SURAKSHA CATALYST

INDO-AMERICAN CYBER ALLIANCE

Preparing for Frontier AI in Cybersecurity

A Practical Readiness Framework for Security Teams

Audience: Cybersecurity practitioners, CISOs, and enterprise security leadership

Version: 1.0 | **Date:** May 2026

Contributors

This document represents a collaborative effort shaped by the expertise, experience, and perspectives of leaders across cybersecurity, enterprise security, AI governance, and the security research community. We sincerely thank each contributor for their valuable insights, thought leadership, and commitment to advancing AI readiness and resilience across the global security ecosystem.

- Lt. Gen Rajesh Pant — Former National Security Coordinator (Govt of India)
- Gurdeep Kaur — Former Chief Information Security Officer, PSEG
- Rohit Kohli — Vice President & Deputy Chief Information Security Officer, Genpact
- Harsha Reddy — Chief Information Security Officer, VEG
- Beenu Arora — Co-Founder & Chief Executive Officer, Cyble
- Kamal Shah — Chief Executive Officer, Prophet AI
- Pravin Kothari — Founder & Chief Executive Officer, Pointguard AI
- Sandeep Johri — Chief Executive Officer, Checkmarx
- Sumit Ohri — Chief Information Security Officer, GetInsured
- Nalin Narayanam — Chief Information Officer, AdaptHealth
- Supro Ghose - Chief Information Security Officer, Finzly
- Mayra Ahuja — Grade 9, JP Stevens High School

Table of Contents

Introduction

Security teams are increasingly being asked what steps they should take to prepare for the broader release of highly capable frontier AI systems such as Claude Mythos Preview from Anthropic. This document consolidates that guidance into a structured readiness framework.

Researchers use the term “frontier models” to describe AI systems that advance the state of the art. Development has been concentrated among a small number of organizations — primarily OpenAI, Google, and Anthropic — with NVIDIA, Microsoft, and others contributing through partnerships compute infrastructure, and to the broader ecosystem.

Claude Mythos is an unreleased frontier large language model that has demonstrated exceptional capabilities for complex, multi-step cybersecurity tasks. During internal testing in early 2026, the model demonstrated the ability to autonomously identify large numbers of high-severity zero-day vulnerabilities across major operating systems, web browsers and widely used open-source software libraries. In response, Anthropic launched Project Glasswing in April 2026, a defensive cybersecurity initiative, providing gated access to Mythos Preview to a select group of partners including AWS, Apple, Google, Microsoft, JPMorgan Chase, and Palo Alto Networks to help secure critical infrastructure ahead of any broader release. As of May 2026, the model is not publicly available.

Regardless of the precise timeline or ultimate capabilities of Mythos specifically, the scenario it represents — AI dramatically accelerating vulnerability discovery — is one every security team should be preparing for now.

As of May 2026, these capabilities are already operational. Anthropic and approximately 50 Project Glasswing partners reported identifying more than 10,000 high- or critical-severity vulnerabilities across critical software ecosystems. Open-source scanning alone uncovered 23,019 total potential vulnerabilities, including 6,202 initially estimated as high or critical. Of the 1,752 findings independently triaged by external security firms, 90.6% were confirmed as true positives. Anthropic’s coordinated disclosure dashboard reports that dozens of vulnerabilities have already been patched and assigned CVE or GHSA advisories.

One notable example is CVE-2026-5194 in wolfSSL, a critical certificate verification vulnerability that had not previously been identified through existing security review processes. The vulnerability surge addressed in this document is not hypothetical — it is measurable and accelerating.

Community Survey Insight

When asked “How prepared does your organization currently feel for AI-accelerated vulnerability discovery?”, 58% of CISOs and security practitioners surveyed described their organizations as “Moderately Prepared,” while no respondents considered themselves “Fully Prepared.”

