AI PRODUCT SECURITY IN THE AGE OF MYTHOS · 2026

Chapter 13 · The New AppSec Metric Is Time to Evidence

Standalone reading module for LMS delivery and required reading.

FORMAT	USE	SCOPE	AUDIENCE
Standalone PDF	Required reading	Single chapter	Learners

CHAPTER 13

The New AppSec Metric Is Time to Evidence

20%

OF BREACHES

Vulnerability exploitation appeared in 20% of breaches in 2025, up 34% from the prior year. Exploits remained the most common initial infection vector for the fifth consecutive year.

VERIZON DBIR 2025 · MANDIANT M-TRENDS 2025

The unit of product security in the Mythos era is not the finding. It is the evidence package.

AI can increase finding volume. That is useful only if the team can separate signal from noise and move useful findings through ownership, containment, patching, regression tests, and detection. A finding without evidence creates triage burden. A finding with a repro, affected asset, version, preconditions, exploitability notes, owner, patch path, and test creates action.

The Useless Ticket

An AI-assisted review reports a possible authorization bypass. The ticket says the issue may allow access to another user's records. The title sounds severe. The owner is unclear. The affected version is missing. The repro is not safe to run. The preconditions are vague. The service name does not match the production asset registry. No one knows whether the vulnerable path is internet reachable. The review team multiplies this ticket by dozens. "Vertigo" is what defenders call the moment when AI finding volume exceeds triage capacity.

Engineering asks for proof. Security asks for priority. Leadership asks for status. The ticket waits.

The problem is not that the finding lacks value. The problem is that it lacks enough evidence to become a decision.

In the Mythos era, this gap becomes catastrophic. Faster finding volume creates faster triage burden unless the organization standardizes what decision-ready evidence must contain.

The evidence package is the unit that converts a signal into product behavior.

Why Volume Metrics Lie

Ticket volume measures activity. It does not measure risk reduction.

A security program can generate a higher ticket volume by finding more real issues. It can also generate more tickets because tooling got noisier, because deduplication failed, or because AI-generated hypotheses were not validated. A rising ticket count might mean the organization is in better trouble (finding real issues faster) or worse trouble (drowning in noise).

The opposite is equally true: a falling ticket count might mean risk improved, or it might mean the scanning system broke, or teams stopped filing tickets because the queue is too long.

Large vulnerability backlogs can remain unresolved long after disclosure. Discovery is accelerating. Remediation is not. This asymmetry is invisible in finding counts. A program can report "found 200 high-severity issues" while 199 of them sit in backlog and the organization's exposure actually grew.

A volume-based metric does not reveal whether the security program is helping or just creating work.

> ***Finding count is not a control metric:*** *Executives need to know whether the product-security system can convert signals into decisions faster than the next discovery round.*

Time to evidence measures exactly this: How long does it take from "possible issue" to "engineer can decide whether to patch, contain, or accept the risk"? This is the first hard step in

the response chain. Everything downstream—patching, testing, detection—depends on clearing this hurdle.

Time to evidence does not replace time-to-patch or exposure burn-down. It makes those metrics more trustworthy by proving the organization actually responded to its findings.

FIGURE 1: FINDING TO FIX: THE EVIDENCE PIPELINE CONVERTS RAW FINDINGS INTO DECISION-READY EVIDENCE PACKAGES, ROUTING THEM THROUGH TRIAGE, PRIORITIZATION, OWNERSHIP ASSIGNMENT, PATCH/CONTAINMENT EXECUTION, DETECTION/VERIFICATION, AND CLOSURE TRACKING.

The outside data explains why this matters. Verizon's 2025 DBIR reported vulnerability exploitation in 20% of breaches, up 34% from the prior report. Mandiant's M-Trends 2025 reported exploits as the most common initial infection vector for the fifth consecutive year, at 33%. This is not paperwork. Vulnerability evidence is part of how organizations decide whether the front door is still open.

The better executive question is not "How many issues did we find?" The better question is: "How fast can we turn signals into decisions?"

The Evidence Package

An evidence package standardizes what a decision-ready finding must contain. It transforms a weak signal into something an engineer can act on.

Weak finding: "Possible authorization bypass in account API endpoint. Model detected that user role parameters might not be validated correctly. Could potentially allow privilege escalation."

