Everforth
Apex Systems

WHITE PAPER

AI SECURITY IN THE AGE OF AGENTIC AI

Threats, Realities, and the Path to Disciplined Defense

Contents

Executive Summary

The operational meaning of artificial intelligence in cybersecurity has fundamentally changed. Advances such as Anthropic's Project Glasswing and the Claude Mythos Preview are not simply productivity milestones, they are signals that vulnerability discovery, exploitability analysis, secure code review, and software assurance are becoming faster and more scalable for attackers and defenders alike [8][9]. At the same time, senior leaders are being forced to confront practical, near-term concerns: what AI adoption means for insurability, how governance translates into enforceable control, and how to enable innovation without introducing risk they cannot defend to boards, regulators, or insurers.

Security leaders have long approached AI with warranted skepticism, curbing adoption while creating some moderately effective safeguards within their span of control. That approach of caution has served a purpose, but it is no longer sufficient on its own. The adversarial community does not wait for institutional readiness. AI amplifies attacker economics around weaknesses that already exist in most environments: identity sprawl, software supply chain opacity, cloud misconfiguration, insufficient monitoring, and slow validation cycles.

Everforth Apex's point of view is straightforward: organizations must treat AI as both a force multiplier and a first-class attack surface. AI introduces new risks, but the greater danger is how it accelerates the identification and exploitation of existing ones. Defending today's enterprise requires protecting traditional cyber terrain: identity, network, endpoints, cloud, applications, data, and supply chain, while simultaneously protecting the AI itself: models, prompts, retrieval pipelines, agent permissions, orchestration layers, and output-driven actions.

This paper outlines four critical areas of AI Security, our perspective on the market trends, and examples of where Everforth Apex is helping clients make an immediate, measurable impact.

Together, these four areas form a continuous control loop, defining intent, enforcing behavior, validating outcomes, and enabling safe innovation at scale.

Domain	Everforth Apex Answer
Operational Enforcement	Purposeful, custom deployment of agentic AI into security operations
Continuous Validation	Technical defense through autonomous, continuous penetration testing and tactical red teaming
Defensible Governance	Risk management through governance and compliance that meet board, audit, and regulatory expectations
Controlled Enablement	Enterprise AI governance that enables innovation without sacrificing control

The Race of AI, The Machinery, and the Crew Responsible

For several years, AI adoption within the CISO's office moved slowly. Security leaders were skeptical, and rightly so. But skepticism alone does not prevent others from adopting first and asking questions later. Some companies and organizations, guided by security and legal counsel, prohibit AI usage entirely. Others encourage it, even rewarding employees for innovating with whatever tools they could access. Databricks, for example, a leader in the data cloud, announced publicly they would reward employees for AI-driven innovation through internal recognition, cash bonuses, and a supportive culture, including peer-to-peer awards, career development funds, and high-impact project spotlights[1]. Employees could receive cash bonuses of up to $1,000 bonuses for their innovations.

While some organizations put up strict barriers to AI usage and some, like Databricks, seemed to open the flood gates, most enterprises fell somewhere in the middle. Business and technology teams of some companies that sit outside the purview and control of the CISO's office raced against competitors to use AI in whatever ways they could. Their adoption of AI had the best of intentions, to innovate, accelerate, and cut costs. But the dynamic got out of hand quickly, and soon became what resembles something like a Formula 1 race: fast, loud, and full of risk. Users across organizations could deploy AI models with relative ease, while security teams were left to manage the consequences. While some enterprise teams race ahead with AI usage at what feels like the speed of light, security teams have been expected to serve as the pit crew, working to fix or mitigate problems in motion, and prevent crashes or catastrophic failures.

AI adoption did not break existing security controls, it bypassed them. *Innovation moved faster than governance and security could assert enforceable authority.*

To clarify, we don't believe security teams have been standing still on the usage of AI. Many have indeed leveraged AI to fight their own battles, but it has mostly come through the usage of SaaS platforms. The truth is many security teams have simply been the benefactors from the broad movement of enterprise-scale security software vendors embedding large language models (LLMs) early on. The SaaS platforms that CISOs had already invested in were enhanced with features that automated tasks like report writing, reduced manual effort, and turned complex engineering tools into low-code and no-code interfaces. This gave security teams an accessible early win for incorporating AI into their workflows.