This highlights a growing industry awareness of AI-driven security risks, while also underscoring the significant maturity gap that still exists in operational readiness, automation, and AI-native defense capabilities

The Readiness Framework: Reduce → Automate → Accelerate

This document organizes recommended actions around three sequential imperatives, each mapped to a stage below:

- Reduce — Systematically shrink the attack surface and improve the speed of vulnerability resolution before AI-assisted adversaries can exploit it.
- Automate — Embed proactive, AI-assisted testing throughout the development lifecycle so defenders identify weaknesses before adversaries do.
- Accelerate — Evolve security architectures to operate at machine speed, with adaptive controls and autonomous response capabilities capable of matching AI-driven offensive systems.

Framework Readiness Assessment

The majority of surveyed CISOs and practitioners reported that **all stages of the framework represent equally important gaps** within their organizations today, reinforcing the need for a holistic and balanced AI security transformation strategy.

Stage 1: Reduce — Improve Vulnerability Handling

The objective is not to deploy more tools but to engineer an operational model capable of absorbing significantly higher vulnerability volumes without collapsing. Organizations should critically assess whether current processes can scale under crisis conditions — mass zero-day campaigns, supply chain compromises, or AI-driven attack surges.

1.1 Continuously Reduce the Attack Surface

Minimize exposure through disciplined asset rationalization, network segmentation, system hardening, and elimination of unnecessary services. Every component removed or isolated is an opportunity that an adversary no longer has and reduces the ability of AI in chaining vulnerabilities to create exploits.

1.2 Reset Vulnerability Tolerance Thresholds

Move beyond traditional risk-scoring models. Establish days of exposure — the elapsed time from identification to confirmed remediation — as the primary resilience metric. When AI systems can weaponize vulnerabilities at scale, the window between discovery and exploitation narrows dramatically.

Leading programs are converging on single-digit-day remediation for critical externally facing vulnerabilities, with many targeting under 72 hours from patch availability to production deployment. Organizations should benchmark their mean-time-to-remediate against these targets and define escalation thresholds for when remediation lags—especially for internet-facing assets

1.3 Modernize Patch Management for Scale and Speed

Redesign patch management to support emergency deployment scenarios, autonomous prioritization, and high-volume remediation during active threat campaigns. Legacy quarterly-cycle workflows are inadequate for the environment ahead.

This shift is already underway. Oracle has moved from quarterly to monthly patch cycles, and Microsoft expects monthly patch volumes to rise—both driven by AI-enabled vulnerability discovery. Palo Alto Networks also reports that most of its May 2026 advisories across more than 130 products stemmed from frontier AI scanning. In short, patch volumes are rising as a structural trend, not a temporary spike.

> ***Industry Perspective from Checkmarx***
>
> *AI writes code at machine speed — and Checkmarx One secures it at the same speed. Its agentic platform embeds SAST, DAST, SCA, and AI supply chain security directly into the CI/CD pipeline, with Triage Assist autonomously prioritizing findings by real-world exploitability, not static CVSS scores.*

1.4 Prioritize Runtime and Business Context Over Static Severity

Evaluate vulnerabilities based on real-world exploitability, reachability within the environment, exposure pathways, and operational criticality — not CVSS scores alone. A critical-severity finding with no viable exploit path is categorically less urgent than a medium-severity vulnerability on a publicly exposed, business-critical system.

1.5 Sustain Layered Defensive Controls

Patching cannot be the sole mitigation strategy. Maintain continuous investment in segmentation, detection engineering, and compensating controls to create resilience when patching is delayed, infeasible, or outpaced by new findings.

Vulnerability Remediation Readiness

100% of respondents reported patching critical vulnerabilities **within one week**, highlighting the growing emphasis on rapid remediation and operational responsiveness in modern security programs.

Stage 2: Automate — Proactive Security Testing

Frontier AI will likely accelerate vulnerability discovery for defenders and adversaries simultaneously. Organizations that embed proactive, AI-assisted testing throughout their development and operational lifecycle will find weaknesses first. Stage 2 is not about testing more — it is about moving validation left and making it continuous.

