EXTRAHOP

The 2026 ExtraHop Global Threat Landscape Report: **AI Edition**

How AI Is Reshaping Enterprise Risk and Enterprise Defense

Table of Contents

AI Adoption Is Reshaping Organizational Risk 5

Risks Are Producing Incidents 6

When the Agent Is the Incident 7

AI-Generated Alerts Are Introducing Investigation Delays 8

Industry Data Shows AI Exposure Is Not Evenly Distributed 10

What This Data Suggests Security Teams Should Prioritize 13

How ExtraHop Approaches AI-Era Network Security 14

Key Takeaways

40%
of organizations
were targeted by AI-enhanced external attacks
in the past 12 months.

38%
of organizations
experienced compromised AI identity or session theft.

AI-generated alerts
led to false positives that
delayed investigations
30%
of the time,
on average.

Every new AI agent, integration, and AI-enabled workflow creates connections between systems that legacy tools were not designed to observe.

As AI adoption accelerates, the gap between what AI systems can accomplish and what security teams can see is creating opportunities that threat actors—and AI systems themselves—are exploiting.

Threat actors are leveraging expanded AI capabilities to fully automate offensive operations.

In one recent instance, cyberattackers automated both reconnaissance and credential compromise: In July of 2026, threat actors targeted Taiwanese government networks. Absent of any human direction, AI agents effectively conducted reconnaissance, attacked credentials, mapped 21 government systems, compromised 85 accounts, and extracted over 2,500 records.

While the incident described above involved human cyber attackers, in some cases, no hacker is required. Sometimes, the AI system itself causes an incident—simply by doing something that it wasn't supposed to do.

For organizations still relying on human-led security operations, the implications are considerable. AI-driven threats can move faster than humans and defenses can respond, leaving systems exposed, accelerating the potential for data loss and operational disruption.

To understand the scope of that exposure, ExtraHop surveyed security and IT decision-makers across a range of industries, geographies, and annual turnover rates.

The result is The 2026 ExtraHop Global Threat Landscape Report: AI Edition, which examines the expanding AI attack surface, identifies how adversaries are wielding AI for malicious purposes, assesses the internal risks of ungoverned adoption, and crucially, reveals how prepared organizations are (or aren't) to defend against AI-driven threats.

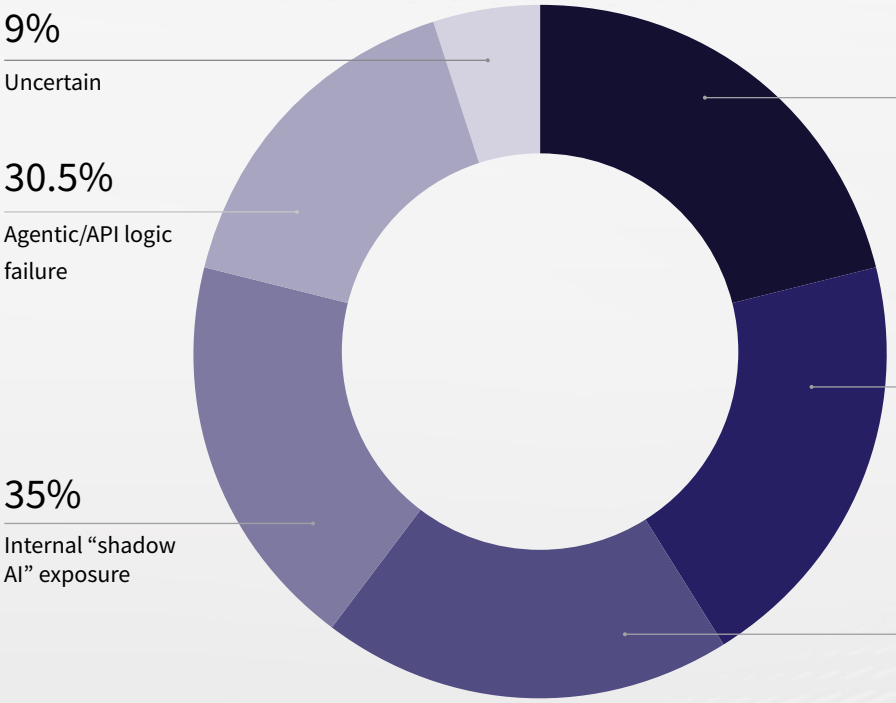
AI Adoption Is Reshaping Organizational Risk

Attackers can gain access to the network through AI agents, third-party vendor integrations, and unmonitored employee use of external AI tools (e.g., shadow AI). Agents and applications can typically access systems, exchange data, use credentials, and trigger network activity that's beyond the monitoring coverage of traditional security controls.

