

Project Glasswing

Frontier AI: A New Era of Cyber Resilience

PREPARED BY
Visa Cybersecurity

SEP 2026

Frontier AI: A New Era of Cyber Resilience

India's digital payments ecosystem has become one of the largest and most dynamic in the world, supported by Digital Public Infrastructure, progressive regulation, private sector innovation, and widespread consumer adoption. India's total **Digital Payment transactions¹ surged from 6,365 crores in CY 2021 to 26,762 crores in CY 2025², and total digital payments value grew from INR 1,675 lakh crores in CY 2021 to INR 3,144 lakh crores in CY 2025.**

This rapidly expanding digital landscape is driven by ubiquitous smartphone penetration, transformative payments infrastructure, and growing public trust in secure & seamless cashless transactions.

As digital payments converge with advanced AI models such as '**Frontier AI**', that can analyze, reason, and take autonomous decisions across complex tasks, the ecosystem is becoming faster and smarter. While advanced AI models are making payments more intelligent, secure, inclusive, and context-aware, attackers are using the same technology to create deepfakes, sophisticated phishing campaigns, and automated attacks at scale. Recent incidents across the country underscore that **cybersecurity is no longer just a technology issue, but an ecosystem imperative.**

India has seen a significant surge in cyber incidents in recent years³

₹22,495cr	Lost to cyber fraud in 2025 — 2.81 million cases, up 24% year over year.
76%	Of all losses were from investment scams; digital-arrest scams accounted for 9% .
~1/sec	1930 helpline logged 32.4M calls in 2025 — a victim roughly every second.

Recent cyber incidents such as ransomware attacks, data breaches, deep-fake-enabled fraud and identity attacks, and supply chain attacks on Financial Institutions and Third Party Service Providers demonstrate that cyber risk in financial services is no longer confined to individual institutions but can be propagated through multiple connected parties.

As threats become more automated and increasingly AI-enabled, traditional approaches to cybersecurity are proving inadequate. In a market like India, where digital adoption spans diverse consumer segments, maintaining trust requires a combination of advanced AI-driven security, secure design, consumer awareness, and ecosystem-wide collaboration.

Building Collective Cyber Resilience: Key Imperatives

Recognizing that cyber incidents can affect not only individual institutions but also public trust and systemic stability, the **Reserve Bank of India has progressively strengthened its cyber governance and resilience framework.**

While these frameworks provide the foundation, cyber resilience cannot be achieved through regulation alone. Protecting India's digital payments will require collective action across the ecosystem. Four imperatives which stand out:

1. Ecosystem intelligence and Secured Design

No single participant has complete visibility since cyber threats move across institutions, sectors, and borders. Hence, the ecosystem needs:

- Faster, structured mechanisms to share threat intelligence, fraud typologies, and emerging attack methods in near real time, using privacy-preserving protocols across key ecosystem players.
- Security embedded from the product design stage through threat modeling, penetration testing, and privacy impact assessments before launch.

2. AI-driven defense with responsible governance

As fraud evolves towards social engineering and AI-enabled attacks, institutions need to strengthen both technology and oversight:

- Deploy AI to detect anomalous transactions, account takeover, synthetic identities, and mule networks, while ensuring that models remain explainable, and auditable.
- Implement adaptive authentication that considers device, behavior, location, and transaction context - moving beyond static credentials.
- Prepare defenses against AI-enabled threats including deepfakes, voice cloning, and phishing automation.

3. Resilient operations and third-party risk management

Institutions need to enhance both their ability to prevent cyber threats and detect-respond-react from them quickly:

- Maintain 24x7 security and fraud operations, regularly tested through cyber drills, ransomware simulations, and crisis exercises with clear escalation protocols.
- Strengthen due diligence, audit rights, and continuous monitoring across technology providers, cloud platforms, and aggregators, requiring critical vendors to demonstrate incident response capability and business continuity readiness.

4. Human vigilance and governance accountability

Institutions need to establish strong governance around AI including executive ownership and stakeholder awareness:

- Coordinate sustained, multilingual campaigns on phishing, OTP fraud and QR-code scams, across banks, payment networks, fintechs, and government, tailored for first-time digital users, senior citizens, small merchants, and gig workers.
- Treat cyber resilience as an enterprise risk. Boards and senior management should regularly review cyber risk posture, fraud losses, and incident readiness, with accountability clearly defined and investment aligned to the pace of digital growth.

Visa Vulnerability Agentic Harness (VVAH) | Reflections from Project Glasswing

It is in this context that **Visa participated in Anthropic's Project Glasswing to test frontier AI models for defensive cybersecurity**. The work evaluated how advanced AI systems could identify, reason about, and help remediate vulnerabilities across high-criticality applications, internet-facing services, and foundational platforms. One of Visa's central observations was that the key constraint in cybersecurity goes beyond simply finding vulnerabilities, towards validating, prioritizing, and fixing them quickly enough. This insight shifted Visa's emphasis towards Mean Time to Adapt (MTTA), the time from AI-discovered exploitability to a validated production fix.

To operationalize these lessons, **Visa built and open-sourced VVAH**, a multi-model, agentic security pipeline that moves from attack-surface mapping and threat modeling through deep analysis and adversarial verification to structured reporting, remediation, and fix validation.

