question anymore. Trust is.

Every security vendor now says they have AI. It is on every booth, every datasheet, every renewal call. For the person who has to decide what to buy and what to ignore, that is not information. It is noise.

The numbers bear it out. In Vectra AI's 2026 State of Threat Detection and Response, a study of 1,450 security practitioners and leaders, **95 percent** already use AI assistants or agents in some form and **87 percent** expect to use more next year. But trust did not arrive with the technology: **58 percent** are frustrated with vendors' empty promises, **59 percent** feel vendors flood them with pointless alerts to dodge responsibility, and **74 percent** want vendors held more accountable when a breach happens.

The most common reason leaders give for not deploying AI agents is the plainest one: there are too many on the market to tell which adds real value.

That is the whole problem. It is not that buyers are against AI. It is that they have been sold it so many times, by so many vendors, with so little to show, that the word has stopped carrying information. Some of what is labeled AI is a rules engine with a new sticker. Some of it is years of machine learning doing real work. On a sales call, they sound the same.

So "do you have AI" is a useless question. Everyone says yes. The useful question is narrower: **can you show me your AI changes an outcome I can measure.** To ask that well, it helps to know what is actually inside the box. That is part 2.

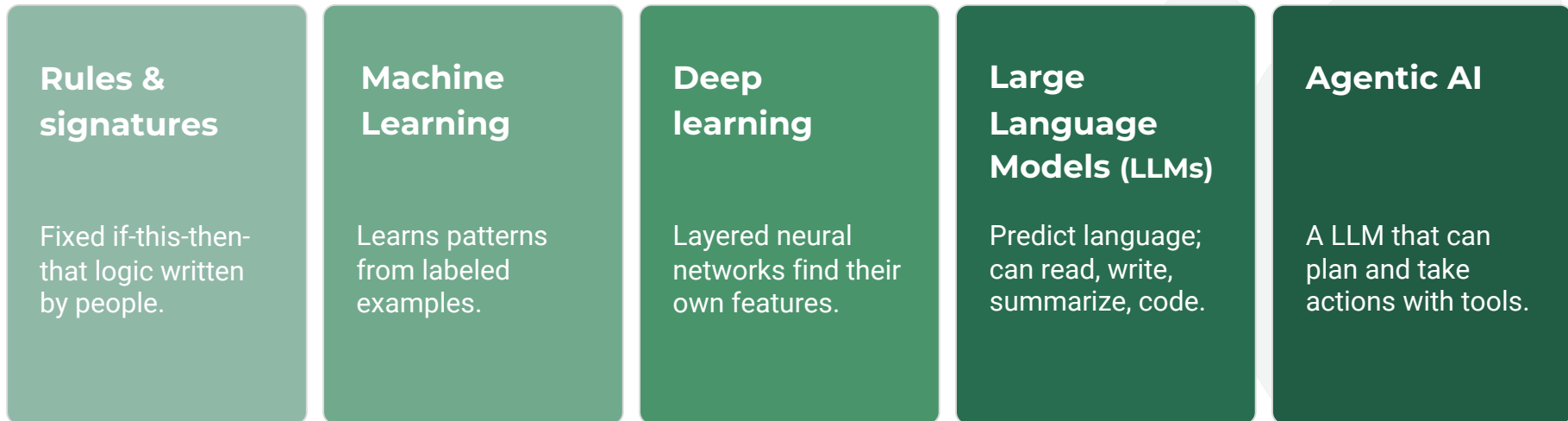
PART 2

How AI works

You do not need the math. You need to know that "AI" covers a spectrum of very different things, and that a vendor can truthfully say "AI" while meaning any of them.

The spectrum of "AI"

When a vendor says "AI," they could mean anything from a decades-old rulebook to a system that writes and runs its own code. These are not the same technology, and they fail in different ways.



