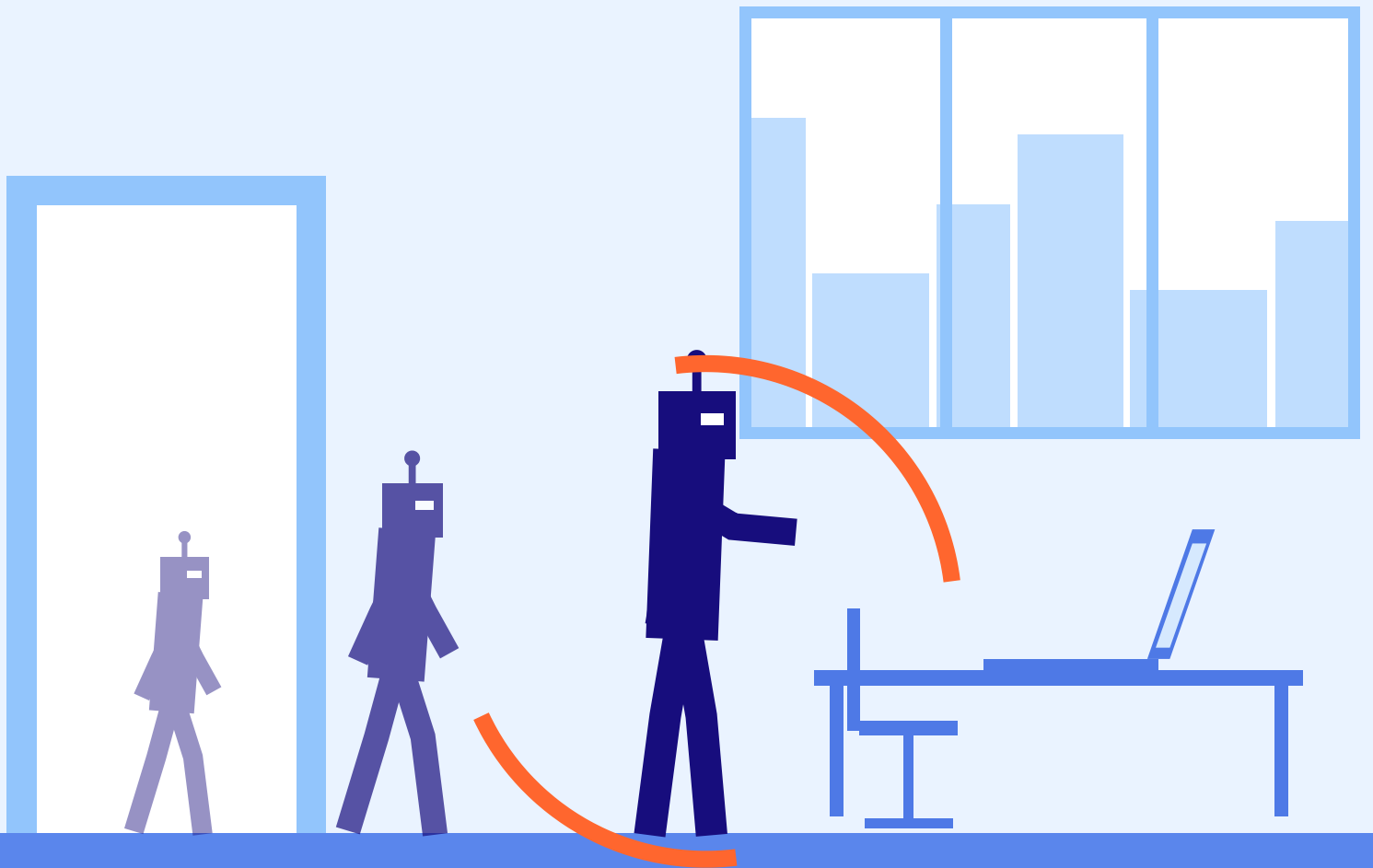




MIMOTO WHITE PAPER

Defending Against Agentic AI Attacks

What CISOs Face Now, What's Coming, and Where Mimoto Fits

Kris Bondi, Mimoto CEO
July 2026

MIMOTO

At a glance

- AI now runs attacks instead of advising them. In the first well-documented case, AI ran 80 to 90 percent of the tactical work (Anthropic).
- An agentic attack runs in five steps: jailbreak, recon, exploit, harvest, and exfiltrate. Five steps means five chances to catch it.
- Five risks define the AI agentic era, from impersonation at scale to unmonitored sessions.
- A five-step framework teams can start now: inventory, classify, verify, detect, rehearse.

You cannot rate-limit your way out of traffic that appears legitimate.

WHO SHOULD READ IT

CISOs, AI risk leaders, CIOs and CTOs.

