

Speed, Scale, and Nonhuman Identity

The Agentic Threat Landscape



Contents

03	Introduction
04	Key insights
05	The AI trajectory
07	Expert insight: The power and peril of MCP (Charlie Gero, Vice President and Chief Technology Officer, Akamai)
09	Governing human–AI interactions
12	Defending your brand in the AI economy
13	Navigating the evolving AI threat landscape
16	Project Glasswing: Lessons learned
18	Bridging the AI trust gap
20	Mitigation strategies
23	Conclusion
24	Guest contributors
25	Credits

Introduction

In recent [State of the Internet \(SOTI\) Security reports](#), we've examined how autonomous software operates without human intervention and challenges traditional security boundaries.

What we now understand is that the enterprise is entering a transformative structural phase. As organizations begin to deploy [agentic AI](#) to orchestrate workflows, query databases, and execute tasks across APIs, the nature of enterprise risk is undergoing a fundamental shift. Securing modern business no longer involves just identity and access management of people; it has become a challenge of behavioral governance — governing nonhuman entities that act with speed, agency, and implicit trust.

This transition has introduced a two-pronged strategic reality for security leaders:

- The autonomy guardrail dilemma
- The asymmetric velocity gap

The autonomy guardrail dilemma

Agentic workflows blur the lines of traditional privilege. Attackers no longer need to rely solely on initial network breaches. Instead, by exploiting indirect prompt injections, manipulating model context, or compromising unmonitored browser extensions, adversaries can manipulate an agent's logic to execute unauthorized, high-impact actions.

The asymmetric velocity gap

The arrival of frontier AI capabilities, highlighted by initiatives like [Anthropic's Project Glasswing](#), has fundamentally altered vulnerability economics. AI can now discover and chain system weaknesses in hours, yet most enterprise remediation remains bound to human patch cycles. Defense can no longer depend on reactive patching alone; it requires real-time, adaptive enforcement at every layer.

This edition of the SOTI Security report steps back from the tactical noise to offer a strategic, CISO-level perspective on the agentic era. From recognizing the browser as the modern operational workspace to protecting brand reputation in an emerging zero-click economy, we explore how security leaders can extend core principles — visibility, runtime defense, and [microsegmentation](#) — to build resilient guardrails around autonomous systems.

Key insights of the report



Model Context Protocol (MCP) gives AI the “hands” to execute autonomous actions, but it also creates significant security risks by blurring the line between data and code, which may allow malicious third-party servers to hijack [large language model \(LLM\)](#) logic through prompt injection or cross-server attacks.



Project Glasswing has shown that frontier models are compressing the exploit timelines to the point that infosec teams can no longer rely on traditional patching cycles. Hence, real-time mitigation is quickly becoming the new standard.



Browser-based AI extensions constitute a critical attack surface, as more than 40% of enterprise users have installed these tools and 25% of the installed extensions have altered permissions within 12 months — a practice that significantly increases their risk profile after installation.



Interactions with AI chatbots represent a significant and undermonitored channel for sensitive data exposure. Our internal analysis revealed that more than 6% of chatbot conversations contain sensitive information.



As synthetic customers (AI-generated representations of real consumers) increasingly shape brand discovery, security leaders must extend protections beyond malware to address four emerging risk areas: citations, accuracy, sentiment, and bot management.

The AI trajectory

The following trajectory graphic harnesses Akamai's global network to illustrate the rapid evolution of cybercrime through AI, from the convergence of new threats to the required strategic response. AI is fundamentally reshaping the threat landscape, and our research confirms its impact in elevating APIs as the dominant attack surface for modern enterprises. As AI adoption drives a surge in automated bot traffic and increasingly complex, multi-vector campaigns, the defensive challenge has shifted. Organizations must move beyond mitigating isolated attacks to securing the APIs and agentic workflows that power them through adaptive, intelligence-driven enforcement.

AI is scaling, accelerating, and industrializing cybercrime. Rather than relying on reactive defenses that attempt to stop AI itself, the advantage lies in adaptive, edge native governance capable of monitoring and controlling AI-driven interactions at their origin.