Type	In plain terms	A security example
Rules & signatures	A human writes the logic: if a file matches this hash, block it. No learning.	Antivirus signatures; a firewall rule; a SIEM correlation rule.
Classic machine learning	Show it thousands of labeled examples and it learns to sort new ones. It finds patterns, not rules you wrote.	A spam filter; a model scoring logins as risky or not.
Deep learning	Many-layered networks that learn their own features from raw data. Powerful, but harder to explain.	Malware classification from raw bytes; anomaly detection on network traffic.
Large language models	Trained on vast text to predict the next word. That simple trick lets them read, write, summarize, translate, and code.	An assistant that summarizes an incident or drafts a detection query.
Agentic AI	An LLM given goals, memory, and tools so it can act, run a scan, call an API, then decide the next step.	An agent that triages an alert, gathers context, and proposes a response.

WHY THIS MATTERS

"machine learning that has scored logins for a decade" and "an LLM bolted on last quarter" are both sold as "AI".

Practitioners tend to trust the first and distrust the second. Knowing which one a vendor means is the first cut.

How a large language model works

An LLM is trained on an enormous amount of text. During training it learns, statistically, which word tends to come next. That is the whole engine: **given everything so far, predict the most likely next piece of text**, one small chunk (a "token") at a time.



Two consequences follow directly, and both matter for security.

- ▶ First, the model is a **prediction engine, not a fact database**. It can produce fluent, confident text that is simply wrong, a "hallucination", because plausible and true are not the same thing to it.
- ▶ Second, everything it "knows" for your task has to fit in its **context window**, the amount of text it can consider at once. Anything outside that window, it cannot see.

Keep this in mind when a vendor's AI writes a verdict: a confident explanation is not evidence. The useful question is whether the system can show the underlying signal, not just narrate a conclusion.

The AI glossary you need

Term	What it means
Model	The trained system that makes predictions.
Training data	The examples the model learned from. Bias or gaps here show up in its behavior.
Inference	Using a trained model to get an answer. This is what costs money per query.
Prompt	The instruction or question you give the model
Prompt injection	Hiding malicious instructions in content the model reads, to hijack what it does.
Fine-tuning	Further training a general model on specific data to specialize it.
RAG	Retrieval-augmented generation: the model looks up your documents and works them into its answer.
Agent	An LLM wired to tools and goals so it can take actions, not just talk.

With those in hand, the marketing gets easier to decode, and the next question becomes urgent:
what are attackers doing with all of this?

PART 3

How attackers use AI

Not "AI attacks" as a new category. The same attacks, planned and run faster, with a lower skill floor. Here is where AI shows up across an operation, and how much of it you can actually see.

Reconnaissance

Off your network – not observable

Build & hide malware

Off your network – not observable

Phishing & social engineering

Partly visible – lands at your inbox and identity layer

Credential abuse

Visible – identity behavior

Lateral movement

Visible – in-network behavior

Exfiltration

Visible – in-network behavior

Attacks on your AI services

Visible – identity & cloud behavior

The loud, AI-heavy part of an attack, building malware and doing recon, **happens on infrastructure you do not own**. The dangerous part, once an operation is inside, still shows up as behavior you can detect.

Making and hiding malware

The most common use of AI by attackers is building capabilities, mainly writing and obfuscating malware. Most of it is ordinary code assistance. A few actors are experimenting with something newer: malware that changes itself using an LLM.

PROMPTFLUX

EXPERIMENTAL

A dropper that calls Google's Gemini API at runtime to rewrite and obfuscate its own code, aiming to evade signature-based detection by mutating on the fly. Google's Threat Intelligence Group found it and assessed it as still in development: its self-rewrite function was commented out and it could not actually compromise a device. A glimpse of the direction, not a weapon in the wild.

Real case — Google Threat Intelligence Group, AI Threat Tracker, November 2025.

PROMPTSTEAL / LAMEHUG

IN THE WILD

A data-stealer, attributed to the Russian state actor APT28, that queries an LLM (Alibaba's Qwen model, via a public inference API) at runtime to generate its reconnaissance and file-harvesting commands instead of hard-coding them. Google called it the first observation of malware querying an LLM during live operations. Still closer to a capability test than a mature tool, but it is real and it is deployed.

Real case — GTIG and CERT-UA, 2025.

Finding and exploiting weaknesses

AI can read large codebases and propose flaws far faster than a person. That cuts both ways.