While security teams checked the box on AI capabilities through their existing enterprise SaaS platforms, the pace of change around them accelerated dramatically. Enterprise engineering teams have been racing to build and deploy AI into applications. These are often complex models with an ever-changing and opaque downstream effect. For example, software engineers can easily download LLMs from repositories such as Hugging Face and integrate them directly into their code. Unfortunately, they often have a very limited understanding of the vulnerabilities that may exist in the models they are deploying.

They don't know what's under the hood. Hyrum Anderson, in his research, discovered instances where an attacker hid malicious code in the AI models that were injected into corporate software[2]. The AI/ML-enabled malware in these models could run hundreds of thousands of scans in seconds and continuously morph itself to avoid positive deception by antivirus software scans. The malware could evade antivirus software deception and operate free of detection thereafter. Of course, software engineers don't intend to imbed malware into their company's applications, but this highlights the immense risk of the unknown that is threatening organizations today.

Things are beginning to change. We have observed, specifically in the past few months, security leaders have started taking matters into their own hands. The embedded LLMs and "agents" inside SaaS platforms have carried them only so far. The risks their own employees pose are worrisome, but the new capabilities that criminals have – or could soon have – at their fingertips have created a new urgency.

Security leaders are taking action, and they are looking at service providers such as Everforth Apex to help with their mission. The next step requires a more purposeful, mission-aligned approach.

Security teams were given responsibility for AI risk, without authority over how AI was deployed.

Attackers Do Not Need A Brand New Vulnerability Class

Before we begin outlining the four foundational areas of Everforth Apex's AI Security model, it must be explained that well before Glasswing was announced in April 2026, there were already real incidents showing the rapid changes to cybersecurity that AI has created.

In 2024, criminals used a deepfake video conference to impersonate Arup executives, inducing 15 wire transfers totaling HK$200 million [3]. In late 2025 and early 2026, researchers documented an AI-assisted intrusion spanning multiple Mexican government entities, in which a single operator used commercial AI tools to accelerate scripting, exploitation, and data analysis at a scale previously associated with much larger teams [7]. And in February 2026, McKinsey's internal AI platform, Lilli, was autonomously compromised in a responsible-disclosure exercise [4][5].

The McKinsey case deserves particular attention because it was an end-to-end compromise of a live enterprise AI platform, not a laboratory demonstration. The autonomous agent reportedly selected Lilli on its own, mapped the attack surface through public API documentation, identified more than 200 documented endpoints (including 22 requiring no authentication), and recognized that JSON keys were being reflected in database error messages. That allowed the agent to iteratively exploit a SQL injection vulnerability that standard scanning had missed, gain read/write access to the production database in under two hours, and chain that access with insecure direct object reference behavior for further data exposure.

McKinsey patched the unauthenticated endpoints and restricted public documentation shortly after disclosure.

But the lesson was already clear: known application-security weaknesses become materially more dangerous when an autonomous agent can find, reason through, and exploit them at machine speed.

Reported accessible data included 46.5 million chat messages, 728,000 files, 57,000 user accounts, 3.68 million RAG document chunks, and writable system prompts that could have altered platform behavior for thousands of users.

The most dangerous shift in the AI-driven threat landscape is not that attackers need wholly new techniques. **It is that known techniques become cheaper, faster, and easier to apply at scale. Reconnaissance, code analysis, social engineering, exploit chain reasoning, misconfiguration discovery, and workflow mapping can all make an adversary's attacks easier and more effective with AI assistance,** and it can happen at machine speed autonomously.

This makes weak identity practices, fragmented telemetry, incomplete asset inventories, stale patch backlogs, and poor architecture more dangerous than ever before. The organizations most at risk are not only those with significant AI exposure, they are any organization whose existing execution gaps can now be targeted at machine speed.