This paper draws on Visa's experience deploying frontier AI in real-world, high-stakes environments and shares practical approaches for managing cybersecurity threats in an increasingly AI-enabled world. It is designed to support banks, payment aggregators, fintechs, merchants, policymakers, and regulators as we work towards strengthening trust and resilience across the digital payments ecosystem.

India's digital payments journey is a testament to what we can achieve when innovation, policy, and industry come together at scale. The continued success of India's digital payments ecosystem will hinge on our ability to maintain trust, security, and resilience.



Suresh Sethi

Group Country Manager, Visa India & South Asia
Visa

Sources

1. Definition of 'Total Digital Payments' as per RBI's Payment System Report dated May 18, 2026 includes retail digital payments such as NEFT, IMPS, NACH (credit, debit and APBS), card payment transactions (excl. cash withdrawal), PPI payment transactions (excl. cash withdrawal), UPI (including BHIM & USSD), BHIM Aadhaar Pay, AePS fund transfer and NETC (linked to bank accounts) and RTGS.
2. RBI's Payment System Report dated May 18, 2026.
3. News articles in 'ThePrint' dated 21st February 2026 and 'The New Indian Express' dated 12th April 2026, based on data sourced from Ministry of Home Affairs (MHA)

Table of Contents

1	1. Executive Summary
3	2. Mythos Impact on Cybersecurity
5	3. What Visa Did: Visa Vulnerability Agentic Harness
6	3.1 Visa's Participation in Project Glasswing
6	3.2 Vulnerability Discovery and Remediation Orchestration
7	3.3 Visa Vulnerability Agentic Harness
9	3.4 Findings and Architectural Validation
10	4. What We Learned: Lessons from Visa's Deployment
11	4.1 Validation and Structured Artifacts Are the Bottleneck to Fixing Risk
11	4.2 Guided and Calibrated Reasoning Beats Raw Scanning
11	4.3 Agentic Programs Operate at Portfolio Scale; Copilots Do Not
14	4.4 Chaining & Exploitability Is Where Frontier Models Add Most Value
12	4.5 Keep Decisions Deterministic and Audit Traceable
13	5. How Visa Is Responding
14	5.1 Three Strategic Priorities
16	5.2 Agentic Control Plane
16	5.3 Mean Time to Adapt (MTTA)
15	6. Securing the Ecosystem – Implications for Partners & Clients
16	6.1 Third Party and Open-Source Risk
16	6.2 Fraud, Business Logic, and the Seams Between Systems
16	6.3 Regulatory and Controls Alignment
17	7. Conclusion
18	Appendix: Non-Negotiable Architectural Security Practices



1. **Executive** Summary

Autonomous, reasoning-capable AI models such as Anthropic's Claude Mythos compress the time between vulnerability discovery, exploit development, and operationalization from weeks or months to minutes. In that world, traditional "scan on a schedule, patch on an SLA" programs no longer track real risk: the constraint shifts from finding issues to validating, prioritizing, and fixing what matters fast enough to stay ahead of machine speed attackers.

Visa operates a global payments network that processes billions of transactions a day and has been hardened over many years with zero trust architecture, layered defenses, and highly automated security operations built for six nines availability. Visa joined Anthropic's Project Glasswing to test those defenses with Mythos at AI speed, gain practical experience running frontier models on high-criticality systems, and design an agentic security architecture that remains audit defensible under strict human governance.

In initial deployments, Visa used Mythos to assess high-criticality applications, internet-facing services, and foundational platforms. Mythos demonstrated system-wide, context-aware analysis, uncovering vulnerabilities deep in the stack and chaining "small" weaknesses into viable attack paths that would traditionally surface only late in penetration

testing. At the same time, Visa's zero trust controls, segmentation, and existing safeguards meant that even issues flagged as critical did not translate into material exploitability in production: the kill chain would have been broken before an external actor could act.

The conclusion is clear: Mythos and similar models are a structural inflection point. To stay ahead, Visa has built an agentic control plane and the [Visa Vulnerability Agentic Harness, an open-source reference implementation](#) for AI-powered vulnerability management. Security teams can inspect the code, adapt it to their environments, and contribute improvements. Visa has also adopted Mean Time to Adapt (MTTA) (the time from AI-discovered weakness to a validated fix in production) as its primary effectiveness metric.

Project Glasswing at a Glance

The Shift



- **Old model:** Periodic testing, manual review, extended remediation cycles
- **New reality:** AI identifies vulnerabilities in minutes; exploitation can follow within hours
- **Key takeaway:** Fewer than 1% of CVEs are actively exploited, so prioritization matters more than volume¹

¹Source: Known Exploited Vulnerabilities Catalog | CISA

Visa's Response



- **Deployed** Mythos across high-criticality infrastructure with no material issues identified
- **Validated** zero-trust and defense-in-depth architecture against AI-driven scrutiny
- **Implemented Visa Vulnerability Agentic Harness** for agentic testing, AI-led SAST validation, and intelligent policy-based remediation and validation
- **New metric: Mean Time to Adapt (MTTA)** – inventory freshness, exploitable paths per release, validation cycle time

What You Should Do



- **Establish a vulnerability exploitability management program** using AI to scan critical and internet-facing assets
- **Stand up an agentic control plane** with harness, skills, and human oversight for machine-speed remediation
- **Track MTTA as a core metric** – from exploitability finding to validated fix
- **Enforce higher vendor standards** – require continuous validation and an AI-aware security posture



2. Mythos Impact on Cybersecurity

For decades, enterprise security programs assumed that discovering subtle, chained vulnerabilities required specialist expertise and substantial manual effort. That assumption shaped how organizations tested their systems, from scheduled SAST and DAST scans to penetration tests and bug bounty cycles, and the metrics they used, such as mean time to detect, patch SLAs, and raw CVE closure counts.

