AI-Armed Adversaries

The Expanding Attack Surface in Financial Services

A Dark Arts White Paper | May 2026

> **"The question is no longer whether vulnerabilities exist. It's whether you can find, fix, and prove it before adversaries exploit them."**

In May 2026, Google's Threat Intelligence Group identified what it believes was the first zero-day exploit developed with AI assistance, in the hands of a cybercrime group, aimed at mass exploitation. GTIG chief analyst John Hultquist called it *"probably the tip of the iceberg,"* with a capability trajectory that is *"pretty sharp."* For financial services leaders, the issue is no longer whether vulnerabilities exist. It is whether the organisation can find, prioritise, fix, and evidence remediation before adversaries exploit them. This white paper examines why annual assurance models no longer match the threat environment financial services firms actually face, and what financial services leaders need to know about closing the window.

The Exposure Window: From Months to Zero Days

'Vulnerability exploitation' means an attacker breaking in through a known flaw in software, often one the vendor has already warned about, rather than tricking an employee.

Every disclosed vulnerability opens a window of exposure that stays open until the fix is deployed. The threat decides how fast that window can be attacked. You decide how fast you can close it. The problem is that those two speeds have moved apart.

The figures below describe where vulnerability exploitation already stood in 2025, before AI is factored in. The next section shows what AI now adds to it.

The Core Problem

The Verizon 2025 DBIR analysed 22,052 cyber incidents across 139 countries. Vulnerability exploitation as an initial access vector reached **20% of breaches**, a 34% increase year-on-year. Exploitation of edge devices and VPNs rose nearly **eightfold**. For espionage-motivated breaches, vulnerability exploitation was the initial access vector **70% of the time**.

The Fatal Arithmetic

The median time to fully remediate edge device vulnerabilities: **32 days**. The median time from disclosure to mass exploitation for those same vulnerabilities: **zero days**. The race to patch is being lost before it begins. Organisations are bringing calendar-based controls to a clock-speed contest.

For an entire class of vulnerabilities, the attacker's window opens before yours can close it.

20%

Of Breaches

Via vulnerability exploitation

8x

Edge Device Exploits

Rise in VPN/edge exploitation

32

Days to Patch

Median remediation time

0

Days to Exploit

Median time to mass exploitation

The Watershed: AI Now Finds and Builds Exploits

On 11 May 2026, Google GTIG identified what it believes was an AI-developed zero-day exploit targeting a popular open-source web administration tool. The vulnerability was implemented in a Python script that enabled two-factor authentication bypass, and bore hallmarks of AI authorship: unusual annotation, auto-generated documentation strings, and a hallucinated CVSS score absent from any vulnerability database.

> "The game's already begun and we expect the capability trajectory is pretty sharp. We do expect that this will be a much bigger problem, that there will be more devastating zero-day attacks done over this, especially as capabilities grow." – John Hultquist, GTIG

The work of identifying exploitable flaws is now being performed by AI systems at machine pace, and in confirmed cases this is already happening in the wild. The asymmetry is not theoretical anymore: offensive capability is accelerating faster than human defenders can keep up.

The Other Side: AI Defenders Got There First

AI does not favour attackers or defenders inherently. It favours whichever side deploys it at scale first. That advantage is no longer assumed to sit with defenders, but this is also where the defensive opportunity becomes real.

The same capability that enables AI to find and build exploits also enables AI to detect, verify, and patch vulnerabilities faster than human teams can do alone. In practice, the question is no longer whether AI changes cybersecurity. It is who operationalizes it first, and at what scale.

Big Sleep Finds a Zero-Day

In July 2025, Google's Big Sleep AI agent, developed by DeepMind and Project Zero, discovered a critical memory corruption flaw in SQLite (**CVE-2025-6965, CVSS 7.2**), affecting all versions prior to 3.50.2. Google described it as the first time an AI agent directly foiled an attempted exploitation in the wild.

CodeMender Closes the Loop

Google's experimental CodeMender agent uses Gemini's reasoning capability to automatically patch critical vulnerabilities once identified. The defensive pipeline is being rebuilt at the same clock speed as the offensive one.

The Regulatory Position: PRA, RBNZ & APRA

Within a single week in May 2026, two prudential regulators on opposite sides of the world issued aligned warnings. PRA chief executive Sam Woods cited frontier AI models' growing ability to identify vulnerabilities, and warned that the resulting pressure to patch at speed is now *"the main driver of outages"* in the financial system. In the same week, the RBNZ stated that models such as Anthropic's Mythos *"could materially amplify cyber risks"*, one of the first times a New Zealand prudential regulator has named a specific frontier AI model in this systemic operational-risk context.