The techniques are not new, but they are enhanced. What is new is the way to deal with these attacks. The previous assumptions that guided security best practices have new AI-driven realities. Everforth Apex's point of view is that defenders should not wait for a dramatic AI-specific incident before acting.

The AI-Driven Reality, up against our previous assumptions, **reveals that AI amplifies the attacker economics around weaknesses that already exist in most environments.**

The AI-driven reality expands the scope of the CISO's office and demands a comprehensive plan: one that spans automation, offensive validation, operational excellence, governance, and regulatory compliance. Business has never faced greater risk, and the tools to manage it have never been more accessible. Everforth Apex is leading the way to help customers mature their security posture in the face of AI-risk.

This whitepaper highlights a representative subset of current Everforth Apex capabilities. Additional capabilities can be incorporated in future revisions. The broader portfolio already includes managed detection and response, threat intelligence, continuous monitoring, Zero Trust engineering, offensive cyber validation, cyber range-based experimentation, and mission-tailored AI delivery patterns that help customers operationalize AI within the stack they already run.

Domain	Previous Assumption	AI-Driven Reality	The Risk
Software Delivery	New services appear at a manageable and controlled pace.	AI-assisted teams create multi-service systems and integrations rapidly and often with limited supervision or oversight. (Exponentially larger risk surface without visibility)	Faster deployments often bypass security architecture reviews. More microservices create a larger "web" of interdependencies, increasing the likelihood of misconfigurations and unmanaged "shadow" applications and potential exposure points.
Vulnerability Analysis	Human experts are the primary bottleneck.	AI increasingly scales triage, exploit reasoning, and remediation paths at machine speed.	AI eliminates the delay between discovery and exploitation. Attackers use AI to automatically generate functional exploits for newly found bugs effectively ending the grace period organizations used to have for patching.
Identity	Governance is centered primarily on people.	Human, machine, and agent identities now all shape the attack surface. (AI Agents must have unique identities and controls like humans)	Managing "Agent" permissions is incredibly difficult. If an AI agent has the power to act on a user's behalf, a single hijacked session can lead to automated, high-speed data exfiltration or unauthorized system changes.
Security Review	Point-in-time gates are sufficient.	Continuous validation and enforcement are required to keep pace with change.	Without continuous validation, an “approved” state becomes stale within hours. Organizations are left with configuration drift until the next review cycle.
SOC Operations	Manual triage absorbs most alert volume.	AI-enabled attacks and growing telemetry demand AI-aware operations and smarter prioritization. (Outpacing Humans)	AI driven attacks create a flood of telemetry that manual triage cannot process allowing sophisticated threats to hide in the volume.
Governance	Periodic approvals define control.	Living controls must govern what data, tools, and actions AI can access in real time.	Static policies can’t govern dynamic AI behaviors. If controls aren’t living organizations can be insecure but appear compliant on paper.

1. Operational Enforcement

Purposeful, Custom Deployment of Agentic AI into Security Operations

The most intuitive starting point for an AI security strategy may be within operations. As AI-entangled security events grow more prevalent, it is increasingly critical for security operations teams to understand the threats, harden exposure, and accelerate remediation. While human judgment remains essential for identifying the challenges an organization faces, machines can absorb the intimidating volume of low-level operational work which can automate the repetitive, the predictable, and the high-frequency.

Automation of security operations functions such as, but not limited to, SOC analyst monitoring, patch management, vulnerability scanning, and incident triage all contain manual, human-operated components that are strong candidates for agentic AI. Enterprise SaaS platforms recognized this early, embedding LLMs into their products and creating agentic solutions to improve speed and usability. However, as these platforms have matured and invested in AI development, their pricing has risen accordingly. While some vendors have embedded AI functionality without raising license fees; others have launched standalone AI modules requiring separate investment. Security leaders must assess whether the return justifies the cost and whether off-the-shelf solutions actually address their unique operational needs.

There is a better path. When it comes to leveraging AI in security operations, leaders can step back being solely SaaS dependent, evaluate their specific operations, and identify where automation genuinely creates value. Our position is not anti-SaaS, far from it. But the everyday operations teams' desires need to meet them where they are. These needs, such as reducing ticket volume, accelerating time to resolution, or cutting down on false positive rates must be evaluated individually and precisely.