Frontier models such as Anthropic's Claude Mythos overturn that assumption. In preview evaluations, Mythos has shown it can perform deep semantic analysis across large, heterogeneous codebases, identify subtle weaknesses, and reason about how they combine into multi-step exploit paths, with independent testing confirming significantly higher success rates on expert level cybersecurity tasks than prior models. These capabilities are dual use: defenders can use them to broaden coverage and validate exploitability, but adversaries can also automate reconnaissance, exploit construction, and fraud operations, collapsing the time between code change and abuse.

Traditional SAST tools remain valuable as a first-pass control for known vulnerability patterns; what changes is that syntactic analysis alone is insufficient against AI-speed adversaries who reason about logic, data flow, and exploit chains that signature-based tools cannot see. The result is a tempo shift: discovery becomes cheap and fast, and the scarce resource is an organization's ability to adapt — measured not in MTTD, but in Mean Time to Adapt (MTTA). Programs that continue to rely on periodic testing, static CVSS based prioritization, and manual remediation will see headline metrics improve while real exposure quietly grows.

Mythos Differentiator

AI-driven vulnerability discovery is accelerating – security teams need machine-speed scale

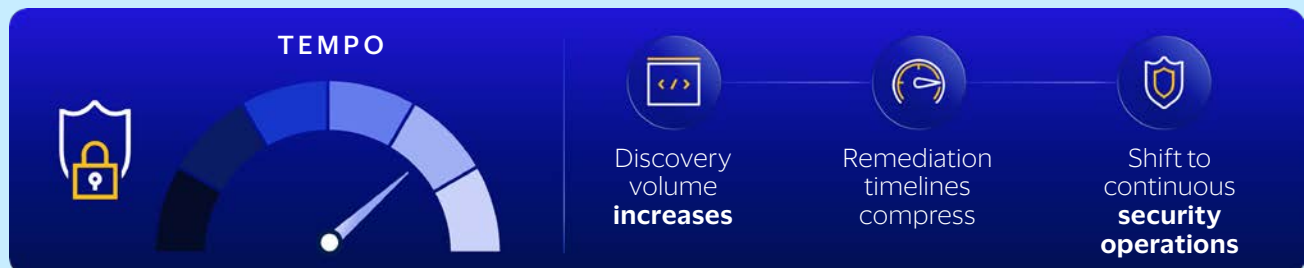
Claude Mythos can autonomously identify, analyze, and chain vulnerabilities



This creates both defensive advantage and systemic risk



Anthropic's Project Glasswing exists to get ahead of the inevitable shift





3. What Visa did:

Visa Vulnerability **Agentic Harness**

3.1 Visa's Participation in Project Glasswing

Visa joined Project Glasswing to evaluate frontier AI capabilities in a controlled defensive program, with tightly scoped use cases and strict human oversight.

The objectives were to:

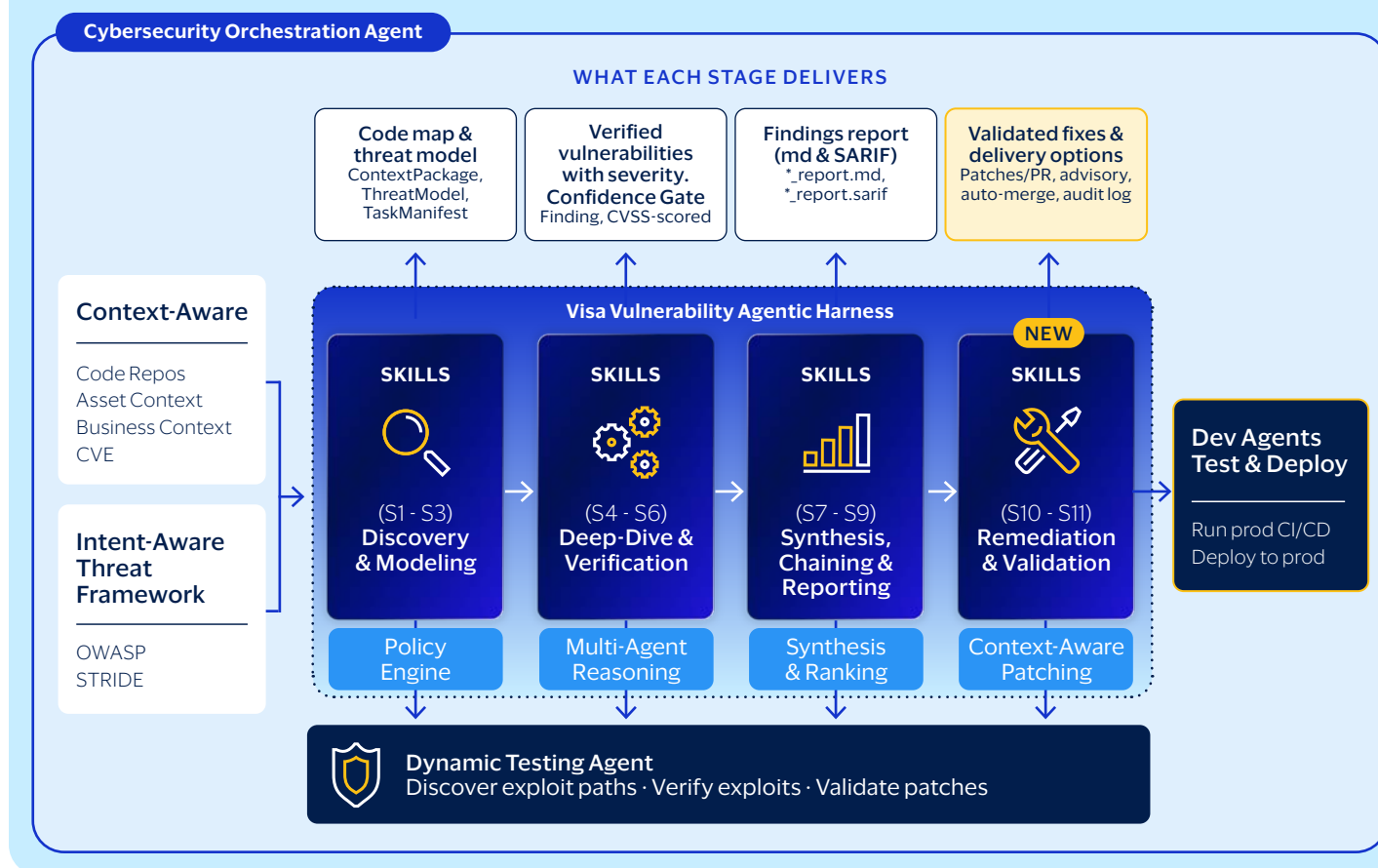
- Test Visa's zero trust and defense-in-depth architectures under AI-driven scrutiny while maintaining payments grade resiliency, auditability, and regulatory alignment.
- Build practical experience orchestrating autonomous agents against high-criticality systems and turn those lessons into a repeatable defensive operating model.

Visa prioritized high-criticality, internet-facing applications and infrastructure based on an internal asset criticality model, then expanded coverage to broader portfolios and foundational platforms.

3.2 Vulnerability Discovery and Remediation Orchestration

To move seamlessly from vulnerability discovery to a validated fix in production, Visa built a full lifecycle orchestration layer that connects multiple classes of AI agents under a unified agentic control plane. The orchestration layer governs how models are aimed, how findings are synthesized, and how remediations are created, validated, and deployed, without compromising the resilience and availability that global payments demand. The design reflects a deliberate architectural principle: AI reasoning is a component inside a governed pipeline, not the pipeline itself.

Visa AI Cybersecurity Agent Orchestration



At the core of this layer is the Visa Vulnerability Agentic Harness, a governed, version-controlled pipeline that directs frontier AI models through structured security tasks while enforcing deterministic controls, policy gates, and human oversight at every stage.

The harness provides planning, memory, governance, and orchestration. Agents within it act through skills, purpose-built capabilities that supply the domain expertise and business actions each agent needs to do real work: reading code, querying CMDB, researching vulnerabilities, running exploit chains, verifying, and reporting findings.

The orchestration layer coordinates four primary classes of agents:

- **Discovery and reporting agents**

These agents operate through the Visa Vulnerability Agentic Harness to ingest code, asset, and business context, perform threat modeled deep analysis, and emit structured, prioritized findings with replayable attack paths and remediation guidance.

- **Dynamic testing agents**

Once a candidate vulnerability is surfaced, dedicated dynamic testing agents attempt to construct and execute real exploit chains against Visa's environment, including discovering additional paths beyond the initial threat model. This adversarial challenge step ensures that only genuinely exploitable findings advance, and that, after remediation, patch quality is measured by verifying both closure of the original weakness and absence of new exploit paths.

- **Remediation agents**

Developer agents consume structured findings and generate targeted remediations, submit them for automated or human validation, and support testing and deployment through Visa's release pipeline, preserving six nines availability and Golden Rules of Change.

- **Fix Validation agents**

Dedicated validation agents verify that the fix resolves the underlying issue without introducing new risk by replaying the exploit path, confirming it fails, and only then closing findings that meet evidence-based closure criteria.

Together, these agents compress the full lifecycle from AI-discovered weakness to validated production fix — the core driver of Mean Time to Adapt (MTTA). An LLM abstraction layer decouples this governance architecture from any specific model or vendor, so Visa can swap or combine providers without changing the control plane. The discovery, reasoning, remediation,

and validation stages within this orchestration layer are executed through a structured, 11-step pipeline. Section 3.3 describes how that harness works.

3.3 Visa Vulnerability Agentic Harness

The Visa Vulnerability Agentic Harness is Visa's purpose-built execution environment for AI-powered vulnerability discovery, prioritization, remediation, and fix validation at enterprise scale. Its core function is to safely direct frontier AI models, including those from Anthropic and OpenAI, through a structured, governed sequence of tasks: mapping the attack surface, reasoning about exploitability, challenging findings adversarially, and producing structured artifacts that feed directly into policy-driven remediation and validation workflows.