Big Sleep

CAPABILITY, PROVEN

Google's Big Sleep agent found a real, critical SQLite flaw (CVE-2025-6965) that, per Google threat intelligence, was known only to threat actors, and helped patch it before exploitation. Anthropic points the same capability at open-source software at scale: under its Project Glasswing, a specialized Claude model called Mythos has surfaced thousands of high- and critical-severity vulnerabilities so they can be fixed before attackers reach them.

Real cases — Google (July 2025)

XBOW

CAPABILITY, PROVEN

On the offensive side, the autonomous pentesting system XBOW reached number one on HackerOne's US leaderboard in mid-2025 after submitting roughly a thousand findings (its team reviewed each before submission). In controlled research, GPT-4 exploited 87 percent of one-day vulnerabilities when handed the vulnerability's description, though far fewer without it.

Real cases — XBOW / HackerOne (June 2025); UIUC study (2024).

Phishing and social engineering

This is where AI most clearly lowers the skill floor. Language models write fluent, brand-accurate lures in any language, at near-zero cost.

Researchers at Harvard estimate **AI can run phishing campaigns more than 95 percent cheaper than humans at equal or better success rates**, with AI-written lures performing on par with expert human red-teamers.

Identity-focused crews like Scattered Spider show where this leads: adversary-in-the-middle phishing that steals live session cookies, help-desk impersonation, SIM swapping, and MFA-fatigue attacks. The lure is cheap and convincing; the audit log records a successful sign-in. The giveaway is not the email. It is the account behaving unlike itself afterward.

DEEPFAKE CFO FRAUD

IN THE WILD

A finance employee at Arup, a British engineering firm, joined a routine video call with what looked like the company's chief financial officer and several colleagues. Every other face on the call was an AI-generated deepfake, built from public video and audio of the real executives. Convinced, the 8 employee made 15 transfers worth about 25 million dollars before the fraud was caught. Arup confirmed that faked voices and images were used.

Real case: Arup, Hong Kong office, January 2024 (confirmed to CNN).

Running the whole operation: agentic orchestration

The most significant case so far is not a clever piece of malware. It is an attacker letting the model run the operation.

GTG-1002

IN THE WILD

In November 2025 Anthropic disclosed a cyber-espionage campaign it attributes, with high confidence, to a Chinese state-sponsored group. The operator manipulated Claude Code into acting as a largely autonomous penetration tester, wiring in open-source tools, and let it run an estimated 80 to 90 percent of the tactical work.