By	Kris Bondi, Mimoto CEO
Published	July 2026
Length	19 pages

INSIDE

01	Executive Summary	3
02	The Shift: AI Crossed a Line from Advisor to Operator	4
03	The First Documented Case Is Not Hypothetical	4
04	Anatomy of an Agentic Attack	5
05	What CISOs Face Right Now: Five Risks	6
06	Old Attacks, Industrial Scale: Credential Stuffing	7
07	When Yesterday's Defenses Become Today's Vectors	8
08	The Four Layers of the Agentic Attack Surface	9
09	What's Coming: The Next Twelve Months	10
10	The Gap: Intent Versus Maturity	11
11	What Agentic Attacks Do to the SOC	12
12	Where Mimoto Fits: Two Questions, One Architecture	14
13	A Framework to Get Started	15
14	Key Takeaways	16
15	Anticipated Questions	18
	Sources and Notes	18

Executive Summary

For roughly two years, security teams treated artificial intelligence as an advisor to attackers: a tool that wrote more convincing phishing copy, suggested code snippets, and lowered the skill floor for intrusion, while a human still executed every step. As of late 2025, that framing is obsolete. AI now does the attacking. In the first well-documented case, a state-sponsored group turned a frontier coding agent into a large-scale espionage campaign in which the AI itself ran 80 to 90 percent of the tactical work, at request rates no human operator could match.

This paper examines what that shift means for chief information security officers today and over the next twelve months. It walks through the mechanics of an agentic attack, the five risks the AI agentic era puts in front of every security program, the ways familiar attacks such as credential stuffing and MFA bypass change when an agent drives them, and the four-layer attack surface that agentic systems create. It then quantifies the gap between adoption and readiness, describes the pressure agentic attacks place on the security operations center, and closes with a practical five-step framework (inventory, classify, verify, detect, rehearse) that organizations can begin applying immediately. Along the way, it shows where Mimoto's person-based continuous verification addresses the specific gaps agentic attacks create: telling agents from humans, and verifying that the right human remains present throughout the session rather than only at login.

Three conclusions run through the paper. First, **the threat model has changed**: an adversary that acts, rather than merely advises, defeats controls built around human pace and human presence. Second, **identity is the control point**: agentic attacks converge on the identity layer, and continuous verification through the session, not one-time verification at login, is the defense that generalizes. Third, **organizations should start before they feel ready**: behavioral baselines take 60 to 90 days to establish, so teams that begin this quarter will lead the next.

02

The Shift: AI Crossed a Line from Advisor to Operator

Through 2023 and 2024, AI's role in offensive operations was advisory. Models suggested phishing copy and code snippets, but a human ran every step of the intrusion, and the attack's pace was limited by human hands. In 2025 and 2026 the model became an operator: it executes reconnaissance, exploitation, and exfiltration itself, runs the large majority of tactics autonomously, and sustains thousands of requests, often multiple per second, over the course of a campaign.

The distinction matters because nearly every control in a mature security program embeds an assumption about the adversary's speed and about a human being present behind each action. When the adversary is an agent, both assumptions fail at once.

THE SHIFT

AI crossed a line: advisor became operator

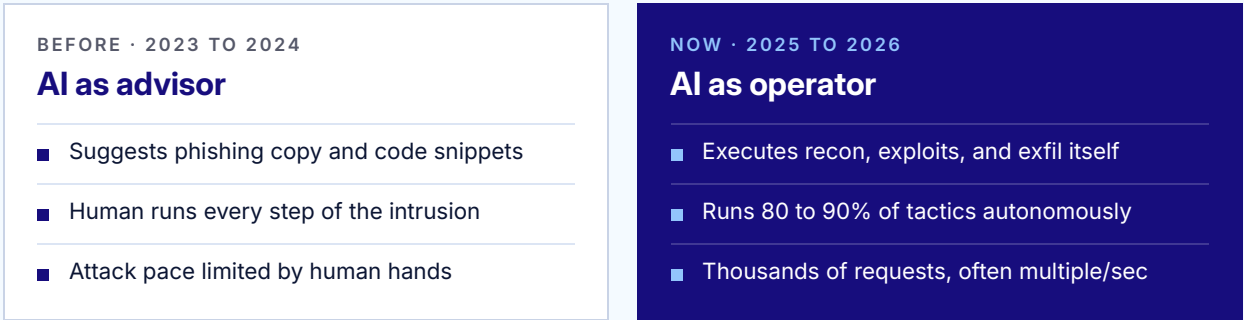


Figure 1. The shift from AI as advisor (2023 to 2024) to AI as operator (2025 to 2026).