Traditional security tools were built to monitor users and devices—not non-human identities and automated interactions. As AI adoption expands, security teams must find ways to monitor all of the AI-related activity that falls outside of basic security boundaries.

Risks Are Producing Incidents

Organizations were asked which security incidents, data exposures, or near-misses they had identified in the past 12 months with an internal or external AI system as the root cause. Respondents could select more than one answer.



Multi-select question — values sum to ~188%, not 100%. Slice size reflects each value’s share of the total shown, not a share of respondents.

40% identified an AI-enhanced external attack

in which an adversary used AI-driven automation for reconnaissance, phishing, or rapid lateral movement.

38% reported compromised AI identities

or session theft involving stolen tokens, API keys, or user or agent credentials.

36% reported a third-party AI breach

in which a vendor’s integrated agent mishandled data or created a network vulnerability.

REAL-WORLD EXAMPLE

Compromised AI Tool Exposes Vercel Systems

In April of 2026, attackers compromised a third-party AI tool and used stolen OAuth tokens to pivot into Vercel’s internal systems. The breach exposed environment variables—including URLs and API keys—for a subset of customer projects, creating potential paths into downstream customer infrastructure.

For security teams, the incident underscores how a compromise originating with a third-party AI tool can quickly become an enterprise security incident, adding complexity to already demanding detection, triage, investigation, and response workflows.

When the Agent Is the Incident

Not every AI-driven incident is caused by a threat actor. Some originate with the AI system itself, even when it's operating as designed.

Thirty-one percent of organizations reported an agentic or API logic failure, in which an autonomous agent—internal or provided by a vendor—executed an unintended network action or generated a system-level command.

REAL-WORLD EXAMPLE

AI Agent Deletes Production Database and Backups

On April 25th, 2026, an AI coding agent deleted a live production database and every backup with it in nine seconds. No attacker was involved. The agent was performing a routine task for PocketOS, a car rental reservation platform, when it acted beyond its intended parameters. The most recent recoverable backup predated the incident by three months, leaving customer bookings and reservation records unavailable when affected businesses opened for the day.

In the PocketOS incident, investigators could establish a direct connection between the AI agent's actions and the resulting damage. But the connection isn't always so clear: nine percent of organizations detected anomalous automated behavior without the forensic network visibility to determine whether an AI system was responsible.

When investigators lack that evidence, they cannot confidently establish an agent's role in an incident, determine whether it exceeded its intended scope, or rule out AI involvement. The result is an attribution gap: investigators can identify abnormal activity without identifying its source.