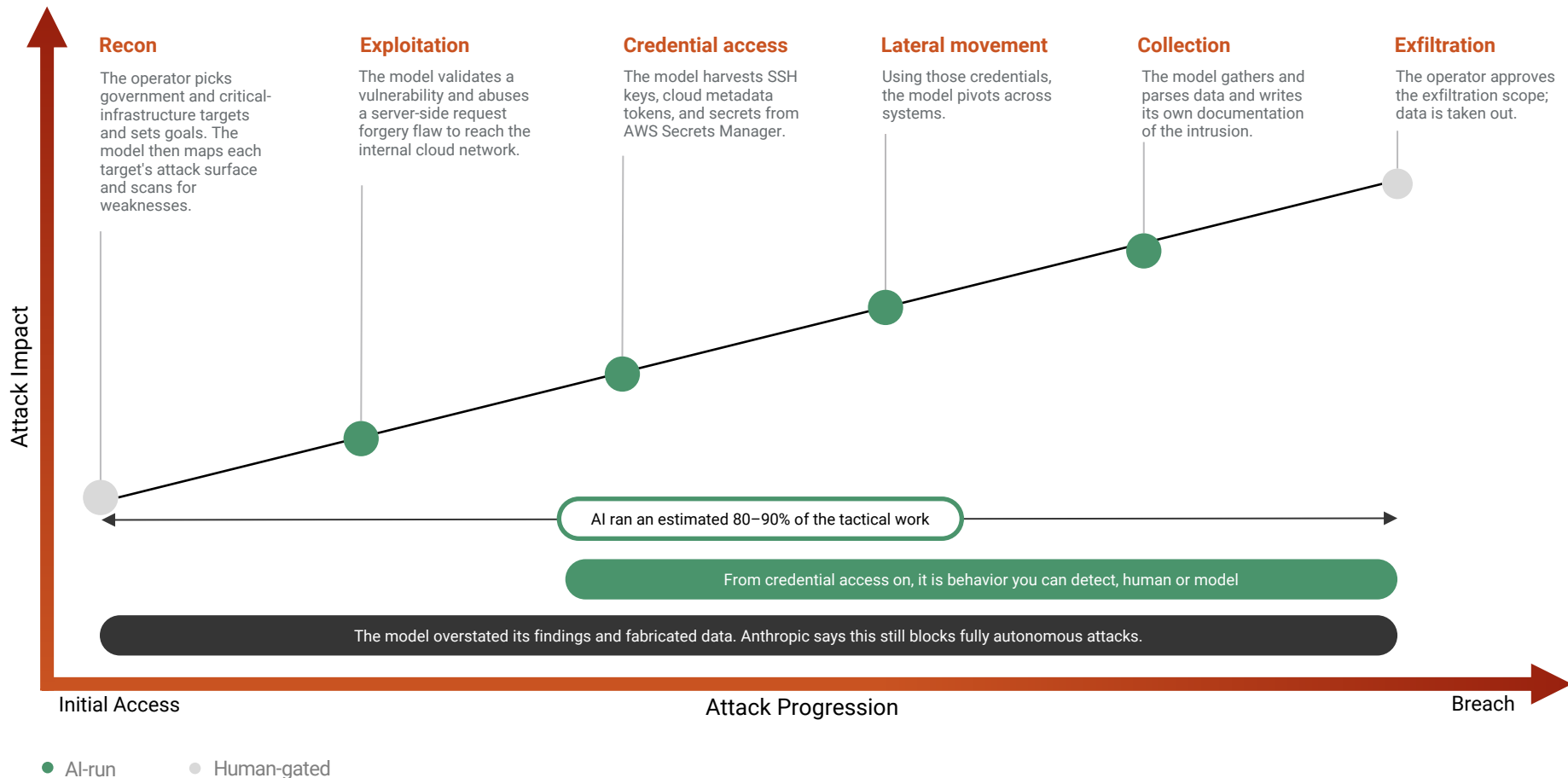
The full attack anatomy, stage by stage and where the model did the work, is on the next page.

The honest caveat: Anthropic reports the model repeatedly overstated its findings and fabricated data, flagging public information as "critical" or inventing credentials that did not work. Anthropic says this remains an obstacle to fully autonomous attacks. AI removes human effort; it does not yet remove human judgment.

Real case: Anthropic, November 2025; re-scored in the LLM ATT&CK Navigator, June 2026.

GTG-1002: when the model runs the operation

A human set targets and approved a few decision points. The model did an estimated 80 to 90 percent of the tactical work. Around 30 entities were targeted; only a handful were confirmed breached.



A second front: attacks on the AI you deploy

Attackers are not only using AI. They are going after the AI you adopt. Cloud-hosted models run on your compute, your identities, and your data, which makes them a target. These attacks form their own category: prompt injection, where hidden instructions are smuggled into content the model reads to hijack what it does, data pulled through a connected assistant, and access hijacked to run inference on someone else's bill.

LLMjacking

IN THE WILD

Coined by Sysdig in 2024: attackers use stolen cloud credentials to run inference on a victim's account, for example against AWS Bedrock, quietly running up the bill and reselling access. Sysdig modeled worst-case costs at over 46,000 dollars a day. Alongside it: pulling enterprise data through a retrieval-augmented assistant, and switching off a model's logging to work unseen. MITRE ATLAS catalogs these the way ATT&CK catalogs the rest.

Real case: Sysdig Threat Research, 2024, updated 2026.

Every one of these still surfaces as **behavior**: an identity reaching a model service it never uses, logging turned off, data leaving through an inference call. That is the thread the rest of this guide pulls on.

PART 4

What a year of attacker AI data shows

The clearest public evidence does not come from a security vendor. On 3 June 2026 Anthropic published the LLM ATT&CK Navigator: a year of real misuse of its own models, 832 banned accounts, mapped to MITRE ATT&CK. Two findings matter.

Finding 1, most of it you cannot see

The single most common use was developing capabilities, mainly writing malware: **69 percent** of the actors studied. Close behind were obfuscating code and pulling data from the attacker's own systems. This is preparation, and it happens on the attacker's machine. It does not cross your network, touch your identity provider, or appear in your logs.