A surgical approach to process analysis reveals the right opportunities for agentic AI deployment, without forcing a one-size-fits-all solution onto a complex, specialized environment.

Alert Fatigue: A Persistent and Solvable Problem

Everforth Apex does not view AI as a replacement for operators, analysts, engineers, or practitioners. AI is best used to improve throughput, prioritization, and consistency. Humans remain essential for escalation, judgment, governance, exception handling, and mission tradeoffs. The goal for all of us is disciplined acceleration. In other words, let machines do what they're best at while humans do where they're best.

For example, alert fatigue remains one of the most persistent pain points in security operations. Despite years of effort, an overwhelming volume of false positives continues to drain analyst capacity and mask genuine threats. Industry data consistently shows that security operations teams spend a disproportionate share of their time investigating alerts that turn out to be benign. That is time that cannot be recovered. It is costly and distracting.

One answer is a purpose-built AI solution tailored to the operational environment. Everforth Apex's Knoesis product directly addresses this challenge. Knoesis is a real-time, AI-powered alert triage and insights capability designed for ITOps and SOC teams. By automatically classifying alerts, prioritizing high-risk events, and surfacing predictive context, Knoesis helps analysts cut through noise and focus on what matters. For environments overwhelmed by the volume of alerts generated across cloud, endpoint, identity, application, and AI workloads, this capability is not a luxury, it is a prerequisite for keeping analysts effective.

Knoesis supports a more scalable operating model in which AI handles first-pass triage and correlation, while human analysts remain focused on investigation quality, response, and adversary understanding. We first built Knoesis for ourselves, but can now customize it to our customer environments to reduce alert fatigue.

Practical Path Forward

The best path to enhanced security operations is rarely a rip-and-replace program. Everforth Apex's preferred approach is additive and mission-aligned: harden the stack the customer already owns, close the most dangerous exposure gaps first, and build operational assurance capable of keeping pace with AI-enabled change. A practical path forward typically follows four phases.

Phase	Key Activities	Primary Outcome
Discover	Inventory models, agents, connectors, repositories, vector stores, service accounts, prompts, data flows, and shadow AI.	A usable map of the AI attack surface and everything it can touch.
Harden	Apply Zero Trust, least privilege, secret control, segmentation, secure delivery, governed retrieval, and action boundaries.	A reduced blast radius and a secure-by-design baseline.
Validate	Assess AI applications and supporting platforms; run CTEM, red and purple teaming, continuous scanning, and rapid remediation.	Evidence that controls hold under adversarial conditions.
Operate	Run continuous monitoring, prioritized triage, threat intelligence, governance reviews, resilience exercises, and managed operations as needed.	Continuous assurance that can keep pace with machine-speed change.

For operations teams, AI becomes both the threat and the solution to that threat. SaaS platforms provide value, but their broad, generalized applications cannot address the unique operational structures of every organization. The path forward requires evaluating each organization's structure and needs individually, then applying precise, cost-effective, agentic AI that solves real problems and genuinely enhances security operations and security posture.

2. Continuous Validation

Technical Defense Through Autonomous, Continuous Penetration Testing and Tactical Red Teaming

The pace of AI advancement has heightened concern among security leaders about the number of vulnerabilities that could soon be discovered and exploited at scale. Anthropic's Project Glasswing and the Claude Mythos Preview threw gasoline on a fire that was already burning hot [10][11]. To get ahead of this, security teams must go on the offensive. Traditional penetration tests have increasingly become a compliance checkbox performed to satisfy cyber insurance requirements or regulatory mandates, then filed away. In the new AI-driven world, where zero-day vulnerabilities can be discovered and exploited by autonomous agents in hours, that approach is no longer adequate. Clear visibility into an organization's AI security and governance risks becomes critical.

Validate Continuously Under Adversarial Conditions

AI increases the need for testing. Organizations need to know whether their controls hold when confronted with prompt manipulation, API abuse, misconfiguration, data leakage, retrieval poisoning, privilege escalation, tool chaining, and other adversarial behaviors that exploit both classical and AI-specific weaknesses. Continuous Threat Exposure Management (CTEM), AI application assessments, red and purple teaming, and rapid remediation loops all have a critical role to play.