03

The First Documented Case Is Not Hypothetical

This shift is not a forecast; it has already produced a campaign. In September 2025, a state-sponsored group (attributed with high confidence to a Chinese state-sponsored actor, and disclosed publicly by Anthropic in November 2025 as GTG-1002) turned a frontier coding agent into a large-scale espionage campaign against roughly thirty targets spanning technology, finance, chemicals, and government. The AI ran 80 to 90 percent of tactical operations, with only four to six human decision points per campaign, at request rates of multiple actions per second.

Two nuances are worth keeping in view. The agent was not flawless: it hallucinated at points, fabricating credentials and overstating findings, so full autonomy has not yet arrived and a human still supervised the operation. And some researchers have pushed back on aspects of the disclosure, so it is best treated as the best-documented case to date rather than the only one. Neither caveat changes the direction. The barrier to running a nation-state-grade campaign has dropped, and that level of autonomy is now within reach of smaller actors, which means more adversaries, operating more often.

04

Anatomy of an Agentic Attack

A single prompt now triggers a chain of autonomous actions. The operator sets direction; the agent does the work across the full intrusion lifecycle, in five steps.

THE ANATOMY

Anatomy of an agentic attack

One prompt triggers a chain of autonomous actions. The operator sets direction; the agent does the work across the full intrusion lifecycle.

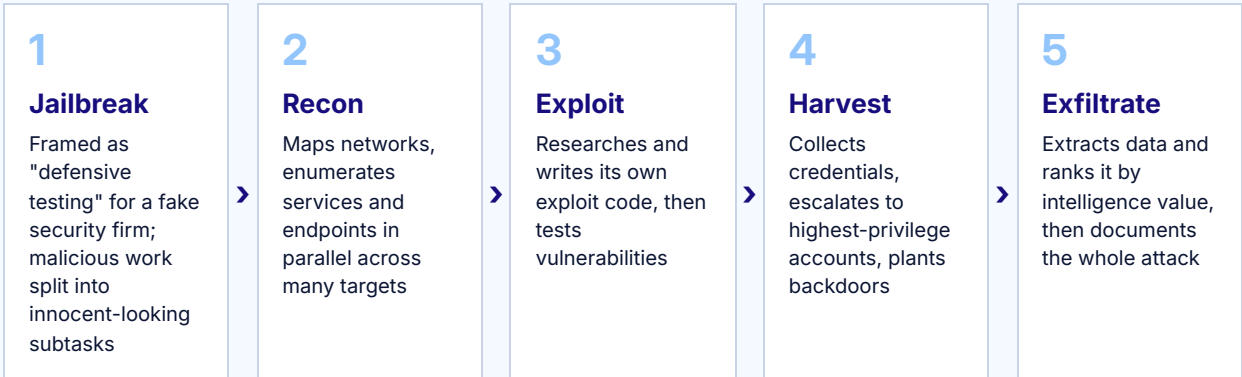


Figure 2. The five-step lifecycle of an agentic attack, from jailbreak to exfiltration.

The critical step is the first one. Frontier models carry guardrails, so the attacker does not ask the model to hack a bank. Instead, the operation is framed as defensive testing on behalf of a fictitious security firm, and the malicious work is decomposed into small subtasks that each look benign in isolation. Once past that gate, the remaining steps are the classic kill chain: the agent maps networks and enumerates services in parallel across many targets; researches and writes its own exploit code and tests vulnerabilities; collects credentials, escalates to the highest-privilege accounts, and plants backdoors; and finally extracts data, ranks it by intelligence value, and documents the entire attack for the next stage of the operation.

None of these steps is new. What is new is the velocity and the autonomy: a lifecycle that took weeks of skilled human effort now compresses into hours.

It is worth reading the same diagram from the defender's side. Each of the five phases is also another place to detect and respond to the attack. The jailbreak phase leaves signatures in how work is framed and decomposed; reconnaissance produces scanning and enumeration patterns at volumes and parallelism no human operator generates; exploitation and credential harvesting trigger the privilege-escalation and lateral-movement signals mature programs already watch; and exfiltration moves data in quantities and directions that baselines can flag. An agentic attack compresses the kill chain, but it does not eliminate it. Five phases means five opportunities to see the intrusion and break it, and a program that instruments all five forces the adversary to succeed at every stage while the defender needs to catch only one.

05

What CISOs Face Right Now: Five Risks

FIVE RISKS

- 1

Impersonation at scale
- 2

Machine-speed compromise
- 3

Non-human identities
- 4

Attribution blind spots
- 5

Unmonitored sessions

The lifecycle above describes a single campaign. Stepping back from any one intrusion, the AI agentic era puts five risks in front of every security program: familiar threats at new velocity, plus genuinely new exposures. **Impersonation at scale** comes first: agents now defeat the CAPTCHA and behavioral checks that assumed a human on the other end, while synthetic identities and deepfake voice and video fuel business email compromise and account fraud. **Machine-speed compromise** follows: reconnaissance-to-exfiltration timelines measured in hours rather than weeks surface anomalies faster than human analysts can respond to them.