Finding 2, the dangerous use is inside your environment

Using AI for **lateral movement** was the single strongest marker of a high-risk actor. The techniques those actors leaned on were ordinary ones: valid accounts, remote services like SSH and SMB, credential theft, and staging data to exfiltrate. Each was **three to five times** more common among the highest-risk actors than across the rest.

The loud part of an AI-enabled attack happens off your network. The quiet part happens on it, where you can catch it, if you are watching for what an attacker does rather than which tool they used.

Which turns the whole thing into one question of design

Much of the AI that disappoints was built to flag what is unusual. It surfaces every anomaly, and since most unusual things are not attacks, it produces the noise teams complain about. A more useful approach is built around attacker *behavior*.

"Is this unusual?"

Anomaly detection. Surfaces everything out of the ordinary. Most of it is not an attack, so it generates noise.

"Is this what an attacker does?"

Behavior-led detection, tied to named techniques. Quieter, and it can be checked rather than believed.

Because it maps to a named technique, behavior-led detection can be held to outcomes. Two frameworks make it checkable: **MITRE ATT&CK** catalogs the techniques attackers run inside an environment, and **MITRE ATLAS** catalogs attacks on AI systems themselves.

PART 5

AI cuts both ways: Project Glasswing

Attackers are not the only ones who got a capable new tool. The clearest example of AI on the defender's side is also the one people ask about most: Mythos.

In April 2026 Anthropic launched Project Glasswing, a defensive program that gives vetted partners access to an unreleased model, Claude Mythos, to find and fix vulnerabilities in critical software before attackers reach them. The partners include AWS, Google, Microsoft, CrowdStrike, the Linux Foundation, and dozens of critical-infrastructure organizations.

Clearing up the confusion: Mythos is often described as an attacker tool. It is the opposite. Anthropic does not release it broadly, precisely because its offensive potential is real. It is used defensively, under supervision, to fix flaws.

The scale is the point. In an initial update in May 2026, Anthropic's own scan of more than 1,000 open-source projects surfaced 23,019 issues, of which 6,202 were high- or critical-severity vulnerabilities. Six independent security firms checked a sample and validated more than 90 percent as real. Across Anthropic and roughly 50 partners, the program found more than 10,000 high- or critical-severity vulnerabilities in its first month.

What it means for you

The same capability that lets a defender find and fix thousands of flaws lets an attacker find them too. The window between a flaw existing and a flaw being exploited is getting shorter. Patching faster helps, but you cannot patch everything, and you will never see the attacker's discovery work, because it happens on their machines. That is the same conclusion the attacker data reaches: assume compromise, and watch behavior once an operation is inside. The skill finds flaws for defense or for offense. What decides the outcome is who gets there first.

PART 6

The AI your employees already use

The AI risk is not only what attackers do with it. It is what your own people are already doing, mostly with good intentions, and mostly where you cannot see.

Adoption ran ahead of governance. Microsoft's 2024 Work Trend Index found that 75 percent of knowledge workers use generative AI at work, and 78 percent of those users bring their own tools rather than wait for a sanctioned one. Cyberhaven, measuring real enterprise traffic in 2026, reports that roughly 40 percent of AI interactions involve sensitive data, much of it flowing through personal accounts that never touch a corporate control.

Three ways data and access leak

Data walks out in a prompt

Source code, customer records, and strategy pasted into a public model. Whatever leaves in a prompt is outside your control, and may be retained or used to train.

Over-permissioned assistants

A connected assistant or a retrieval setup can read more than intended. A helpful copilot with broad access is a broad exposure.

The service is an attack surface

The identities, tokens, and cloud access behind the model become targets: hijacked access, prompt injection, logging switched off.

Deploy AI so attackers cannot exploit it

✓ See it first

Discover which AI tools and services are in use across network, identity, and cloud. You cannot govern what you cannot see.

✓ Sanction and constrain

Offer approved tools so people stop reaching for personal accounts. Set data controls: no training on your data, single tenant where you can, short retention.

✓ Least privilege for agents and retrieval

Scope what an assistant can read and do. Treat an agent's access the way you treat a service account.