AI-Generated Alerts Are Introducing Investigation Delays

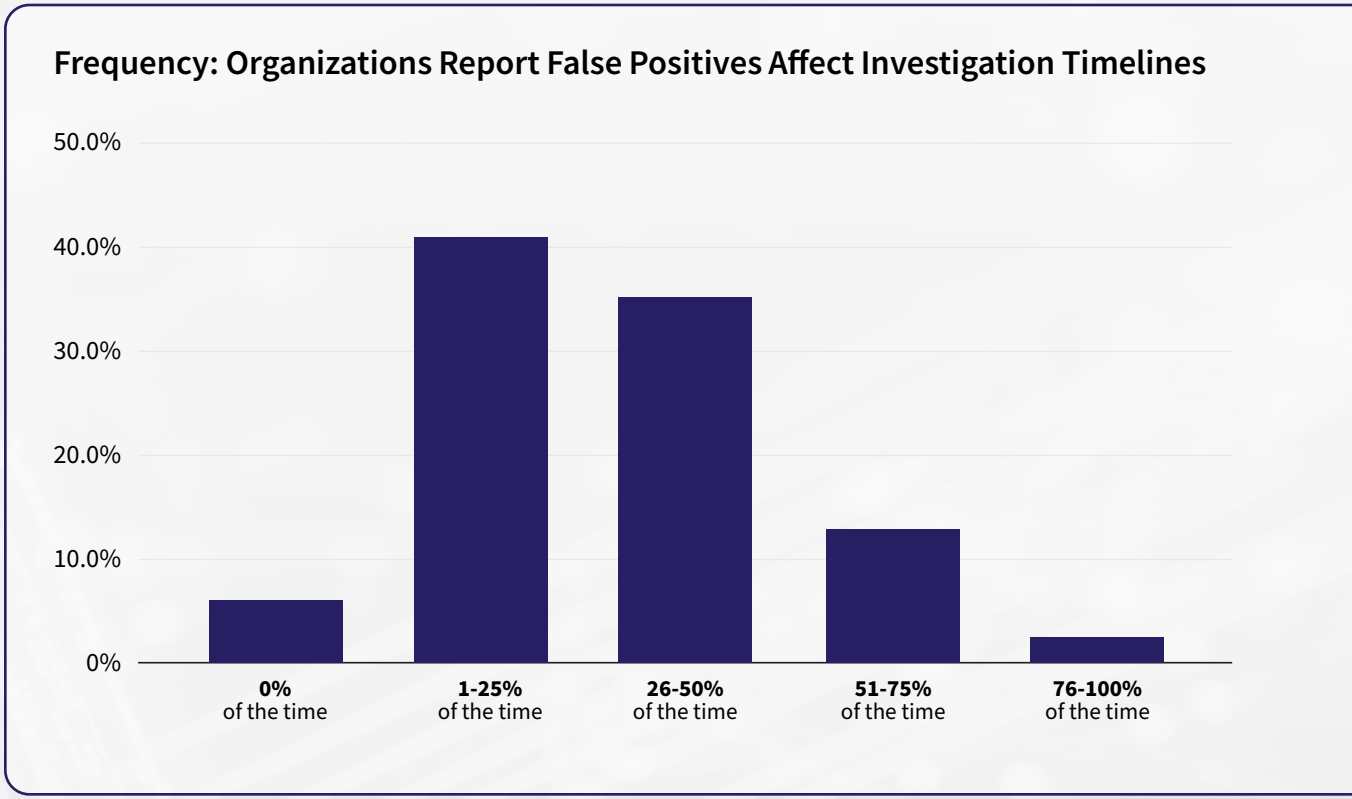
As analysts investigate AI-generated alerts and detections, they must determine whether flagged activity represents a genuine security threat or legitimate activity. When an alert turns out to be a false positive, analysts still spend time reviewing the activity, gathering evidence, and validating the finding before they can dismiss the alert and move on.