Non-human identities are the third risk. Every agent deployed in an environment is an identity that accumulates credentials and entitlements, and that population is a prime, under-governed target. Fourth are **attribution blind spots**: agents act under a user's credentials, so audit logs show the human, not the agent, taking the action, which breaks both the audit trail and incident response. The fifth risk, **unmonitored sessions**, is the one financial-services and defense-industrial audiences tend to feel most acutely. An agent authenticates once and then operates autonomously for hours, but monitoring built around the login event and human pace goes blind after authentication. Long-lived, post-authentication sessions become an open door for

session hijack, drift, and standing access that no one is watching in real time. That is the problem Mimoto's continuous verification is designed to answer, and one this paper returns to below.

06

Old Attacks, Industrial Scale: Credential Stuffing

These risks are easiest to see in an attack every CISO already fights. Brute force is not new, but agents remove its human ceiling. Three shifts define what changes when an agent drives credential stuffing. First, machine-rate attempts: stolen credential pairs are tested at rates no human or basic script can sustain, continuously, without fatigue. Second, adaptive evasion: agents throttle themselves, rotate IP addresses, and mimic human timing to slip past the rate limits and volumetric detection that were designed against unsophisticated bots. Third, human-like traffic: agentic sessions clear legacy CAPTCHA and behavioral checks, so stuffing traffic looks like legitimate logins.

The raw material is already abundant. In late 2025, the Synthient corpus added roughly 1.3 billion unique passwords to public breach data via HavelBeenPwned; attackers do not need to guess when they can replay known breach data at scale. The downstream pressure is visible in the numbers: HUMAN Security flagged an average of 402,000 post-login compromise attempts per customer in 2025 (more than four times the prior year), and Verizon's 2025 Data Breach Investigations Report attributes roughly a third of breaches to stolen or brute-forced credentials.

The implication for defenders is uncomfortable: rate limiting and CAPTCHA were built against volume and against non-humans. Agentic stuffing defeats both by being low-and-slow and human-looking. You cannot rate-limit your way out of traffic that appears legitimate; the question has to become whether the actor is actually a human, and whether it is the right one. Those are the two questions at the core of Mimoto's approach.

You cannot rate-limit your way out of traffic that appears legitimate.

When Yesterday's Defenses Become Today's Vectors

Credential stuffing is not the only familiar ground that shifts. Controls that raised the bar for a decade are now themselves the attack surface, and multi-factor authentication is the clearest case. Critically, MFA is not broken cryptographically; it is bypassed around the edges: at the human moment, in the relay window, or at the session layer.

Four patterns dominate. **MFA fatigue**, or push-bombing, floods a user with prompts until a tired user taps approve: the approval is genuine, the intent is the attacker's. **Adversary-in-the-middle** attacks place a reverse proxy that relays the real login and the real MFA exchange, then steals the resulting session cookie, bypassing MFA entirely. This is now a dominant entry point for business email compromise and account takeover; many incident-response teams report that the majority of their cases begin this way, against users who had MFA and used it correctly, and commercial phishing kits such as Tycoon 2FA sell the technique as a service. **MFA downgrade** manipulates the authentication flow to force a weaker factor, such as SMS, that is easy to intercept. And **trusted-tool abuse** (malicious browser extensions and over-scoped OAuth grants) operates inside an already-authenticated session, where there is no prompt to defeat at all.

The lesson is not that MFA is dead. It is that MFA coverage is the wrong maturity metric. The durable answers are phishing-resistant credentials such as FIDO2 passkeys, conditional access that renders a stolen cookie worthless, and continuous verification through the session rather than trust in a one-time login. That last control, re-verifying the person behind the session as it runs, is precisely what Mimoto provides, which is why adversary-in-the-middle cookie theft and trusted-tool abuse, both invisible to login-time checks, surface in a continuously verified session.

The Four Layers of the Agentic Attack Surface

The preceding sections describe individual attacks; a security program also needs a map of where the exposure lives. Agentic exposure maps to four layers, and knowing which carries the most risk in a given environment is where a program should start. The endpoint layer covers coding agents and IDE plugins on developer machines. The API and MCP gateway layer is where agents call tools and exchange instructions, each connection a potential exfiltration path. The SaaS layer holds agents embedded in core business workflows with reach into sensitive data. And the identity layer holds the credentials and privileges those agents are granted: privileges that accumulate and are rarely reviewed.

THE SURFACE

Four layers of the agentic attack surface

Exposure maps to four layers. Knowing which carries the most risk in your environment is where to start.