Strong evidence package: "Account API v3.2.1 `/accounts/{id}/admin-settings` endpoint allows authenticated user with

role=support_viewer to modify admin fields when account_id comes from user input. Vulnerability does not exist in v3.3+.

Repro: Private harness in infra repo `/tests/auth-bypass-repro.py`. Requires valid support_viewer session. Preconditions: target account must use legacy role system (affects ~300 of 15,000 customers).

Reachability: Internet-reachable via customer portal. Requires valid authentication. Exploitability: High—single request can escalate viewer to admin on affected accounts.

Affected deployments: us-east-1 (50 customers), eu-west-1 (30 customers). Version exposure: exactly v3.2.1. Older versions use different auth path (safe). Newer versions fixed the bug.

Owner: Accounts Platform team (assigned platform owner in the private tracker). Patch option: upgrade to v3.3 (ready, tested). Containment option: disable legacy role system immediately (breaks API for 300 customers temporarily).

Regression test: Added to auth test suite. See PR #8923.

Detection: Log query monitors `/accounts/{id}/admin-settings` writes from support-viewer roles.

Exception: A large enterprise customer requested delayed upgrade to Q4. Exception expires 2026-09-30. Reviewed by VP Security (signed off).

Decision needed by: 2026-06-15 (before Q3 release)."

An evidence package should make three decisions easy:

- **Decision 1**: Should this be patched (upgrade), contained (disable path), accepted (exception), or escalated (executive)?
- **Decision 2**: Who owns the next action, and do they have authority to act?
- **Decision 3**: What proof—patch applied, exception expired, detection fired—proves the risk changed?

The difference between weak and strong is actionability. Weak findings create questions. Strong packages enable decisions.

Executives often think they are buying detection. They are actually buying triage debt. Every scanner, model, red-team exercise, bug bounty, SBOM alert, dependency advisory, and AI-assisted review creates signals the organization must process. More signal is useful only when the evidence loop can absorb it. Verizon's 2025 DBIR reported that only 54% of edge-device vulnerabilities were fully remediated during the year; median remediation was 32 days. That is the real remediation rhythm most organizations operate at. Without evidence packages, without clear ownership, without a decision path, that rhythm stalls further.

Risk Only Changes When Behavior Changes

> ***Finding risk does not change by documenting it:*** *Risk changes when a vulnerable path is patched, an exposed version is removed, a feature is disabled, a detection is deployed, or an exception is accepted with an expiry date.*

Risk changes when one of these happens:

- A vulnerable path is patched
- An exposed version is removed
- A feature is disabled
- A tool scope is reduced
- A retrieval path is authorized correctly

- A regression test is added
- A detection is deployed
- A release is blocked
- An exception is accepted by the right owner with an expiry date

The evidence package is valuable because it makes one of those outcomes possible.

The Metrics That Matter

Evidence is not only for security findings; it is also how the company proves that product behavior matches the claims it has made to customers, regulators, auditors, and the market.

Time to evidence measures how long a signal takes to become decision-ready.

Time to owner measures how quickly the right team accepts accountability.

Time to containment measures how quickly exposure is reduced.

Time to patch measures how quickly the fix ships.

Time to regression test measures how quickly the fix is made durable.

Remediation velocity: how much of discovered exposure is actually closed. Today, roughly 54% of edge-device vulnerabilities are remediated within a year, while time-to-exploit has fallen to roughly 44 days. Remediation lags behind discovery.

Exploitability burn-down measures how fast reachability and impact shrink.

Control coverage measures how much of the high-risk surface is actually governed.

Exception age measures how long unresolved risk lingers without review.