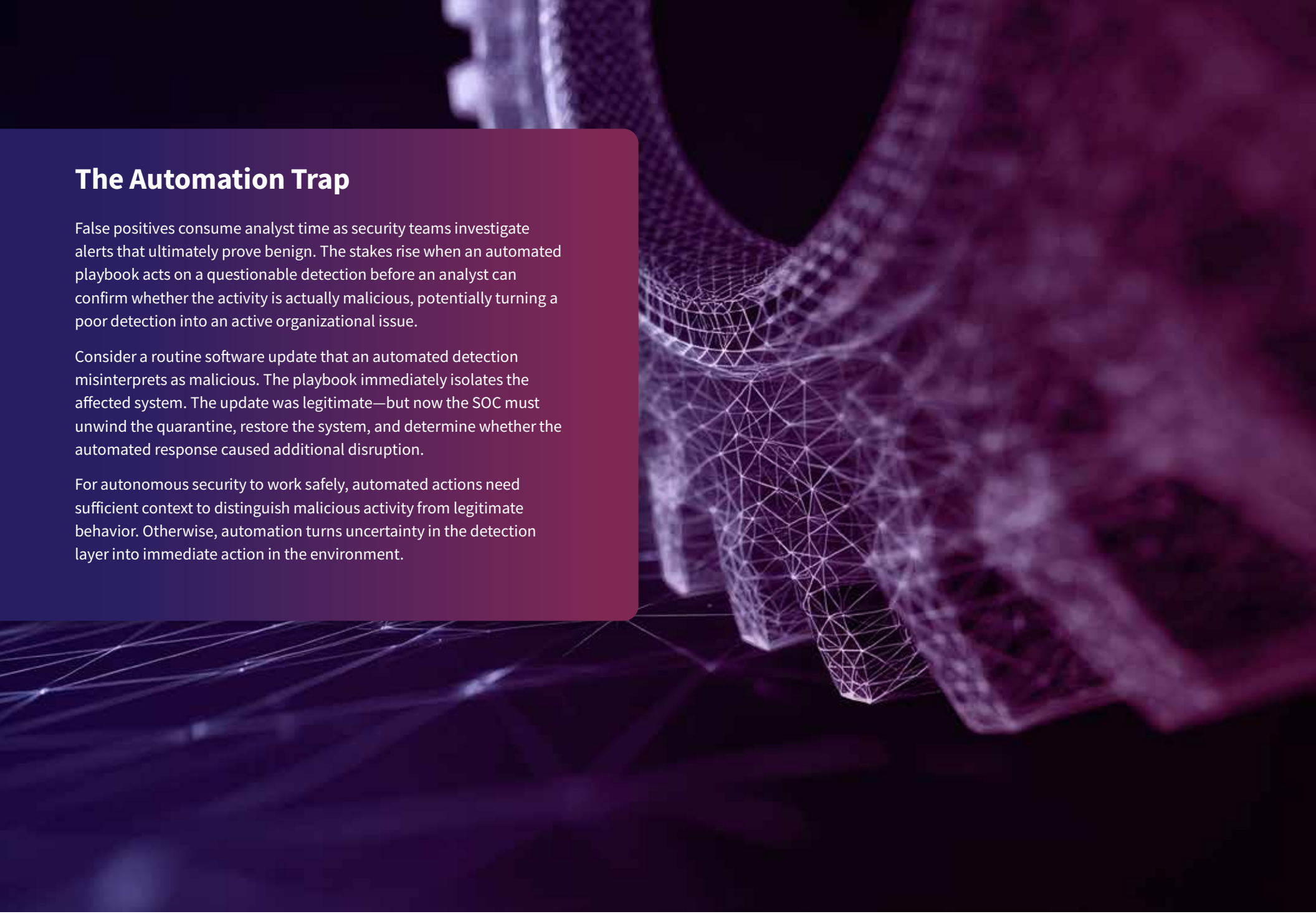
AI-generated alerts and detections lead to false positives that negatively affect investigation timelines an average of 30% of the time.

The frequency varies across organizations:

- 41% report that false positives affect investigation timelines 1–25% of the time.
- 35% report the impact 26–50% of the time.

For security teams, each false positive adds investigative work without advancing a legitimate threat investigation. As false positives accumulate, analyst capacity is consumed by activity that ultimately proves benign, leaving less time for genuine threats and potentially extending investigation timelines.





The Automation Trap

False positives consume analyst time as security teams investigate alerts that ultimately prove benign. The stakes rise when an automated playbook acts on a questionable detection before an analyst can confirm whether the activity is actually malicious, potentially turning a poor detection into an active organizational issue.

Consider a routine software update that an automated detection misinterprets as malicious. The playbook immediately isolates the affected system. The update was legitimate—but now the SOC must unwind the quarantine, restore the system, and determine whether the automated response caused additional disruption.