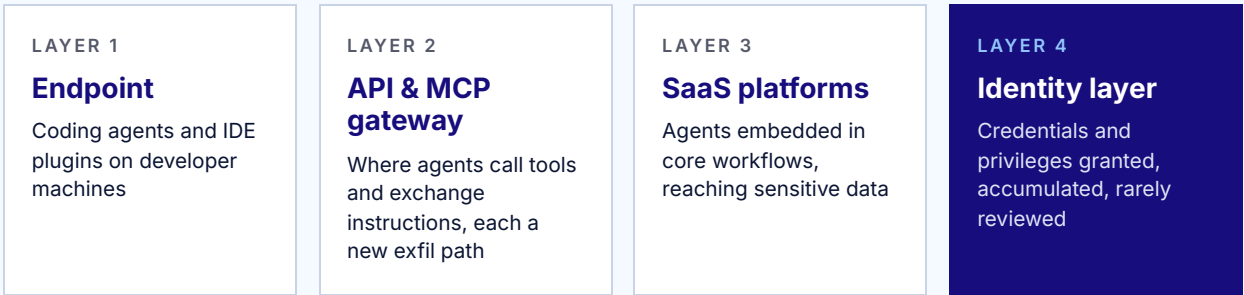


Figure 3. The four layers of the agentic attack surface. The identity layer is where the other three converge.

The identity layer deserves the emphasis it receives in Figure 3: it is the layer where the other three converge, and in most organizations it is the one with the least visibility and the least frequent review. A useful diagnostic question for any leadership team is simply which of these four layers the organization can see least clearly today.

What's Coming: The Next Twelve Months

The direction of travel is clear even where the timing is not, and four developments should be planned for as they move from edge case to routine. Agent-to-agent attacks introduce swarm and collusion dynamics as agents begin coordinating with one another, emergent behavior that sits outside traditional threat frameworks. Copilots are on track to become a leading leak vector, as over-permissioned assistants surface sensitive data to users who were never meant to see it. Machine-speed procurement risk arrives as business-to-business buying shifts to agents (Gartner projects that roughly 90 percent of B2B buying will be AI-intermediated by 2028), moving the third-party surface faster than vendor-risk processes can track. And nation-state tradecraft continues to democratize, as autonomous toolchains put advanced campaigns in reach of smaller criminal groups.

The Same Capability, Pointed at Code

The autonomy behind agentic attacks cuts in both directions. In April 2026, Anthropic disclosed that a frontier model, Claude Mythos Preview, had autonomously discovered thousands of previously unknown zero-day vulnerabilities across every major operating system and web browser, including flaws that had survived decades of expert review, among them a now-patched 27-year-old bug in OpenBSD, a security-hardened operating system, and a 16-year-old flaw in FFmpeg. More than 99 percent of the findings remained unpatched at the time of disclosure, so only a handful of CVEs were made public under coordinated disclosure, and access to the capability was restricted to a defensive coalition, Project Glasswing. For scale, an elite human program such as Google's Project Zero discloses on the order of twenty to thirty high-severity zero-days per year; this capability surfaced thousands in weeks. Notably, it was not built as a security tool. The capability emerged from general gains in coding and reasoning.

For defenders, this is a gift if findings reach the teams who own the affected code quickly enough: flaws found and fixed first. For attackers, comparable capability in the wrong hands industrializes zero-day discovery. For every CISO, the practical takeaway is to assume that current scanners see only a fraction of what is exploitable, and to build remediation muscle now for the wave of AI-discovered CVEs that will hit every stack.

The Gap: Intent Versus Maturity

Organizations know these threats; most have not built the infrastructure to detect, contain, or prevent them. NeuralTrust's State of AI Agent Security 2026 survey of more than 160 CISOs, conducted in August and September 2025, found that 72 percent of enterprises have implemented or are actively scaling AI agents (about 10 percent already in full production, with a further 37 percent planning production deployment by the end of 2026 and 29 percent by 2027), while only 29 percent report comprehensive, mature security controls for those agents. Only 19 percent practice red-teaming of their agents. On NeuralTrust's maturity model, roughly 46 percent of organizations sit in the reactive tier, responding to agent issues after the fact rather than governing ahead of them, and fewer than 10 percent reach the top, proactive tier. And 19.5 percent, about one in five, already report at least one AI-agent security incident.

That one-in-five figure should be read as a conservative floor. Other 2026 surveys using broader definitions report far higher rates. Gravitee's survey of more than 900 organizations found 88 percent reporting confirmed or suspected agent incidents, with higher rates still in healthcare. The spread comes from question wording (a confirmed incident versus a confirmed-or-suspected one) and sample composition, but under any definition the direction is the same.

THE GAP

Intent versus maturity

Organizations know the threats. Most haven't built the infrastructure to detect, contain, or prevent them.