The harness ensures that AI reasoning operates within defined boundaries at every stage, with deterministic controls, policy-enforced gates, and human oversight checkpoints that make every decision auditable and every finding replayable. The result is a system that combines the depth and speed of frontier AI with the governance rigor that payments-grade infrastructure demands.

Frontier models are invoked selectively at reasoning-intensive stages, while deterministic stages execute without LLM calls to preserve auditability and cost efficiency. This multi-model design means no single provider is a dependency, and the harness routes each task to the most capable model available at that stage.

Each stage is implemented as a set of reusable AI and deterministic "skills" (for example, attack surface mapper, AppSec threat modeler, vulnerability research strategist, adversarial reviewer, and exploit strategist) that can be independently tuned, versioned, and audited.

Visa is open sourcing the [Visa Vulnerability Agentic Harness as a reference implementation for AI-powered vulnerability management](#). This allows security teams to inspect the code, adapt it to their environments, and contribute improvements, using open-source transparency and collaboration to strengthen defenses and keep the harness aligned with emerging attack techniques.

Open-Source Visa Vulnerability Agentic Harness



Stage



Frontier AI-informed



Standardized inputs such as batch repository data, GitHub Enterprise metadata, CMDB records, and CVE plus control feeds flow into the harness. Across four phases and eleven stages, the pipeline combines deterministic controls with Mythos-informed reasoning to produce structured reports, SARIF artifacts, and validated candidate fixes that feed developer and deployment systems downstream.

PHASE 1

Discovery and Modeling (S1-S3)

- S1 – Explore the map and attack surface**
Build an attack surface map by ingesting code, asset & business context, data classification, external exposure, and asset criticality, normalizing repository inputs, and preparing the workspace for analysis.

- S2 – Model threats in business context**
Apply STRIDE and OWASP-aligned threat modeling to prioritize flows, components, and trust boundaries that merit detailed investigation, explicitly linking technical assets to business context.
- S3 – Strategize and prioritize the hunting plan**
Decompose the codebase across API boundaries and authentication controls, perform taint analysis on data flows, and spin up specialist sub-agents for specific components or services where deeper reasoning is warranted.

PHASE 2

Deep Dive and Verification (S4-S6)

- S4 – Research vulnerabilities by specialized lens**
Run multiple independent reasoning chains in parallel for each target and surface findings

only when there is sufficient convergence (an n-vote pattern), reducing the risk of single-path hallucinations and improving confidence scores.

- **S5 – Enforce policy before spending verification budget**
Apply explicit scope rules, severity floors, and context thresholds (e.g. PCI) so every suppression or promotion decision is traceable to a specific policy, enabling auditability and repeatability of the filtering logic.
- **S6 – Verify and adversarially challenge each finding**
Challenge surviving candidates using frontier-model-driven adversarial verification focused on exploitability in Visa's environment, tracing full exploit chains across trust boundaries and data flows rather than isolated code fragments.

PHASE 3

Synthesis, Chaining, and Reporting (S7–S9)

- **S7 – Distill high-volume findings based on business-calibrated signals**
Deduplicate across components, microservices, and batch runs, then normalize severity using CVSS enriched with business context (asset criticality, exposure, exploit availability, and environmental controls) to produce Visa-specific risk tiers and priority bands (P1–P4).
- **S8 – Synthesize findings into exploit chains and severity**
Chain individual findings into multi step exploit paths, re rank them by real world exploitability, and capture CWE mappings, attack narratives, and remediation context so engineers and AI developer agents can see “how the attack actually works,” not just single flaws in isolation.
- **S9 – Report findings into change and deployment pipelines**
Emit machine readable outputs (including SARIF, report bundles, and batch summaries) and automatically upload them into Visa's ingest APIs and internal platforms, wiring each finding into the same pipelines that control change windows, validation, and deployment.

PHASE 4

Remediation and Validation (S10–S11)

- **S10 – Remediate prioritized findings with policy driven, ready to apply candidate fixes**
Use CWE, language and framework specific playbooks to generate candidate diffs for high priority findings under strict policy gates. Diffs are

written into an isolated patches workspace and are never auto applied to the scanned repository, with kill switches and scope rules preventing unauthorized changes.

- **S11 – Validate that every fix works as intended**
Run each candidate patch through a six level gate ladder in an ephemeral git work tree – syntax, build, rescan, tests, adversarial “defeat the fix” review, and regression scan – and surface only those patches that clear every gate as recommended fixes, with an auditable remediation trail.

Within the broader orchestration layer, this eleven-stage harness is the discovery, reasoning, and remediation engine that transforms raw repositories and enterprise context into prioritized, exploit-validated work items and validated candidate fixes. Section 3.4 describes how these structured artifacts drive Visa's automation-first remediation pipeline and MTTA-focused operating model, closing the loop from AI-discovered weakness to validated production fix.

3.4 Findings and Architectural Validation

AI-driven assessments using Mythos and other frontier models identified concrete hardening opportunities across legacy applications, older libraries, and complex inter-module interactions that would have been impractical to uncover via manual or fragmented review alone, or through traditional SAST/DAST pipelines without the Visa Vulnerability Agentic Harness. Many findings surfaced recurring patterns (framework-level flaws, misconfigurations, and business logic weaknesses) prompting Visa to refactor its SSDLC pipeline by shifting security further left, strengthening testing frameworks, and integrating exploitability signals and reachability into the developer feedback loop, rather than treating each item as an isolated bug.