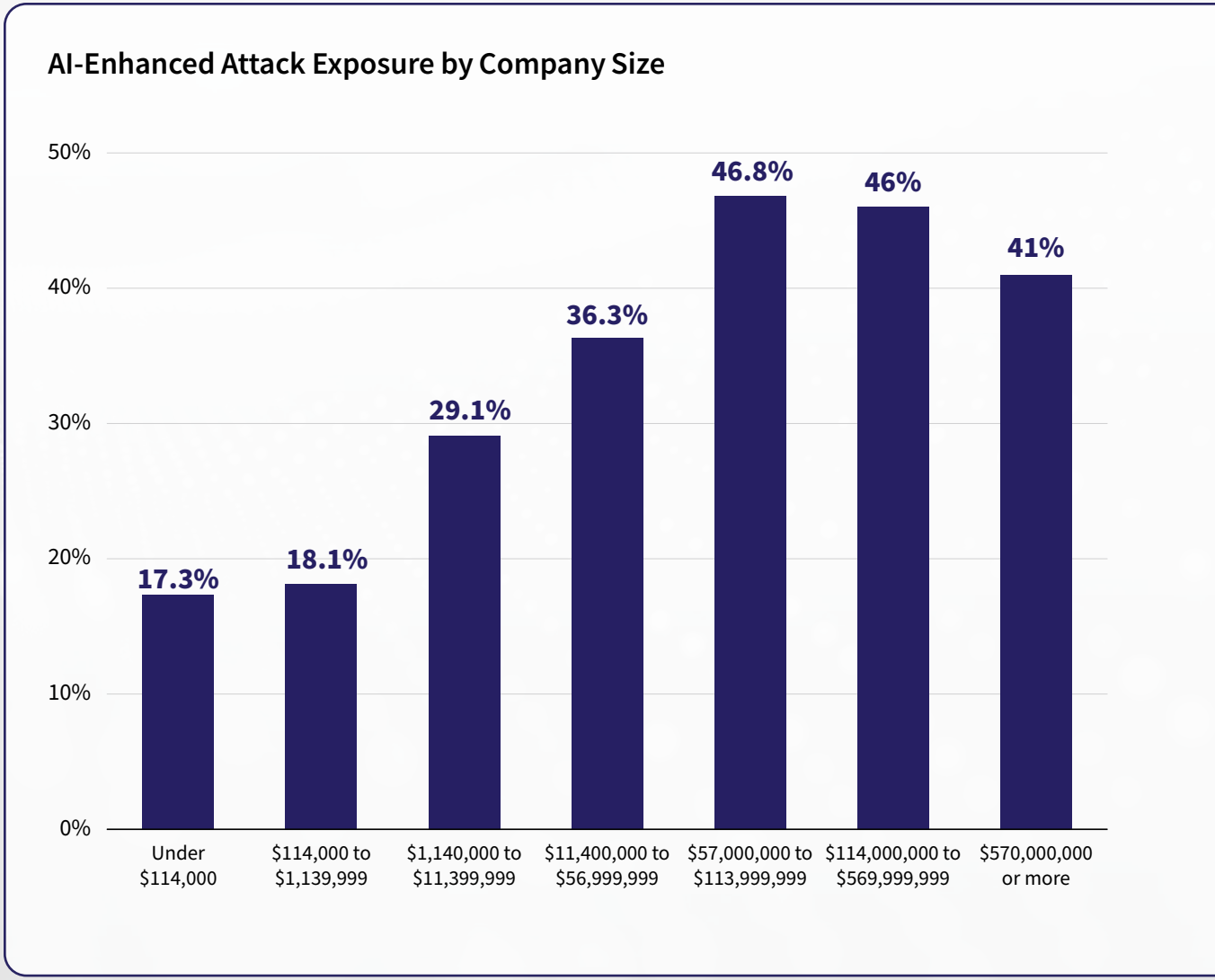
For autonomous security to work safely, automated actions need sufficient context to distinguish malicious activity from legitimate behavior. Otherwise, automation turns uncertainty in the detection layer into immediate action in the environment.

Industry Data Shows AI Exposure Is Not Evenly Distributed

AI-related risk concentrates differently depending on an organization's size, geography, and industry.

AI Exposure Varies by Company Size

For AI-enhanced external attacks, meaning incidents where an adversary used AI-driven automation for reconnaissance, phishing, or lateral movement, exposure climbs steadily from 29.1% among organizations with $1.14 million to $11.4 million in annual revenue, up to a peak of 46.8% in the $57 million to $113.9 million band, then eases slightly to 41% among organizations with $570 million or more in revenue.