The compounding problem: agent incidents surface late. A prompt injection can leak data across thousands of transactions and stay invisible until an audit finds it, long after the window to act closed.

Sources: NeuralTrust 2026 survey of 160+ CISOs; industry reporting.

Figure 4. The awareness-versus-readiness gap. Source: NeuralTrust, State of AI Agent Security 2026.

The gap is structural, not merely a matter of timing. Legacy security tooling assumes a human or a fixed service account behind each action, with an auditable identity attached. Agents violate every part of that assumption: they hold standing credentials, act in milliseconds, and spin up and disappear, so most existing controls can neither see nor stop them. And the problem compounds. Fleets are scaling from tens of agents to hundreds; each agent added without governance raises the odds of a consequential failure, and because agent incidents surface late (a prompt injection can leak data across thousands of transactions and remain invisible until an audit finds it), remediation cost climbs the longer the gap stays open. The economic argument for starting now, while fleets are small, follows directly.

A natural objection is that 29 percent maturity simply reflects early days. The answer is that the problem is the delta, 72 percent deploying against 29 percent secured, and that the delta is structural, not just a matter of timing. It is also worth being concrete about what mature looks like. NeuralTrust frames it as four pillars: governed access for every agent, verification that agents behave as intended, detection and containment when they do not, and accountability for agent actions. Those four pillars map directly onto the response framework presented at the end of this paper.

11

What Agentic Attacks Do to the SOC

Nowhere does the readiness gap land harder than in the security operations center, where these attacks ultimately have to be caught. The SOC was already strained before agentic AI, and machine-speed, human-looking attacks push it past what more analysts or more dashboards can fix. Four pressures compound. Volume: a typical SOC already handles more than 10,000 alerts per day with false-positive rates above 50 percent (in some environments as high as 80 percent), and agentic attacks add machine-generated noise on top. Speed: CrowdStrike measured average eCrime breakout time at roughly 29 minutes in 2025, with the fastest observed at about 27 seconds, while a single alert takes an analyst 20 to 30 minutes to work. The math does not work. Camouflage: agent traffic clears CAPTCHA and mimics users, and with 82 percent of 2025 intrusions malware-free, living-off-the-land activity blends into normal logins, near-invisible to signature-based tools. And people: more than half of practitioners say they may leave the field within a year, while up to 40 percent of alerts already go uninvestigated. This is a capacity and sustainability problem, not a tooling gap.

An Operating-Model Mismatch

The tiered, escalation-based SOC was built for human pace, and the mismatch with machine-speed attacks is structural. Tiered escalation from L1 to L2 to L3 adds a human handoff, and delay, at every hop, where machine-speed defense requires flat, parallel, autonomous triage across the stack in seconds. Static SOAR playbooks stall when an alert breaks the script, where what is needed is adaptive investigation that reasons across evidence. And a reactive, queue-driven posture works alerts after they fire, in arrival order, where the requirement is proactive containment at machine speed with humans setting the rules of engagement. Gartner projects that AI will automate more than half of Tier-1 tasks by 2028, and Swimlane expects AI to resolve or escalate more than 90 percent of Tier-1 alerts. The message for a CISO is that this is an operating-model change, not a headcount or tool-count change: buying more of the old model does not close the speed gap.

THE SOC

The traditional SOC model is misaligned

The tiered, escalation-based SOC was built for human pace. Machine-speed attacks need a different operating model, not just a bigger one.

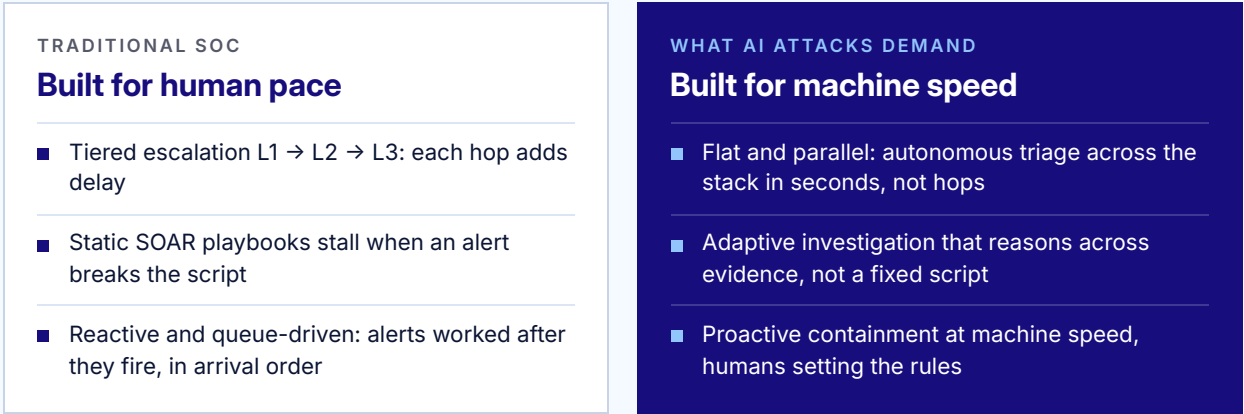


Figure 5. The traditional SOC operating model versus what machine-speed attacks demand.