Discover the AI Attack Surface

Before an organization can secure AI, it must know where AI already exists. That means inventorying models, assistants, code repositories, vector stores, system prompts, agent workflows, SaaS copilots, third-party APIs, CI/CD pipelines, service accounts, tool connectors, and the sensitive data those components can reach. Shadow AI discovery and attack surface mapping are foundational. Many of the highest-risk pathways are undocumented or poorly governed.

The output of this phase should be an inventory meaningful to operators: what exists, what it can access, how it is authenticated, what data it touches, what actions it can take, and which business or mission processes depend on it.

Harden Architecture and Control the Blast Radius

Once the environment is mapped, the next step is structural exposure reduction. That means least privilege for human and non-human identities, segmented and observable application design, secured APIs, protected secrets, governed retrieval pathways, system prompt control, secure-by-design delivery, and clear separation between experimentation and production.

Zero Trust provides a practical framework for applying continuous verification and data-centric control in environments where assumptions of trust break down quickly. But Zero Trust can sometimes be subjective. Organizations should also define explicitly what must always require human review or escalated approval. Not every agent action carries the same risk. Mature environments make that distinction explicit and enforce it in controls, not just policy.

Extended Red Teaming for the AI Era

Security leaders must look beyond standard penetration testing and deploy extended red teaming tailored to the AI environment. The expansion of AI platforms and APIs introduces new attack vectors that require discovery-led testing across AI endpoints, model endpoints, vector databases, agent frameworks, and orchestration infrastructure.

Everforth Apex provides improved visibility into AI assets and risks that extend well beyond standard API inventories. Our AI-era penetration testing is curated for this environment, covering up to seven AI-specific domains including model security, AI APIs, RAG pipelines, data poisoning, and governance integrity. Most importantly, we simulate agentic attacks, mirroring how autonomous adversaries actually operate, rather than relying on static, point-in-time testing models.

Point-in-time penetration tests may satisfy compliance requirements, but they can create a false sense of security. Everforth Apex recommends, and itself employs, an autonomous, continuous penetration testing model that ensures ongoing, real-time visibility into the most critical risks threatening client systems.

Implementing an AI Security Operating Model

Where red-teaming capabilities are already mature, Everforth Apex focuses on implementing a comprehensive AI Security Operating Model across platforms, teams, and business units.
In a recent and ongoing client engagement, Everforth Apex partnered with a large enterprise to implement field-level data protection across the data landscape and establish an AI Security Operating Model that bridges accelerated AI adoption with enterprise security requirements.

We recognized that the client's centralized AI platform represented a strategic inflection point for embedding security-by-design principles. Working across the client's security, engineering, and governance teams, Everforth Apex developed architectural patterns, threat modeling methodologies, and operational frameworks specifically tailored to AI workloads.

Through joint assessment of AI initiatives against emerging threat vectors and collaborative creation of reusable security artifacts, this engagement enabled the client to scale AI development velocity while maintaining consistent security controls. Everforth Apex's partner-led execution model, encompassing architectural oversight, governance, and delivery, ensures both traditional data protection and emerging AI workloads adhere to enterprise security standards and compliance requirements.

3. Defensible Governance

Risk Management Through Governance and Compliance that Meet Board, Audit, and Regulatory Expectations

According to this poll by CyberOne, most cybersecurity leaders cannot fully account for how AI is interacting with their organization's data. This is alarming but not surprising. In highly regulated sectors such as healthcare, financial services, and the public sector this represents an unacceptable level of institutional risk.

Organizations find themselves at varying stages of AI security maturity. Many lack a clear inventory of LLM assets in their environment, have limited visibility into how those models interact with sensitive data, and have not assessed the downstream impact of that exposure. An asset inventory of LLMs is a necessary first step, but governance cannot stop there. It must be injected into the organization's secure development lifecycle and embedded in DevSecOps processes with sufficient authority and enforcement to function as a real operational control, not a policy document with a badge and a flashlight.