2.1 Assume Vulnerability Discovery Will Accelerate

Operate under the assumption that AI systems will substantially reduce the cost and complexity of identifying exploitable weaknesses across large software estates. Planning and resourcing decisions should reflect this assumption before it is confirmed by an incident.

Anthropic has launched the Cyber Verification Program (CVP), which provides verified security professionals with adjusted safeguards on Claude's public models for penetration testing, vulnerability research, and red teaming. CVP does not provide access to Mythos. Security teams should evaluate whether enrollment could support their internal security testing and research efforts.

XBOW, an autonomous offensive security research company that evaluated Mythos Preview, described the model as "a major advance" and "substantially better than prior models" at identifying vulnerability candidates during its testing.

2.2 Test Critical Systems Continuously

Shift security validation from event-driven exercises to persistent, ongoing activity. High-risk applications and infrastructure should undergo automated adversarial testing continuously — not only in response to scheduled assessments or detected incidents.

2.3 Embed AI-Enabled Testing Throughout the SDLC

AI-assisted code review, threat modeling, fuzz testing, and exploit simulation should be integrated directly into developer workflows and CI/CD pipelines. Security testing must be a native part of software delivery, not a downstream gate.

2.4 Validate Runtime Exploitability, Not Only Static Findings

Prioritize remediation based on real-world reachability, active execution paths, data exposure potential, and business impact. A finding that cannot be reached from an attacker's realistic position in the environment is categorically different from one that can.

2.5 Adopt Secure-by-Design Engineering Principles

Align with secure-by-design and secure-by-default practices — including principles published by CISA and equivalent national cyber agencies — to reduce the systemic introduction of weaknesses during development.

2.6 Evolve Testing, Detection, and Response to Machine Speed

Detection engineering, triage, remediation orchestration, and compensating control deployment must increasingly operate with significant automation. Human oversight remains essential, but manual processes alone cannot serve as the primary response mechanism.

2.7 Continuously Validate Software Supply Chain Integrity

Extend validation beyond internally developed code to include third-party libraries, open-source dependencies, CI/CD pipeline components, container images, and build environments. Supply chain compromise remains one of the highest-leverage attack vectors available to sophisticated adversaries.

The scale of AI-driven software supply chain findings makes this increasingly urgent. Through Project Glasswing, Anthropic has scanned more than 1,000 open-source projects and identified 23,019 potential vulnerabilities, including 6,202 initially estimated as high or critical severity. Independent triage reported a 90.6% true-positive rate.

Two examples illustrate the breadth of impact. One is CVE-2026-5194 (CVSS 9.1), a certificate forgery vulnerability in WolfSSL — a cryptography library embedded across firmware, IoT devices, and enterprise software — where a single flaw can cascade across hundreds of downstream products and dependencies. Another is Mozilla's use of Mythos Preview to identify and remediate 271 vulnerabilities in Firefox 150, reportedly a 10x increase over findings generated by a prior Claude model during testing of Firefox 148.

Security teams should stress-test their software supply chain response capabilities against realistic scenarios in which dozens of high-severity vulnerabilities simultaneously impact widely used open-source dependencies within a single patch cycle.

2.8 Assessing Readiness for Stage 3

Stage 3 is a qualitative shift—from embedding security into workflows to redesigning architecture for machine-speed operations. Readiness indicators include:

- Automated security testing in CI/CD for all critical applications, not select teams
- MTTR for critical vulnerabilities consistently under one week with automated patch deployment
- Current, validated SBOM covering direct and transitive dependencies
- At least one AI-driven surge tabletop completed with lessons documented
- Executive sponsorship and budget for 12–18 month security architecture investments

Organizations that do not yet meet these criteria should close Stage 1 and Stage 2 gaps first. Stage 3 capabilities are force multipliers—but only when the fundamentals are solid.

CI/CD Security Automation Maturity

When asked “Does your organization currently have automated security testing embedded in CI/CD pipelines?”, approximately 50% of respondents indicated their pipelines are fully automated, while the remaining 50% reported partial automation.