RBNZ Cyber Incident Reporting (April 2024)

Banks, non-bank deposit takers, and insurers must report material cyber incidents within **72 hours**, with periodic reporting and annual self-assessments.

FMA Operational Resilience (COFI Act 2022)

Licensed entities must ensure critical technology systems are operationally resilient and notify the FMA within **72 hours** of a material incident. Boards are expected to continuously enhance their understanding of operational risk.

Privacy Act 2020 & APRA CPS 230/234

An AI-developed exploit compromising customer records may trigger multiple regulatory clocks simultaneously, including Privacy Act notification where serious harm is likely. Australian-parented NZ banks may also need to align group controls with APRA CPS 230 and CPS 234 expectations.

Why Annual Pen Testing Is No Longer Enough

The Decay Problem

An annual penetration test remains a necessary control, and for many firms a mandatory one. The issue is not that it has no value. It is that its value is concentrated on the day the report lands, and decays every day after.

An annual penetration test covers a two-to-four week window against a defined scope. Every day after it concludes, code changes, dependencies update, new CVEs are disclosed, and the threat landscape evolves. By the time the next annual test begins, the assurance value of the previous report has decayed to near zero.

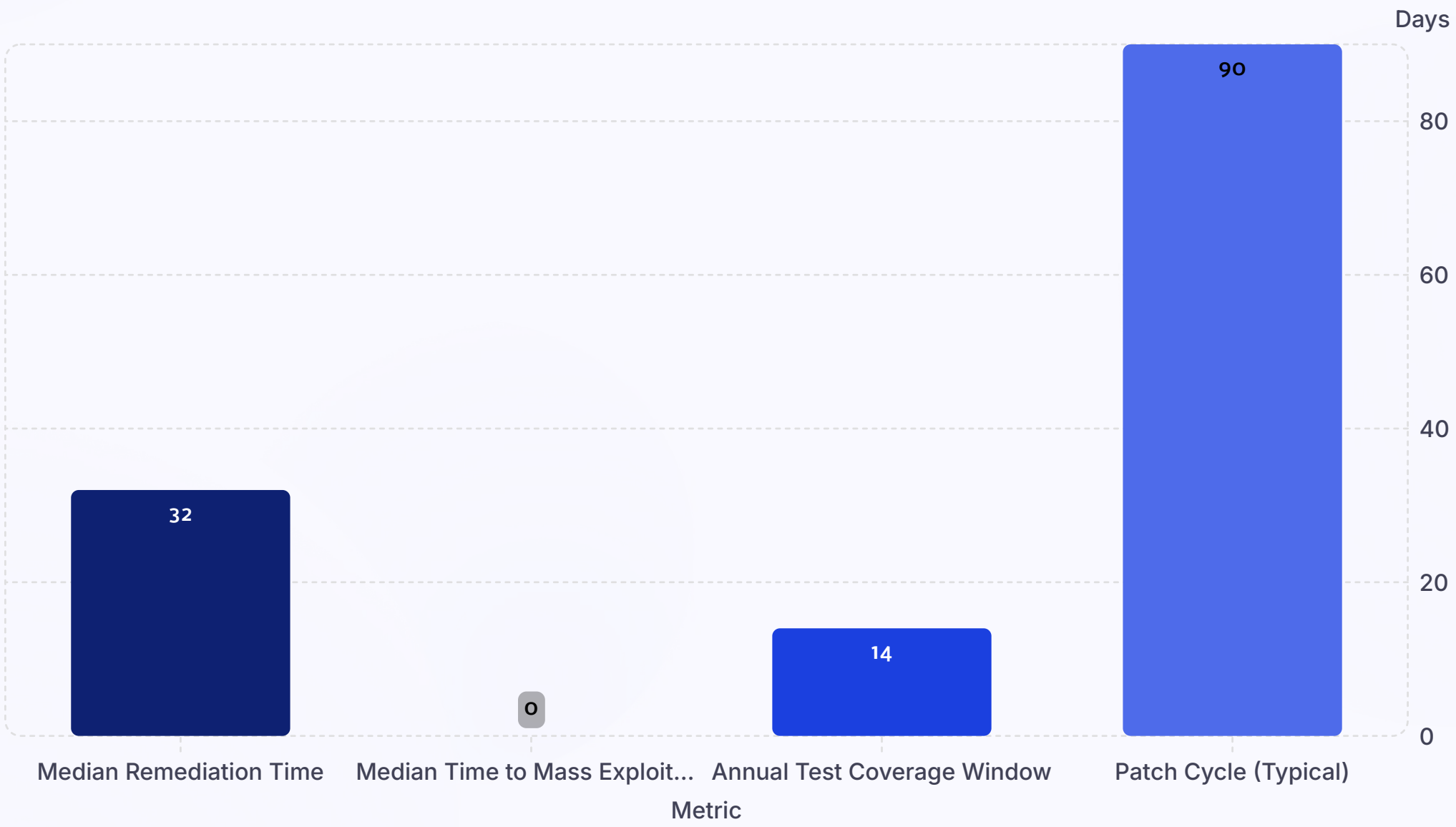
The 2025 DBIR found only **54% of edge device vulnerabilities** were fully remediated within the year - not a tooling problem, but a coverage and triage problem.