> *"If a regulator walked in tomorrow and asked you to reconstruct the decision trail of every agent that touched regulated data in the last 90 days, could you?"* ***When this question was posed to 30 CISOs, only two said yes.***

Secure by Design Must Extend to AI

Security-by-design principles must encompass AI-enabled workflows, not just conventional software. That means secure architecture, clear trust boundaries, least-privilege access to data and tools, hardened APIs, tested deployment pipelines, and policy-aligned retrieval and action logic. AI systems should inherit the same rigor expected of production systems, not be treated as exceptions because they are new or strategically important. In our findings, it's not that AI is being given special treatment, it's that it operates in shadows.

Fortunately, organizations do not need to build governance frameworks from scratch. Global standards for scalable, sustainable AI adoption already exist, including ISO/IEC 42001 and NIST's AI Risk Management Framework [12]. Everforth Apex's AI Security Services help organizations secure and govern AI as it scales across the enterprise, aligning to these frameworks through a tiered approach that supports every stage of AI maturity.

Our services span AI Security Governance Starter, AI Security Control Builder, AI Security Assurance Program, and vCISO/vCAIO leadership. These offerings enable organizations to assess risk, implement controls, maintain assurance, and embed executive oversight. Collectively, these services address the board, audit, and regulatory expectations that most enterprises are currently underprepared to meet.

Data Loss Prevention: Back on the Agenda

Adjacent to AI governance is a renewed and urgent need to revisit Data Loss Prevention (DLP) strategy. As enterprise AI tools proliferate, sensitive information and intellectual property face greater risk of unauthorized exposure than at any point in recent memory. DLP is back on the CISO's priority list, and for good reason.

Everforth Apex is actively engaged in helping clients mature their DLP posture in parallel with AI adoption. In one recent engagement, Everforth Apex helped a client deploy an enterprise AI chat product to employees while simultaneously aligning governance controls and implementing a DLP product and strategy as the DLP layer, ensuring that AI enablement and data protection advanced together rather than in conflict.

Governance Must Operate at Machine Speed

Governance is not a one-time exercise. The output of an assessment tells you where your organization was yesterday, not where it will be tomorrow. In an AI-driven environment, governance must be a living control surface governing data access, model use, testing boundaries, human review requirements, action authority, and third-party risk in real time.

Organizations that rely on point-in-time approval processes alone will find that operational reality diverges too quickly from documented intent. Governance frameworks must be designed to evolve continuously, with automated enforcement mechanisms that can keep pace with the speed at which AI systems change and interact. As regulatory pressure around AI intensifies globally, from the EU AI Act to sector-specific guidance from financial regulators and healthcare authorities, organizations that invest in living governance frameworks now will be far better positioned to demonstrate compliance, maintain audit readiness, and respond to regulatory inquiries with confidence.

4. Controlled Enablement

AI-Assisted Development Changed the Math of Cyber Exposure

AI coding tools have already transformed software delivery. Teams can now produce in days what previously took weeks: multiple services, APIs, data pipelines, front-end layers, automation scripts, and infrastructure-as-code assembled into one working system. The result is both impressive and concerning. We have more capable software, produced faster, with a larger blast radius if security and operational discipline do not
keep pace.

For security and platform teams, the consequences are cumulative. Each new service introduces runtime dependencies, access paths, secrets, networking relationships, logging requirements, identity decisions, and potential misconfigurations. Add retrieval pipelines, model endpoints, agent tools, and third-party integrations, and a so-called prototype can quickly become a material attack surface with real access to real data and
real actions.

Building applications has accelerated. Secure deployment, validation, monitoring, and governance are now the gating functions. Organizations that do not modernize these layers will find that AI speed creates operational drag as quickly as it creates innovation.