This indicates that while automation is becoming a standard practice in modern software delivery, most organizations are still operating in hybrid models where manual validation and human-in-the-loop processes continue to play a significant role.

Stage 3: Accelerate — Addressing Long-Term Frontier AI Risk

The long-term impact of frontier AI will extend well beyond traditional software vulnerabilities. The attack surface will increasingly encompass business processes, AI agents, APIs, identity systems, and organizational decision-making itself. Security architectures built around static controls, periodic assessments, and human-centered response models will become progressively less effective against adversaries operating at machine scale and speed.

3.1 AI-Enabled Defensive Automation

Progressively automate detection, triage, investigation, and remediation workflows to reduce dependence on human response speed. Defensive systems must correlate telemetry, prioritize risk in context, and initiate containment actions in near real time. Practical starting points include automated alert triage, AI-assisted threat hunting, and orchestrated containment playbooks for low-ambiguity, high-urgency scenarios.

This is already producing operational results. Anthropic reported that a Glasswing partner financial institution used Mythos Preview to help detect and prevent a fraudulent $1.5 million wire transfer involving a compromised email account and spoofed phone calls.

The shift from purely human-centered security operations toward AI-augmented defense is already underway for early adopters. Organizations participating in initiatives such as Project Glasswing are beginning to operationalize AI-assisted detection, investigation, and response capabilities at a scale that traditional workflows alone may struggle to match.

AI Threat Detection Maturity

The majority of respondents indicated their AI-specific threat detection capability is “**Developing**” reflecting an early but progressing stage of maturity across the industry.

> ***Industry Perspective from Prophet AI***
>
> *Prophet AI is the agentic AI SOC Platform that ensures alerts don't die in a backlog, applying autonomous reasoning to surface only what genuinely requires human attention. It is the operational layer that makes the transition from human-centered to AI-powered investigation, response and threat hunting practical today, not theoretical tomorrow.*

3.2 Adaptive Policy Enforcement

Evolve security policies from static rule sets into adaptive controls that continuously adjust based on context, behavior, threat intelligence, and operational risk. Future attacks will exploit gaps between identity systems, cloud infrastructure, APIs, AI agents, and data platforms. Context-aware, dynamically enforced controls will be essential to maintaining resilience across interconnected ecosystems.

3.3 Continuous Verification

Move toward continuous validation of users, devices, workloads, APIs, agents, and transactions rather than relying on one-time authentication. This principle extends beyond human users to AI agents, service accounts, and automated pipelines. Autonomous offensive platforms will increasingly leverage stolen credentials, session hijacking, and synthetic identities.

> ***Industry Perspective from Pointguard AI***
>
> *PointGuard AI discovers every AI model and agent in your environment, continuously assesses their security posture, and enforces runtime guardrails at machine speed. Its Agentic Gateway provides the centralized control point for MCP, API, and agent-to-agent traffic that Stage 3 of this framework requires.*

3.4 Runtime Monitoring and Behavioral Visibility

Security architectures must provide deep runtime visibility into applications, AI agents, APIs, infrastructure, and user behavior to detect malicious activity as it unfolds. Future frontier AI systems will likely adapt their offensive behavior dynamically during attacks, making static scanning and periodic assessment insufficient for detecting sophisticated, multi-stage campaigns.

> ***Industry Perspective from Cyble***
>
> *As frontier AI compresses the window between disclosure and exploitation, defenders need external intelligence to know which exposures are actively targeted. Cyble's Blaze AI*

> *continuously monitors the surface, deep, and dark web, translating threat signals into prioritized, actionable intelligence; ntegrating natively with SIEM and SOAR so intelligence is operationalized at the moment of detection.*

3.5 Resilient Infrastructure Design

Architect enterprise systems with compartmentalization, redundancy, graceful degradation, and rapid recovery capabilities to limit the blast radius of autonomous attacks. This includes not only technical redundancy but operational continuity planning — ensuring critical functions can continue in degraded modes while security events are being contained.