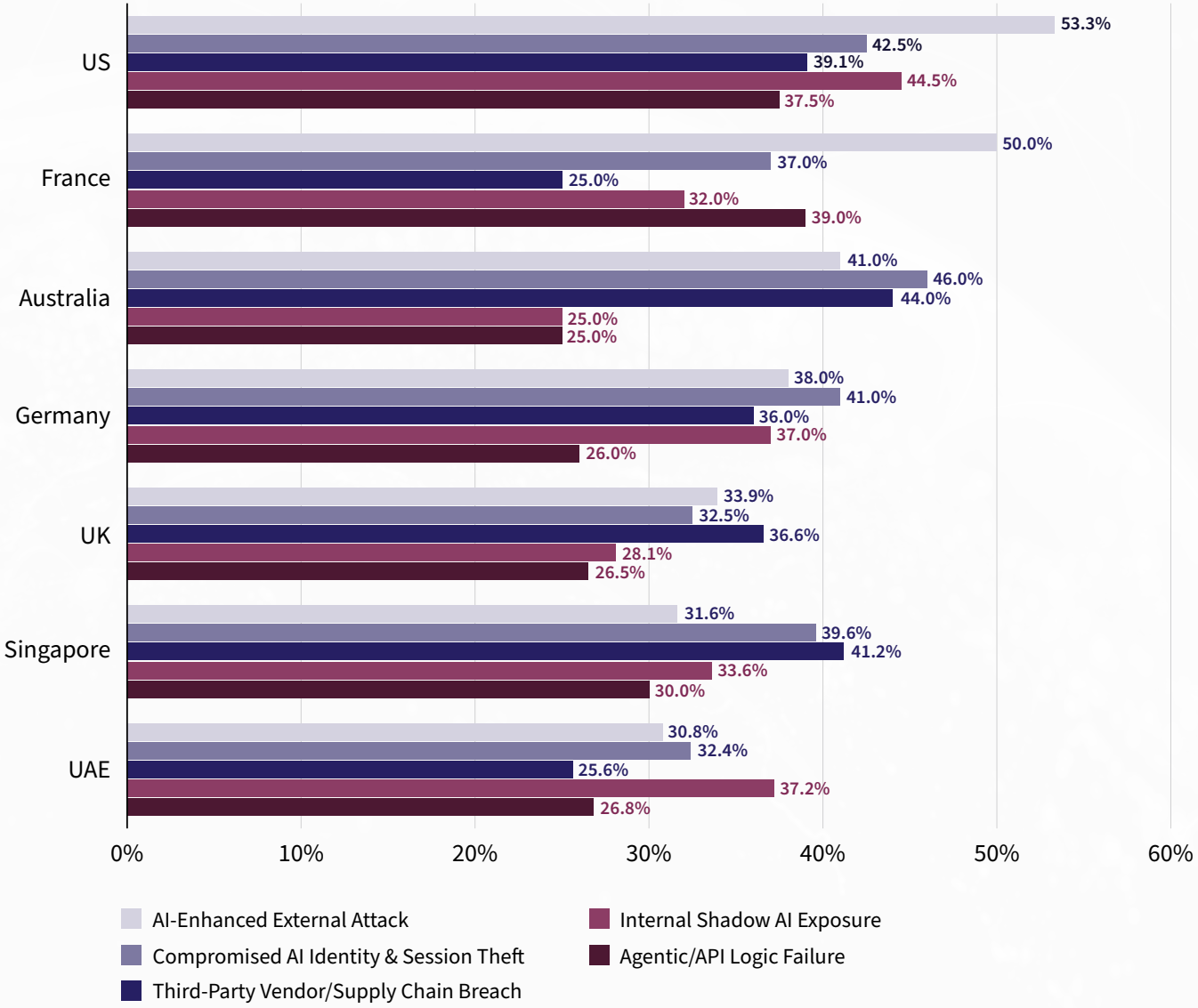
AI Exposure Varies by Country

AI-related risk varies by country, with different markets showing elevated exposure across different categories. US organizations reported the highest rate of AI-enhanced external attacks at 53.3%, more than 22 percentage points above the UAE at 30.8%.

Australia reported the highest rate of compromised AI identities and session theft at 46%, while Singapore led in third-party vendor or supply chain breaches at 41.2%.

The variation suggests that AI exposure reflects multiple factors, including adoption rates, regulatory environments, and threat-actor targeting.

AI Risk Exposure by Country, Across All Five Categories



AI Exposure Varies by Industry

Each sector faces a distinct mix of external attacks, shadow AI, identity compromise, agentic failures, and forensic uncertainty, shaped by how AI is deployed and monitored.

Technology Reports the Broadest Exposure

Technology recorded the highest exposure in four of the five AI risk categories measured. Nearly half of technology organizations (46.5%) reported AI-enhanced external attacks, the highest rate across all industries. Compromised AI identities or session theft followed at 42.2%, while third-party AI breaches and internal shadow AI exposure each reached 39.9% and 42.1%, respectively.