TotalSight™: An End-to-End AI Operating Model

To address this challenge, Everforth Apex Apex developed TotalSight™, an end-to-end operating model for AI that unifies the entire lifecycle from intake through development to governance. Rather than managing a fragmented set of tools and processes, teams operate within a single framework organized around three pillars:

- **Business Prioritization:** Uses Business Catalyst scoring, feasibility analysis, and LiftOff/GotoProd estimates to rank and select high-value use cases before a single line of code is written.
- **Rapid Build:** Combines citizen-developer tooling with assisted pro-code development, AutoScrum, and support for any Git-based platform (GitHub, GitLab, Azure DevOps, and more) to accelerate delivery without sacrificing structure.
- **Automated Governance:** AI Watchtower continuously monitors performance, cost, and compliance within a secure, enterprise-grade architecture keeping policies in force across environments as AI workloads evolve.

AI Watchtower: Continuous Governance in Production

As a core module of TotalSight™, the AI Watchtower module provides continuous oversight of AI performance, cost, and compliance. It tracks accuracy and efficiency, surfaces misclassification patterns, and ensures that governance policies remain in force across environments without requiring manual intervention for every change.

For CIOs and CISOs, AI Watchtower provides the visibility and control needed to govern LLM usage across the enterprise in a unified, systematic way. As AI governance requirements escalate, a centralized management capability becomes not just valuable, but essential.

What once required months of fragmented development now happens in weeks, with risk, cost, and compliance visibility maintained throughout the lifecycle. Software-only solutions address part of the problem but frequently create new ones. Services paired with AI accelerators, rather than standalone SaaS products, represent the gold standard for implementing secure AI usage without generating overhead and new technical debt.

Conclusion

The AI-driven adversarial world is the operating environment cybersecurity leaders face today. Glasswing-class capability signals a permanent shift in how software is built and attacked. AI systems now accelerate code creation, vulnerability discovery, exploit reasoning, and security analysis. As a result, the time between weakness and consequence continues to shrink.

These conditions require a security model that operates at the same velocity. Human-paced control cycles alone cannot keep up when adversaries function at machine speed. Organizations need operating models that combine automation with oversight, allowing security teams to manage scale while retaining accountability for critical business decisions.

Everforth Apex's point of view is that effective defense in this environment requires AI in the loop to reduce time to triage, validate exposure continuously, and enforce governance in production environments. Humans remain responsible for setting intent, thresholds, and approval boundaries, while automated systems execute within those constraints. This approach supports enterprise control, traceability, and insurability by preserving evidence of what occurred, how decisions were made, and how risk
was reduced.

Across the four domains described in this paper, the operating model is consistent.

1. Agentic operations apply AI to absorb volume, prioritize risk, and accelerate triage so human teams can focus on judgment, investigation quality, and response.
2. Offensive validation replaces periodic testing with continuous adversarial assessment, ensuring that assumed security conditions are verified under real operating conditions.
3. Governance and compliance evolve from static documentation to enforceable controls with audit-ready records that support regulatory review, board oversight, and insurance requirements.
4. Enterprise AI control enables innovation at scale while maintaining visibility into what AI systems can access, the actions they can take, and how outcomes are monitored over time.

Across these domains, Everforth Apex applies a practical approach: precise assessment, structural hardening, continuous validation, and operational assurance. This allows organizations to move faster with confidence, supported by controls designed for environments where both innovation and threat activity now operate at machine speed.

References

1. Databricks: Data + AI Strategy: People Focus. January 2024.
2. Cisco: Cisco teams with Hugging Face for AI model anti-malware. August 2025.
3. South China Morning Post. Arup confirmed as victim of HK$200 million deepfake scam. May 17, 2024.
4. CodeWall. How We Hacked McKinsey's AI Platform. March 2026.
5. The Register. AI agent hacked McKinsey chatbot for read-write access. March 9, 2026.
6. Gambit Security. A Single Operator, Two AI Platforms, Nine Government Agencies. April 2026.
7. SecurityWeek. Hackers Weaponize Claude Code in Mexican Government Cyberattack. April 2026.
8. Anthropic. Project Glasswing. April 7, 2026.
9. Anthropic. Claude Mythos Preview. April 7, 2026.
10. UK AI Security Institute. Evaluation of Claude Mythos Preview's cyber capabilities. April 13, 2026.
11. Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign. November 13, 2025.
12. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0).