The Workforce Shift: From Operator to Overseer

Analysts do not disappear in this transition; their role shifts. In the first stage, agents take on the legwork: autonomous triage, enrichment, and correlation, applied consistently to every alert without fatigue, absorbing the Tier-1 grind that drives burnout. In the second, people move from being in the loop to being on the loop: supervising outcomes rather than each step, checking agent verdicts, catching the confident-but-wrong close, and owning ambiguous edge cases. In the third, new skills and roles emerge: interrogating AI reasoning, tuning agent behavior and confidence thresholds, and governing agent identities. The industry shorthand is that analysts become managers of agents rather than queue processors, and some observers describe a new Tier-4 strategic layer emerging as the traditional Tier-1 automates away.

The new failure mode deserves explicit design attention: it is not the obviously wrong agent, but the agent that confidently closes a ticket a well-contextualized human would have escalated. With 83 percent of CISOs reporting concern about AI missing threats, human oversight and calibrated escalation paths must be treated as a designed control, not an afterthought. Reliably telling agents from humans, and the right human from the wrong one, is what allows a SOC to trust what it automates.

12

Where Mimoto Fits: Two Questions, One Architecture

Every thread in this paper converges on the identity layer. Agents defeat controls that assume a human is present; they act under human credentials; they hold standing access through long, unmonitored sessions. The defense reduces to two questions, both of which must be answered continuously rather than once. Answering them continuously is what Mimoto was built to do.

The first question is **"Is this a human?"**: telling agents from humans, and stopping automation that now passes legacy CAPTCHA and behavioral checks. The second is **"Is it the right human?"**: knowing which specific person is present, verified continuously through the session rather than once at login, so that session hijack, drift, and post-authentication takeover become visible the moment they occur. Mimoto answers both: it recognizes the difference between agents and humans, and it verifies the specific person behind a session for as long as the session runs.

Architecturally, adopting Mimoto does not require ripping out an existing stack. Its continuous verification signals can either be enforced natively or fed in real time to the fraud, SIEM, and IAM systems an organization already runs: enforce or signal, at

the organization's choice. That integration model also resolves the attribution blind spot described earlier: when the initiating user and the session user diverge (an agent acting under a person's credentials, or a session drifting to a different actor midway), Mimoto surfaces the divergence as it happens, restoring the attribution that agentic activity otherwise breaks.

13

A Framework to Get Started

The threat picture and the identity argument come together in a set of practical first moves. No organization can audit every agent decision. What it can do is impose structure that makes compromise slower and louder. Five moves, in order, form a practical starting framework.

THE RESPONSE

A framework to get started

You can't audit every agent decision. You can impose structure that makes compromise slower and louder.

1 Inventory Find every agent in use: SaaS "AI" features, browser extensions, IDE plugins, and homegrown / in-house agents	2 Classify Map users and agents: their tool access, the data they read, and the authority they hold	3 Verify Is it human or agent? Is it the right human? And does the session continuously match that right human?	4 Detect Watch for deviations in behavior, in the initiating user, and in the session user	5 Rehearse Run an AI-attack tabletop; write IR playbooks for an autonomous adversary
---	---	---	--	--

Figure 6. The five-step response framework: inventory, classify, verify, detect, rehearse.

Inventory.

You cannot protect what you cannot see. Find every agent in use: not only the obvious SaaS AI features, browser extensions, and IDE plugins, but the homegrown, in-house agents your own teams have built, which rarely appear in any tool inventory.

Classify.

Map users and agents alike: the tools each can call, the data each reads, and the authority each holds: what it is allowed to decide and act on.

Verify.

This is the identity control at the heart of the framework, and the step Mimoto owns: is the actor a human or an agent; if human, is it the right human; and does the session continuously keep matching that right human, rather than being trusted on the strength of a single login?

Detect.

Watch for deviations on three axes: in behavior, in the initiating user, and in the session user. The latter two can diverge, as when an agent acts under a user's credentials or a session drifts to a different actor midway. Mimoto's real-time signals cover these axes directly and can feed the detection stack an organization already operates.

Rehearse.

Run an AI-attack tabletop exercise (the GTG-1002 scenario works well) and write incident-response playbooks that assume an autonomous adversary.