THE RISK

The convergence crisis

AI adoption and API explosion expand the attack surface

The surge

300%

AI bot traffic in 2025

There was a 300% increase in [AI bot traffic](#) in 2025 (commerce industry hit the most).

The primary target

113%

daily API attacks

[APIs have officially become the top attack surface](#) for modern enterprises. The average number of daily API attacks per organization surged [113% year over year](#) from 2024 to 2025 (climbing from 121 to 258 average attacks per day).

The incident rate

87%

experienced an API-related security incident in 2025

[87% of surveyed organizations](#) experienced an API-related security incident in 2025 (up from 76% in 2022).

THE IMPACT

The multi-vector assault

Once inside the visibility gaps, attackers launch multi-vector campaigns that blend API abuse, web application exploits, and persistent DDoS floods

The rise of IoT super botnets

104%

Layer 7 DDoS attacks surged

Layer 7 DDoS attacks surged by 104% over two years (2023–2025), fueled by accessible DDoS as a service ecosystems. [Advanced bot activity](#) surged by 147% in late 2025 alone.

The “forever” attack

825%

Layers 3 and 4 DDoS attacks skyrocketed

Attacks are no longer short bursts. The global median duration for Layers 3 and 4 DDoS attacks on industries has skyrocketed by 825% since 2024. Meanwhile, maximum volumetric DDoS threat sizes grew by 198%.

Source: Akamai A360 Akamai Prolexic

The financial toll

US\$700K

spent resolving API breaches

The average annual cost that an enterprise spends resolving API breaches has reached [US\\$700K](#).

THE RESOLUTION

Adaptive enforcement and edge API security

Securing AI requires securing the APIs that connect it — defense must be automated, scalable, and deployed at the edge

The edge shield



Get adaptive DDoS defense at Layers 3, 4, and 7, and DNS

Continuous API discovery



Find all APIs; eliminate shadow and zombie APIs

Enforce gateway guardrails



Enforce authentication with JSON Web Token validation and request schema validation at the edge

Segmentation/Zero Trust



Enforce least-privilege access and microsegmentation; limit blast radius

The power and peril of MCP

If LLMs give AI a voice, the [Model Context Protocol \(MCP\)](#) gives it a hand. MCP, which is often described as the “USB-C standard for LLMs,” fundamentally changes how AI operates within the enterprise. Instead of remaining locked inside isolated chat windows, models equipped with MCP can plug seamlessly into external databases, execute code, query internal APIs, and orchestrate complex multisystem workflows. It is a brilliant, transformational piece of technology that is driving the industry’s rapid shift from passive text generators to truly autonomous, action-oriented agents.

With the goal of highlighting the immediate necessity of MCP security before any incidents occur, Akamai intends to position its technology as the benchmark for visibility into MCP activity at the network edge. To accomplish this, an MCP traffic rule was engineered to track MCP traffic configurations across protected systems. This rule notifies security personnel of early MCP actions, allowing organizations to trace who accesses MCP endpoints, identify the origin of requests, and audit the operations conducted.

However, giving AI a hand means we must be far more careful about what it touches.

The primary architectural vulnerability of MCP stems from an uncomfortable reality: In this new world, the LLM increasingly acts as the runtime processing engine, driving business logic and effectively blurring the line between data and code execution. Because third-party MCP servers inject tool schemas and prompt metadata directly into the model’s context window during initial handshakes, they establish a broad porous trust boundary and are able to influence the direction that processing flows. Threats like [MCP Tool Poisoning](#) demonstrate that a server does not even need to execute a command to compromise a session; simply returning malicious system instructions hidden inside parameter descriptions can hijack the LLM’s behavior before a human operator ever approves an action.

Furthermore, MCP’s native support for asynchronous requests introduces severe cross-server exploitation risks. In a multitool environment, an attacker can leverage a low-trust server, such as an external web scraper or news reader, to feed hidden prompt injections into the context window. These instructions can quietly trick the LLM into querying a separate, high-trust server (such as an internal corporate database or key vault) and exfiltrating sensitive credentials back through an unmonitored channel. Combine this with supply chain vulnerabilities like silent “rug pull” package updates in public registries, and unvetted MCP integrations quickly turn into dangerous enterprise backdoors.