At the same time, Mythos confirmed the effectiveness of Visa's security foundations. Issues initially flagged as critical were often rendered non-exploitable by existing zero trust controls, network segmentation, and other defense-in-depth measures that broke the kill chain before an external actor could complete an attack. This combination of newly exposed weaknesses and independently validated strengths reinforced Visa's investment in zero trust architecture and structural controls and directly informed the non-negotiable architectural security practices described in the Appendix.



4. What We Learned: Lessons from Visa's **Deployment**

Visa's experience with Mythos and the Visa Vulnerability Agentic Harness yielded lessons that generalize to any large enterprise evaluating frontier models for security.

4.1 Validation and structured artifacts are the bottleneck to fixing risk

AI-driven scanning can generate high-volume, high-quality findings faster than human validators can review them. Without an automated validation layer, backlogs grow faster than fixes land. Visa found that automated validation using agentic workflows cut review time for many findings from 30–60 minutes of expert time to roughly 15 minutes unattended, while preserving human sign off for high-risk changes. By emitting findings as structured artifacts including CWE, file and line, replayable attack path, and remediation guidance, Visa can drop them into engineering backlogs and then replay them as executable test cases for regression. VVAH addresses this by emitting not only structured findings but also candidate patches and gate-by-gate remediation logs, so teams can validate fixes with the same agentic discipline that is applied to discovery.

4.2 Guided and calibrated reasoning beats raw scanning

The largest lift in finding quality came from threat modeling and context loading before deep analysis, combined with multi agent voting during analysis. When Mythos entered the harness with clear objectives informed by STRIDE and business context such as PCI scope, external exposure, and asset criticality, its reasoning became targeted and high value. Multi-agent voting then reduced noise and created a confidence gradient that shaped triage, where high convergence findings moved directly toward remediation and low convergence items received more intensive challenge.

4.3 Agentic programs operate at portfolio scale; Copilots do not

Chat-based copilots boost individual analyst productivity, one session and one finding at a time. Programmatic agents orchestrated by a harness run unattended across thousands of tickets and environments, enabling continuous validation at a scale no interactive tool can match.

1 KEY TAKEAWAY

Design and fund the validation tier from the outset, and treat findings as structured, replayable data, not free-form tickets.

2 KEY TAKEAWAY

Build threat modeling and context ingestion into your harness and use voting as a calibration loop, not just a filter.

3 KEY TAKEAWAY

Use interactive AI for investigation; rely on agentic, orchestrated systems for portfolio-level coverage.

4.4 Chaining & Exploitability is where frontier models add most value

Earlier tools can spot suspicious patterns; Mythos proved most valuable when asked to reason explicitly about exploitability, including preconditions, trust boundaries, and reachable sinks. Real attacks rarely use a single bug. They chain smaller weaknesses into a working exploit path, combining primitives such as memory corruption, control flow hijacking, and logic errors into a viable end-to-end compromise. Mythos reasons through that chaining process the way a senior security researcher would, not the way an automated scanner does, producing structured exploit chains that made each finding immediately actionable for engineering teams.

4.5 Keep decisions deterministic and audit traceable

Regulators and auditors are not likely to accept “the model said it is fixed” as a closure rationale. Visa requires that each automated decision — which findings are promoted, which patches are recommended, and when a fix is considered closed — is traceable to deterministic policies, evidence checkpoints, thresholds, and documented reasoning chains, and enforces read-only or tightly bounded write scopes for any agent touching production systems.

4 KEY TAKEAWAY

Design a dedicated adversarial verification stage focused on exploitability, not only pattern matching.

5 KEY TAKEAWAY

From day one, ensure the same evidence always produces the same decision, with logs and artifacts that can be explained months later.



5. How Visa is **Responding**

Visa's response to the Mythos era is organized around three strategic priorities, an agentic control plane, and Mean Time to Adapt as the core metric.

5.1 Three strategic priorities

- **Reduce the attack surface**

Shift security as far left as possible in the software lifecycle so exploitable vulnerabilities are designed out before they reach production, through extreme shift-left SSDLC practices, AI-driven AppSec (including AI-augmented SAST and reachability analysis), and continuous agentic validation in build and release pipelines.

- **Remove structural supply chain dependency**

Reduce exposure to high-risk, under-supported open-source and commercial components before they become material issues. Visa uses AI-driven exploitability and exposure metrics to identify vendors and open-source components that introduce disproportionate risk and holds them accountable through sourcing decisions, targeted engagement, and escalation. Visa also participates in Project Lightwell, a multi-year initiative with IBM and Red Hat to harden widely used upstream components using AI-validated remediation and coordinated patching.

- **Refactor defenses to be increasingly autonomous under human governance**

Use agentic AI and automation so detection, validation, and remediation can scale with threat volume and model sophistication, while preserving human oversight and clear risk ownership. This includes closed-loop agentic harnesses for discovery, chaining, remediation, and validation (such as VVAH), intelligent patch orchestration platforms, and automated compensating controls when patches are not yet available.