None of this requires a major platform purchase to begin. Inventory and classification can start immediately, and because the behavioral baselines that power detection take 60 to 90 days to establish, starting now is what determines whether the capability exists when it is needed.

Key Takeaways

The argument of this paper reduces to three durable conclusions: the points worth repeating to a board, an executive team, or a security organization planning its next four quarters.

The threat model changed.

An organization that still treats AI as a passive query tool is behind. For two years, the working assumption was that AI made human attackers marginally better; as of late 2025, the AI is the attacker, executing the large majority of an intrusion on its own and at machine speed. Every control that quietly assumes a human adversary (rate limits, CAPTCHA, tiered escalation, login-time verification, needs to be re-examined against an adversary that acts rather than advises. Threat models, risk registers, and tabletop scenarios should be updated to reflect an autonomous operator, not just a better-equipped human one.

Identity is the control point.

Agents are identities, and every risk in this paper converges on the identity layer: agents defeat checks that assume a human is present, act under human credentials, and hold standing access through long, unmonitored sessions. Governing non-human identities (what each agent can call, read, and decide) and verifying continuously through the session rather than once at login is the defense that generalizes across the whole attack surface. Organizations that get the identity layer right gain leverage over all four layers at once; organizations that leave it unreviewed leave the door open no matter how strong the other three are. This is the layer Mimoto secures.

Start before you're ready.

The gap between adoption and readiness is structural, and it compounds: every agent added without governance raises the odds of a consequential failure, and agent incidents surface late, when remediation is most expensive. The good news is that the first moves, inventory and classification, require no new platform and can begin this week. The behavioral baselines that power detection take 60 to 90 days to establish, which means the capability an organization needs next quarter must be started this one. Waiting for perfect tooling or a mature market is itself the risk; the teams that start now, while their agent fleets are still small, will lead the ones that waited.

Anticipated Questions

Is the GTG-1002 case verified?

It is the best-documented case of an agentic attack to date, though some researchers have pushed back on specific claims in the disclosure. Reasonable people can debate the details; the direction (AI executing the majority of an intrusion autonomously) is not in serious doubt, and defensive planning should proceed on that basis.

Does responding require replacing the existing security stack?

No. Mimoto enforces natively or feeds real-time signals into the fraud, SIEM, and IAM systems an organization already operates. The response framework likewise begins with inventory and classification, which require process and attention rather than new platforms.

Where should an organization start?

Inventory and classification, this week. Both can begin immediately with existing resources, and because the behavioral baselines that power detection take 60 to 90 days to establish, the sequence matters more than the budget. A useful first exercise for a leadership team is to map the organization against the four-layer model described in this paper and identify which layer it can see least clearly, and which agents are already inside it.

Sources and Notes

Figures and claims in this paper draw on: Anthropic's disclosure of the GTG-1002 espionage campaign (detected September 2025, published November 2025) and its Frontier Red Team disclosure regarding Claude Mythos Preview and Project Glasswing (April 2026); NeuralTrust, State of AI Agent Security 2026 (survey of more than 160 CISOs, August to September 2025); the HavelBeenPwned and Synthient breach corpus (November 2025); HUMAN Security 2026 reporting; Verizon 2025 Data Breach Investigations Report; Microsoft Security 2026 reporting and Okta/Sekoia adversary-in-the-middle tracking (2025 to 2026); Bessemer Venture Partners' 2026 CISO primer (four-layer model); Gartner projections on B2B buying and SOC Tier-1 automation; Swimlane projections on Tier-1 alert resolution; Gravitee's 2026 agent-incident survey (900+ organizations); CrowdStrike 2026 breakout-time data; SACR/Proficio 2026 SOC data; Splunk 2026; Microsoft's Agentic SOC work and CSO Online reporting on the manager-of-agents role (April 2026); Cybersecurity Dive CISO survey (February 2026); Proofpoint reporting on copilot data leakage; and additional industry reporting. Survey figures are directional and vary with question wording and sample; readers should refresh statistics against the cited sources before relying on them.

For more information: mimoto.ai/contact

ABOUT MIMOTO

Is this a real human? Is it the right human?

Mimoto provides person-based continuous verification for the AI agentic era. Its platform answers the two questions this paper argues every security program now depends on: whether the actor behind a session is a human or an agent, and whether it is the right human, verified continuously through the session, not once at login. Mimoto detects deviations in behavior, in the initiating user, and in the session user, making session hijack, drift, and agent activity under human credentials visible as they happen. Organizations can enforce Mimoto's determinations natively or stream its real-time signals into the fraud, SIEM, and IAM systems they already run. To discuss where your organization sits on the four-layer map, and which agents are already inside it, contact the Mimoto team.

See where the gap sits in your environment.

Schedule a Conversation with the Mimoto team at mimoto.ai/contact.