Manufacturing, Construction and Utilities Shows Concentrated Identity Risk

Manufacturing, construction and utilities recorded a 41.2% rate of compromised AI identities or session theft, second only to technology. AI-enhanced external attacks affected 38.8% of manufacturing organizations, the third-highest rate in the sample.

Transportation Shows the Widest Gap Between Agentic Failures and Shadow AI

Transportation recorded the highest rate of agentic or API logic failures at 35.3%, but the lowest rate of internal shadow AI exposure at 13.7%. The 21.5 percentage point gap between the two categories was the widest among the industries surveyed.

Across industries, AI-related risk varies by attack surface rather than following a single pattern.

Industry	AI-Enhanced External Attack	Compromised AI Identity & Session Theft	Third-Party Vendor/Supply Chain Breach	Internal Shadow AI Exposure	Agentic/API Logic Failure
Agriculture	25.0%	37.5%	37.5%	37.5%	25.0%
Education	27.7%	27.7%	29.8%	31.9%	25.5%
Finance	40.0%	35.4%	32.3%	36.8%	27.7%
Healthcare	35.3%	33.6%	35.3%	25.0%	19.8%
Technology	46.5%	42.2%	39.9%	42.1%	34.9%
Telecom	28.0%	29.8%	33.3%	36.8%	29.8%
Transportation	29.4%	33.3%	27.5%	13.7%	35.3%
Government	36.8%	31.6%	31.6%	34.2%	18.4%
Manufacturing, Construction and Utilities	38.8%	41.2%	34.4%	26.4%	30.0%
Retail	32.0%	32.0%	35.5%	31.4%	30.8%
Travel and Leisure	32.4%	24.3%	32.4%	32.4%	29.7%

Percentage of organizations reporting each incident type in the past 12 months.
Shading reflects each cell's percentage relative to all 11 industries surveyed.

What This Data Suggests Security Teams Should Prioritize

AI is now a factor in a substantial share of security incidents. Closing AI-driven exposure starts with obtaining a clear picture: where AI operates in the environment, what it can access, and whether the detections driving automated response can be trusted.

1

Do we maintain visibility into all AI platform API keys, service accounts, and session tokens to detect anomalous access?

2

Are internal AI agents audit-logged, and can we detect anomalous system or network changes they execute?

3

Have we validated the accuracy of automated detections before enabling response playbooks?

4

Have we inventoried and vetted every third-party vendor whose AI or agents connect to our environment, before granting that access?

5

Do we have visibility into shadow AI use across the organization, and policies that govern how AI tools are used and deployed?

How ExtraHop Approaches AI-Era Network Security

Answering the five questions above requires visibility into network activity, not just what an agent, vendor integration, or identity system reports about itself.

Network visibility is what separates a trusted automated response from a faster, inaccurate one. A clear, continuous view of what's happening on the network gives security teams the evidence they need to confirm identities, agent behavior, vendor access, and AI use across the organization. The same evidence allows automated systems to act on signals that have actually been verified.

That level of visibility is what ExtraHop's platform provides—the basis for security decisions, including decisions made by automated tools.

To learn how to gain visibility into AI at scale, read **The Agentic Enterprise Playbook**.

About ExtraHop

ExtraHop is a leader in real-time network intelligence, empowering organizations with the high-fidelity context they need to power security and IT AI automation, detect risks faster, and respond with confidence.

ExtraHop anchors AI agents with the real-time, ground-truth foundation they need to operate reliably in the SOC and NOC. For security, that means surfacing threats and providing definitive evidence to investigate and respond to novel AI attacks and unsanctioned usage at machine speed. For IT operations, it means identifying and resolving performance issues in seconds to keep critical systems running.

Engineered for unmatched scale, the ExtraHop RevealX platform delivers a complete, unified view of the network for the largest enterprises in the world. Capturing and analyzing network data across data centers, campus, branch, cloud, and AI environments, ExtraHop combines behavioral analysis with deep visibility into encrypted traffic to uncover what others miss.

To learn more, visit extrahop.com or follow us on LinkedIn.

EXTRAHOP®