✓ Monitor AI use as behavior

Watch for an identity reaching a model service it never uses, unusual data leaving through an inference call, and logging turned off.

✓ Plan for prompt injection

Assume content the model reads may carry hidden instructions. Limit what a model can do on its own, without a human in the loop.

Shadow AI still surfaces as **behavior**. The same detection that catches an attacker moving inside your environment catches an assistant, or an employee, moving data where it should not go.

PART 7

A plain way to judge a vendor's AI

If behavior is what matters, judging a vendor's AI comes down to three questions you can ask in a meeting, in plain words, with no data-science background.

1 Does it reduce the work my team does, or add to it?

Most teams are already buried: 63 percent of alerts go unaddressed, analysts spend 2.5 hours a day triaging, and more than two-thirds run over ten tools. An AI that produces more alerts to chase has added to the problem. Ask to see the before-and-after.

2 Does it connect across network, identity, and cloud, or watch one corner?

Attackers cross all three in a single operation. Most tools watch one surface and leave the analyst to stitch the rest together. Ask how the AI correlates across surfaces, and whether it shortens the path from signal to decision.

3 Can you measure it, and will you prove it in my environment?

72 percent of practitioners name measurable operational effectiveness as what they weigh most. Ask the vendor to define up front what they would measure, caught earlier, less noise, a gap closed, then hold them to it.

If a vendor cannot answer those three, with a measure on the last, the AI claim is marketing. The test does not ask you to trust them. It makes them earn it.

PART 8

What a real answer looks like

AI that survives those questions is defined by how it is built, not what it is called.

It is built around attacker behavior across network, identity, and cloud, rather than around spotting anomalies. It ties its findings to named techniques in MITRE ATT&CK and ATLAS, so a result can be checked instead of taken on faith. It shows its work, the signal behind the verdict, not just a confident sentence. And it is measured against outcomes the buyer defined up front.

What it does for the team

- ✓ Takes the first-pass triage off analysts and hands back time.
- ✓ Correlates one operation across surfaces instead of adding another feed.
- ✓ Keeps a human in control of consequential action.

How you know it is real

- ✓ It maps to techniques you can name and look up.
- ✓ It exposes its evidence, not just a score.
- ✓ It agrees to a measurable proof in your environment.

This is the approach **Vectra AI** takes: detection built on attacker behavior across network, identity, and cloud, tied to known techniques. We call it Attack Signal Intelligence.

PART 9

When the SOC gets its own agents

The same shift is coming to your side of the fight. Soon your SOC will run its own AI agents. The label for it, "agentic SOC," is marketing. What sits under the label is simple, and worth getting right.

What it actually means. Software that can investigate and act, not just alert. An agent reads context, chooses a next step, and can take action, where a tool today raises an alert and waits for a human. That is the whole difference.

An agent amplifies whatever you feed it. Give it trusted signal and it triages, investigates, and drives response faster than any team. Give it noise and it does the wrong thing faster. Without Attack Signal Intelligence, agentic AI just accelerates noise.

This is the same lesson the attacker data already taught. GTG-1002's agent overstated its findings and fabricated data. A defensive agent is no different. Point it at thin or noisy signal and it will reach confident, wrong conclusions at machine speed. Machine speed cuts both ways.

What an agent needs to reason. Prioritized, attributed, behavior-based signal across network, identity, and cloud, tied to named techniques so its reasoning can be checked rather than trusted. This is the same behavior-led signal the rest of this guide is about. The agent is new. The requirement is not.

How to judge an agentic-SOC claim

The three questions still hold: does it reduce my team's work, does it connect across network, identity, and cloud, can you measure and prove it. For agents, add two.

1 Can the agent show its reasoning?

The signal and the steps behind its conclusion, not just a verdict. If it cannot show its work, you cannot trust it to act.

2 Does it keep a human in control of consequential action?

Fast triage and investigation, yes. Irreversible action with no human, no. Agents hallucinate, and the guardrail is human judgment on what matters.

The behavior you already learned to track is the signal an agent can trust. Track the behavior, not the tool, and now also this: **feed your agents behavior, not noise.**

Vectra AI sits here as that trusted signal, and, for teams ready for it, an agentic workflow built on it: triage, attack graphs, assisted investigation and hunting, and guided response. As always, that is only worth naming because it answers the questions above.