5.2 Agentic control plane

Visa is building an agentic control plane, a governance and orchestration layer that will manage identity, permissions, context, and policy for AI cybersecurity agents and developer agents. It ingests business and security context before agents are invoked and coordinates the Visa Vulnerability Agentic Harness, cyber validation agents, and remediation workflows within bounded surfaces and deterministic policies.

5.3 Mean Time to Adapt (MTTA)

Visa treats Mean Time to Adapt as the key operational measure of security effectiveness in the Mythos era. MTTA is the time from AI discovered weakness to a validated fix in production with evidence attached and is tracked along three dimensions:

- **Inventory freshness**

How current and complete your view is of code, configuration, and runtime deployment — including CMDB, living SBOMs, and the mapping from a finding to the real owning service, version, and dependency graph..

- **Exploitable paths per release**

After each release, how many end-to-end attack chains remain possible, not just how many findings were closed. Visa's SSDLC policy assumes every exploitable path will be exercised in production and requires it to be remediated before code is promoted.

- **Validation cycle time**

Time to produce repeatable, evidence-backed proof that a fix works and stays working.

Visa is aligning automation investments, harness design, and vendor expectations around reducing MTTA rather than optimizing legacy metrics alone, with VVAH as the reference pattern for turning AI-discovered weaknesses into validated fixes.



6. Securing the Ecosystem – Implications for **Partners & Clients**

The MTTA imperative does not stop at Visa's perimeter. The same logic that drives Visa's internal program applies to every entity that touches the global payments infrastructure. Payments and banking rely on layered, legacy-rich infrastructure and strict reliability requirements, and the ecosystem must reconcile the probabilistic behavior of agentic AI with deterministic financial controls.

6.1 Third party and open-source risk

A well-defended enterprise remains vulnerable if its vendors and open-source components are weak. Visa is making AI-specific cybersecurity posture a non-negotiable dimension of supplier due diligence, including expectations for continuous vulnerability validation, living SBOMs, and MTTA baselines. Recent work through Project Glasswing has already shown frontier models like Mythos surfacing thousands of previously unknown vulnerabilities across widely used open-source software, including decades-old-bugs in platforms such as OpenBSD and FFmpeg that had survived years of human review. This demonstrates both the power of AI to raise the floor on discovery and the need for enterprises and maintainers to have the automation and capacity to triage and remediate findings at machine speed. Visa uses AI to assess real-time exploitability in context, including asset criticality, exposure, exploit availability, and configuration, rather than relying solely on raw CVSS scores, and works with industry partners, including through initiatives such as [Project Lightwell](#), to share vulnerability data, coordinate upstream fixes, and raise the security baseline of common components.

6.2 Fraud, business logic, and the seams between systems

Frontier AI compresses the distance between technical compromise and abuse of payment flows. The same capabilities that chain code level vulnerabilities can chain weaknesses across onboarding, authentication, authorization, and dispute processes, optimizing fraud campaigns at machine speed. Visa's analysis shows that the most damaging failures are business logic and rule interaction errors where individually secure components combine into insecure outcomes. To address this, Visa operates the Visa Payment Threats Lab (VPTL), a simulation environment that replays real world fraud scenarios

against Visa's actual authorization rules, thresholds, and configurations to identify AI enabled failure modes and produce targeted hardening recommendations.

6.3 Regulatory and controls alignment

These practices largely map to existing frameworks such as PCI DSS and NIST CSF; what changes in the Mythos era is how rigorously and continuously those controls must be tested and evidenced. For AI agent governance, the NIST AI Risk Management Framework provides a structure for managing autonomy, access, oversight, and accountability in an audit defensible way. Visa recommends that clients map current controls to these frameworks with explicit consideration of agentic AI, treat human in the loop audit trails and replayable evidence as first class objectives, and extend vendor assessments to include AI specific posture and MTTA measures.



7. Conclusion

Claude Mythos and similar models mark a fundamental shift in enterprise cybersecurity: they provide powerful defensive capabilities while also creating new risks if comparable systems proliferate among adversaries.

Attackers are already using AI assisted techniques to automate and scale their operations, closing the gap between vulnerability discovery and exploitation. Visa joined Project Glasswing to deploy these tools responsibly and gain the operational knowledge needed to run automated, agentic security programs at global payments scale while maintaining strict human accountability. Early deployments have validated this strategy, uncovering complex vulnerabilities that traditional methods missed and confirming that periodic vulnerability management must give way to continuous, AI enabled security operations.

Organizations that act now to build automated, human-governed security programs, anchored on agentic control planes and closed-loop harnesses that span discovery through validated remediation, with MTTA as a core metric and rigorous ecosystem expectations, will be best positioned to preserve operational resilience as AI-driven threats mature. The window to get ahead of this shift is open but finite. By continuously advancing defensive capabilities to outpace emerging automated risks, Visa remains committed to helping ensure that the global movement of money remains secure, reliable, and resilient for every participant in the payments ecosystem.

Appendix:

Non-Negotiable Architectural Security Practices

Frontier AI collapses the time from vulnerability discovery to exploit development from weeks to minutes. Surviving machine speed exploitation requires shifting from reactive patching to proactive, structurally sound engineering practices where trust is never assumed. Based on empirical data from Visa's Mythos deployments and guidance from Project Glasswing, the following twelve architectural and design practices are treated as non-negotiable for securing critical infrastructure.