Compliance vs. Assurance

SAST tools run more frequently but are poor at finding logic flaws, chained vulnerabilities, and subtle weaknesses that AI-assisted attackers now readily discover.

> ⚠ An annual report that is months out of date by the time it is filed still satisfies the auditor. It no longer gives the business operational assurance. The gap between those two things is where the exposure lives.

The gap is not just a security gap. It is an assurance, remediation, and governance gap. Technology leaders need evidence that detection and remediation are operating at the speed of the threat environment, not at the speed of annual cycles and 32-day patch windows.



None of this replaces the annual penetration test. It wraps around it, closing the gap between one point-in-time snapshot and the next.

What Continuous AI-Assisted Assurance Looks Like

Continuous Scanning

Every code commit, dependency update, and infrastructure change triggers automated scanning. The unit of assurance is the **change**, not the year.

AI-Assisted Code Review

AI agents with code-context capability run in the CI/CD pipeline, identifying logic flaws and chained vulnerabilities that pattern-matching SAST tools miss.

Exploitation-Likelihood Triage

Firms move beyond CVSS scores alone, feeding CISA KEV updates and commercial threat intelligence directly into scanning and triage, prioritising what adversaries are *actually* doing today.

Rolling Window Metric

The metric that matters: the rolling gap between disclosure and remediation for the firm's exposed systems, versus the rolling gap between disclosure and exploitation in the wild.

The Remediation Imperative: Speed of Response

Continuous scanning solves the detection problem. It does not solve the remediation problem. A vulnerability identified at machine speed and remediated at human speed is still exposed for weeks. Without a remediation capability matched to detection throughput, the backlog grows faster than it can be cleared.

Domain-Fluent Engineers

Financial systems remediation requires engineers who understand the regulatory, accounting, and operational context, not just the code. A generic fix inside a loan origination or credit bureau platform can introduce a worse problem than the one being fixed.

AI-Augmented Engineering

AI-augmented engineers can analyse a vulnerability, identify root cause across a codebase, generate candidate fixes, and produce regression test coverage in **hours rather than days**. These tools are commercially available and in production use today.

Test Coverage as Proof

A fix without verifying tests is a hope, not a control. Each fix must be accompanied by test coverage proving the vulnerability cannot recur and that adjacent functionality is intact. AI-assisted test generation makes this practical at the required cadence.

A vulnerability surfaced Monday, scoped Tuesday, fixed and tested Wednesday, deployed by end of week: this is the new standard for firms that have done the work.

Closing the Window

> Defending at human speed against machine-speed adversaries is not a control. It is a hope.

When exploitation took weeks to develop after disclosure, annual pen tests and 32-day patch cycles were defensible. When exploitation begins on the day of disclosure and AI-developed exploits are being identified in the wild, those controls have stopped working. There is no plausible improvement in human patching cadence that closes a 32-day gap when the adversary's window is zero days.

Firms That Act Now

Build continuous AI-assisted assurance and AI-augmented remediation into their architecture before regulatory prescription arrives. They operate from strength.

Firms That Wait

Retrofit under deadline, under regulatory scrutiny, and potentially after a material incident that triggers 72-hour reporting obligations to multiple regulators simultaneously.

The watershed has passed. AI is finding and building exploits in the wild. AI is also finding, patching, and verifying them. The decisive question for leaders is no longer *whether* to engage with this shift: it is whether the firm's posture is operating at the clock speed of the threat environment it **actually faces**.

Close the Window with Dark Arts

If any of this resonates, we'd welcome a confidential conversation about where your exposure window sits today, and what it would take to close it. No pitch; just a conversation with engineers who fix these systems for a living.

Book a confidential Continuous Assurance Readiness Review

Identify where your exposure window is open, where remediation capacity is constrained, and what needs to change before the next material incident.

info@darkarts.it | 0800 DARK ARTS