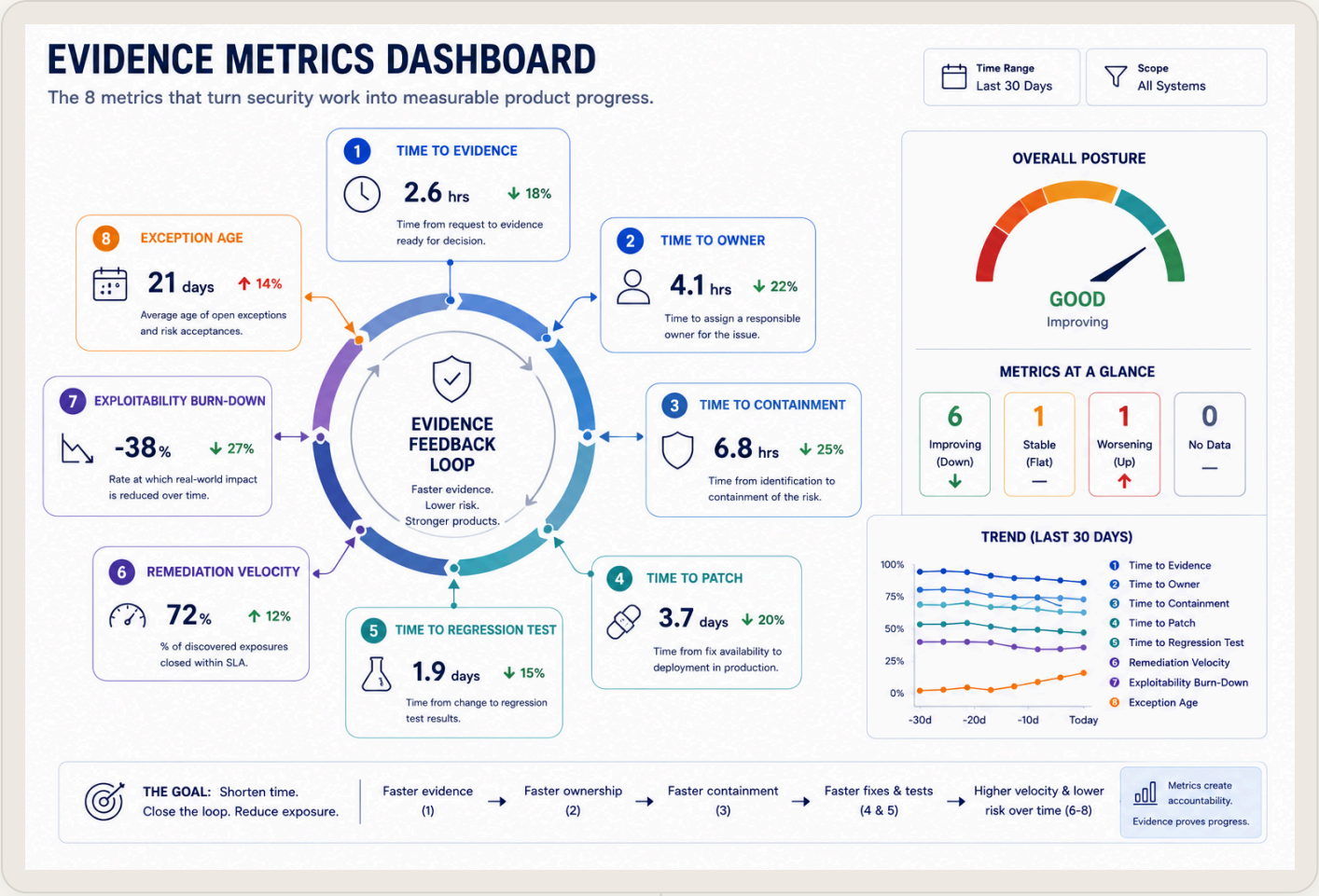

FIGURE 2: EVIDENCE METRICS DASHBOARD: THE EIGHT METRICS THAT MATTER—TIME-TO-EVIDENCE, TIME-TO-OWNER, TIME-TO-CONTAINMENT, TIME-TO-PATCH, TIME-TO-REGRESSION-TEST, REMEDIATION VELOCITY, EXPLOITABILITY BURN-DOWN, AND EXCEPTION AGE—FORM FEEDBACK LOOPS THAT PROVE WHETHER THE CONTROL PLANE IS ACTUALLY REDUCING RISK.

The Evidence Quality Ladder

The full evidence quality ladder belongs in Appendix H. The chapter-level requirement is that AI-era AppSec must measure whether findings become reproducible, owned, impact-scoped, patched or contained, regression-tested, and monitored.

Evidence-package templates, quality-level definitions, weak-to-strong example translations, executive dashboards, and metric tracking procedures – in Appendix H.

Sources

- Verizon 2025 DBIR: https://www.verizon.com/business/resources/reports/dbir/
- Mandiant M-Trends 2025: https://www.mandiant.com/resources/reports/m-trends