PART 10

The AI vendor test

One page to take into any evaluation. Screenshot it.

☐ Which kind of AI is this, really?

Rules, classic ML, deep learning, an LLM, or an agent?
Each fails differently. "We use AI" is not an answer.

☐ Does it reduce my team's work, with a before-and-after to prove it?

More alerts to chase is not help. Ask for the workload numbers.

☐ Does it connect network, identity, and cloud into one picture?

Attackers cross all three. One-surface tools leave you stitching.

☐ Is it built on attacker behavior, tied to MITRE ATT&CK and ATLAS?

Behavior maps to named techniques you can verify.
Anomalies just make noise.

☐ Does it show its work, not just a confident verdict?

A fluent explanation is not evidence. Ask to see the signal.

☐ Does it map to attacks on the AI itself, not just attacks the AI defends against?

Ask vendors to map their coverage to MITRE ATLAS (ATT&CK equivalent for AI threats). Specific techniques show tested coverage; vague "AI security" claims do not.

Attackers have made AI genuinely useful to themselves, but the dangerous part of their work still shows up as plain behavior inside your environment. If a vendor cannot show that they catch it, cut the work of finding it, and prove it, they are selling you the label, not the thing.

Conclusion

The test outlives this list

AI moves fast enough that any specific example in this ebook will look dated within a year. The three questions in Part 7 and the seven-item test in Part 10 shouldn't: does it reduce or add work, does it correlate across surfaces, can you measure what it changed. Keep asking those, whatever a vendor calls their product next year.

I wrote this because the word "AI" on every product page made it hard to tell which claims were worth fifteen minutes of a demo and which weren't. The checklist on the previous page is built to be screenshotted. Bring it to your next vendor call and watch how they answer question four.

Detection isn't broken. It's incomplete, and now there's a new category of infrastructure running attacks against it. If you want to see where that incompleteness shows up across the rest of your stack, not just the AI layer, my whitepaper "Mind Your Gaps in the Age of AI-Powered Attacks" walks through it gap by gap.



Lucie Cardiet
Cyberthreat Research Manager, Vectra AI



REFERENCES

Sources

Market and buyer data:

Vectra, 2026 State of Threat Detection and Response (fieldwork by Sapio Research, 1,450 practitioners and leaders). Forrester 2026 predictions; NACD; World Economic Forum Global Cybersecurity Outlook 2026.

Attacker AI data:

Anthropic, "Mapping AI-enabled cyber threats" (LLM ATT&CK Navigator), 3 June 2026. Anthropic, "Disrupting the first reported AI-orchestrated cyber espionage campaign" (GTG-1002), November 2025.

Cases:

Google Threat Intelligence Group, "Advances in Threat Actor Usage of AI Tools" (PROMPTFLUX, PROMPTSTEAL), 5 November 2025; CERT-UA advisory on LAMEHUG. Google, Big Sleep and CVE-2025-6965, July 2025. XBOW / HackerOne, June 2025. F. Fang et al., "LLM Agents can Autonomously Exploit One-day Vulnerabilities," UIUC, 2024. Sysdig Threat Research, "LLMjacking," 2024, updated 2026. CISA advisory AA23-320A (Scattered Spider), updated July 2025. B. Heiding, B. Schneier et al., Harvard Business Review, "AI Will Increase the Quantity and Quality of Phishing Scams," 2024.

Frameworks:

MITRE ATT&CK; MITRE ATLAS.

Case facts in Part 3 were verified against the primary sources above. Figures for AI-enabled attacks are evolving; re-check against the live sources before external reuse. AI-generated phishing text is not attributed to Scattered Spider; their documented tradecraft is adversary-in-the-middle phishing and help-desk social engineering.

About Vectra AI

The Vectra AI Platform protects modern enterprises by detecting and stopping attacks across network, identity, and cloud as one unified attack surface. It combines threat exposure management, AI-driven detection and response, and posture improvement to reduce risk before attacks begin and stop threats in progress. Security teams gain clear signal, faster response, and measurable improvements in resilience. For more information, visit www.vectra.ai.