Recommended Exercise: AI-Driven Vulnerability Surge Tabletop

One of the most immediately actionable steps any organization can take is a structured tabletop exercise built around the scenarios described in this document. Unlike traditional tabletops that model known attacker TTPs, AI-driven surge exercises test organizational readiness for qualitatively different conditions: simultaneous exploitation of multiple novel vulnerabilities, adversary operations at machine speed, and the failure of manual response processes under volume stress. Exercises should involve security operations and AppSec teams alongside CIO and CISO leadership, infrastructure engineering, software development, legal, communications, and executive leadership.

Exercise Design at a Glance	
Scenario themes	Mass zero-day exploitation campaign across widely deployed OSes and browsers, with variants emerging faster than patch cycles can address. Supply chain compromise of a widely used development dependency, affecting hundreds of internally developed applications simultaneously. AI-assisted credential harvesting and lateral movement exploiting gaps between identity systems and cloud infrastructure faster than detection rules can be written. Coordinated mass disclosure of AI-discovered vulnerabilities across dozens of open-source dependencies simultaneously, requiring parallel triage and patching across hundreds of internal applications within a compressed window.
Participants	Security operations and AppSec • CIO / CISO leadership • Infrastructure and cloud engineering • Software development leadership • Legal counsel and privacy • Corporate communications • Executive leadership

Key questions to explore	At what volume of simultaneous vulnerabilities does our patching process break down, and what is our contingency? How quickly can we move from detection to coordinated containment without manual approvals at every step? What decisions require executive involvement, and are the right escalation paths in place today? How do our legal and communications obligations interact with the speed of response a mass exploitation event demands? Where do our security operations and business continuity plans intersect, and are the owners coordinated?

AI Tabletop Exercise Readiness

Most respondents confirmed they have conducted an **AI-focused cybersecurity tabletop exercise in the past 12 months**, while the remaining participants indicated they are **actively planning to do so**.

Key takeaway: AI-driven scenario planning is increasingly becoming a standard part of modern security preparedness programs.

Final Recommendations

Security teams should begin preparing immediately for a sustained increase in vulnerability discovery volume, accelerated offensive automation, and more capable AI-assisted adversaries. Waiting for the public release of any frontier model before initiating readiness activities is not a defensible posture.

Immediately:

- Review vulnerability handling processes against the Stage 1 criteria and identify gaps.
- Assess patch management capacity against surge-scenario volumes and define contingency thresholds.
- Identify highest-risk applications and begin shifting toward continuous security validation for those assets.

Within the next quarter:

- Conduct an AI-driven vulnerability surge tabletop with the cross-functional participant group described above.
- Establish a roadmap for embedding AI-assisted security testing within CI/CD pipelines for critical development teams.
- Review software supply chain visibility and identify dependencies lacking current integrity validation.

Strategically:

- Begin planning the transition toward AI-native security operations, including evaluation of automation platforms for detection, triage, investigation, response, and threat hunting.
- Incorporate the Stage 3 architectural priorities into security strategy and roadmap planning cycles.
- Build the business case for adaptive policy enforcement and continuous verification infrastructure, recognizing that these capabilities require lead time to architect and deploy.

Organizations that adapt earliest to this transition are likely to be in the strongest position as frontier AI capabilities continue to evolve. The gap between those who prepare proactively and those who respond reactively will widen significantly as AI-assisted attacks and defense capabilities mature.

> ***Executive Perspective***
>
> *In the Mythos era, CIOs and CISOs must assume AI-accelerated attackers can discover and weaponize vulnerabilities in hours, not weeks, and respond by adopting zero trust, strong identity security, and continuous, AI-augmented defense modeled on leading models like Zero Trust which includes AI Agents. They should build always-on vulnerability operations, tightly govern both human and AI agents, and "turn AI inward" to continuously test, harden, and monitor their environments so real-time, exploitable risk drives priorities rather than static vulnerability lists.*