1. **Secrets never live in code:** AI models can instantly scan repositories for obfuscated keys and tokens. Enforce automated pre-commit scanning and blocking of any embedded credentials; secrets belong in hardened secret management systems, not source control.
2. **Authorization is server side, explicit, and mandatory:** Client-side checks and implicit trust models are trivial for AI to abuse. Every transaction must be authorized on the server using explicit, centrally enforced policies.
3. **"Internal" is not a security boundary:** Internal networks are no longer safe zones. Treat all internal traffic as untrusted and verify it continuously, as if it originated on the open internet.
4. **Tenant isolation is centrally enforced:** Relying on individual microservices for tenant isolation creates gaps AI can quickly find. Enforce tenant boundaries through centralized, architecture level controls that cannot be bypassed by a single service.
5. **No raw HTML or script rendering:** To prevent automated cross site scripting and injection attacks, all rendering must use modern, auto-escaping frameworks. Avoid ad hoc HTML or script generation wherever possible.
6. **Cryptography is either correct or not used:** "Homegrown" or bespoke cryptography is trivial for semantic AI models to analyze and break. Use only validated, industry standard cryptographic libraries and configurations.
7. **Inputs are hostile until proven otherwise:** Every input—external user data, internal API calls, or AI agent output—must be strictly validated and sanitized before processing. Treat AI generated content as untrusted by default.
8. **Sensitive data is never logged:** Logs are high value targets. Personally identifiable information, secrets, and security tokens must be scrubbed before data is written to any logging or observability system.
9. **Security decisions fail closed:** When authorization cannot be verified or systems encounter unexpected errors, the default must be to deny access. Do not rely on permissive fallbacks that autonomous attackers can discover and abuse.
10. **Security patterns are centralized, not copy-pasted:** Duplicated security code leads to inconsistent behavior and blind spots. Concentrate critical security logic in centralized, heavily audited libraries and services to provide a single, defensible source of truth.
11. **AI agents are identities:** Any AI agent that calls APIs, reads data, or modifies systems must be treated as a first-class identity, with scoped permissions, least privilege enforcement, full audit trails, and inclusion in IAM governance and incident response plans. Assume attackers will attempt to steal and chain agent credentials.
12. **Design for absence:** Default to removal over defense. Eliminate unused features, dead code, redundant libraries, and legacy protocols. Less code and fewer dependencies mean fewer exploitable paths for AI to construct; simplification itself is a security control.

ABOUT VISA

Visa (NYSE: V) is a world leader in digital payments, facilitating transactions between consumers, sellers, financial institutions and government entities across more than 200 countries and territories. Our mission is to connect the world through the most innovative, convenient, reliable and secure payments network, enabling individuals, businesses and economies to thrive. We believe that economies that include everyone everywhere, uplift everyone everywhere and see access as foundational to the future of money movement. Learn more at [Visa.com](https://www.visa.com).

LEGAL DISCLAIMER

This white paper contains forward-looking statements within the meaning of the U.S. Private Securities Litigation Reform Act of 1995. Forward-looking statements generally relate to future events or our future financial or operating performance, such as statements regarding our AI and cyber risk management practices, commitments and work, including goals, targets, metrics, aspirations, and related strategies. In some cases, you can identify forward-looking statements because they contain words such as “anticipates,” “aims,” “aspires,” “believes,” “commits,” “estimates,” “expects,” “intends,” “may,” “projects,” “plans,” “could,” “should,” “will,” “continue,” “strives,” and other similar expressions. All statements other than statements of historical fact could be forward-looking statements.

Forward-looking statements speak only as of the date they are made, are not guarantees of future performance, and are subject to certain risks, uncertainties, and other factors, many of which are beyond our control and are difficult to predict. These statements may be based on historical or current assumptions, estimates, standards, commitments, methodologies, targets, diligence, third-party information, internal control frameworks, and currently available data, which continue to develop and evolve. We describe risks and uncertainties that could cause actual results to differ materially from those expressed in, or implied by, any of these forward-looking statements in our SEC filings, including our most recent Annual Report on Form 10-K and our subsequent reports on Forms 10-Q and 8-K. In addition, actual results may vary due to changes in the macroeconomic and geopolitical environment, technology, artificial intelligence capabilities and adoption, cybersecurity threats, fraud and threat-actor behavior, regulation and legislation, supervisory expectations, industry standards, stakeholder engagement, third-party technologies and service providers, data availability and quality, and other unforeseen events or conditions. Except as required by law, we do not intend to update or revise any forward-looking statements as a result of new information, future events, or otherwise.

The information in this paper is provided “as is” for informational purposes only and does not constitute legal, regulatory, cybersecurity, technical, risk-management, or other professional advice. Although Visa uses reasonable efforts to include accurate and up-to-date information, Visa makes no warranties or representations, express or implied, as to the accuracy, completeness, timeliness, suitability, non-infringement, or fitness for any particular purpose of any information or recommendation provided herein. Readers are responsible for making their own independent assessments and for obtaining appropriate legal, technical, cybersecurity, regulatory, and other professional advice before implementing any practice or recommendation described in this paper. Visa assumes no liability or responsibility for any errors or omissions in the content of this paper or for any reliance on its contents, to the fullest extent permitted by applicable law.

All brand names and logos are the property of their respective owners, are used for identification purposes only and do not imply product endorsement or affiliation with Visa.



VISA