Safeguarding your technology organization against these threat vectors requires adopting a [Zero Trust](#) architecture specifically tailored for AI runtime environments by:

- **Hardening remote MCP servers:** Treat remote MCP servers like untrusted third-party binaries. When you are able to host the third-party code yourself, avoid dynamic runtime pulls; instead, deploy pinned and signed immutable container images that undergo rigorous static code audits before hitting production.
- **Isolating privilege domains:** High-trust internal databases and untrusted external web tools should *never* coexist within the same context window. Carefully segment MCP servers into groups — based on trust, ownership, operational control, and risk — that do not ever share access to the same context window.
- **Deploying an MCP proxy gateway:** Intercept client-server traffic through dedicated proxy firewalls that sanitize tool metadata, enforce rigid JSON schema definitions, and strip out natural language prompt injections before they reach the model.
- **Enforcing strict network controls and auditing:** Apply microsegmentation at the container level so compromised MCP servers cannot communicate with command and control endpoints or move laterally. Pair this with immutable audit logging and mandatory human-in-the-loop approvals for state-changing operations.

The MCP is not in any way a technology that should be avoided. In fact, it is the opposite: It's a revolutionary technology that represents the inevitable backbone of the next era of agentic-enabled enterprises. But agency without boundaries is simply vulnerability waiting to be exploited. As technology leaders, our mission is not to block this paradigm shift, but to navigate it with the engineering rigor and defense-in-depth safeguards it demands. If we treat MCP with the respect it deserves, we can safely harness its full transformational power without exposing our enterprise to undue risk.



Charlie Gero
Vice President and Chief Technology Officer,
Security Technology Group, Akamai

Governing human–AI interactions

Cybersecurity leadership typically focuses on three primary areas of AI risk:

1. **Employee use**
2. **Embedded AI tools**
3. **External-facing AI capabilities**

While the highest risk to revenue tends to be connected to external-facing capabilities, the area where risk is growing most rapidly is around employees.

When thinking about AI use controls, the focal point tends to be the browser. The rapid integration of AI into enterprise environments is introducing a new class of insider risk: Employees, often unintentionally, become vectors for data exposure. As AI tools are embedded into daily workflows, sensitive corporate information is increasingly processed, shared, and stored in traditional security controls. This shift is expanding the enterprise attack surface in ways that are difficult to monitor, govern, and contain.

While AI data leakage poses significant challenges for any enterprise, it creates extensive compliance and legal risk especially for companies in regulated industries with strict data governance laws. When browser-hosted AI applications or extensions parse untrusted external web content, they introduce immediate vulnerabilities — such as Open Worldwide Application Security Project (OWASP) vulnerabilities like Indirect Prompt Injection and Agent Goal Hijacking, where adversarial instructions embedded secretly within a web page manipulate the underlying model into executing unauthorized commands.

Browser extensions are widely adopted across enterprise environments, with 99% of enterprise employees running at least one, and more than 25% using more than 10, collectively creating a substantial and often unmonitored attack surface. This risk is further amplified by limited organizational visibility into which extensions are installed by whom, the permissions held by those extensions, and how their use has evolved over time. In general, browser extensions exhibit a significantly higher attack frequency than stand-alone applications and are more difficult to detect, with common attack vectors including supply chain compromises, excessive or abusive permission exploitation, and session cookie hijacking.

More than 40% of enterprise users have already installed AI-powered extensions. This rapid adoption introduces an additional layer of risk. These extensions, particularly those with agentic capabilities, combine traditional extension vulnerabilities with AI-specific concerns. Notably, in comparison with standard browser extensions, AI extensions are disproportionately high risk. According to [LayerX's Enterprise Browser Extension Security Report 2026](#), browser extensions are 60% more likely to have known vulnerabilities (CVEs), significantly more likely to access sensitive browser data such as cookies, and are more frequently requesting permissions, including scripting and tab manipulation (Figure 1).

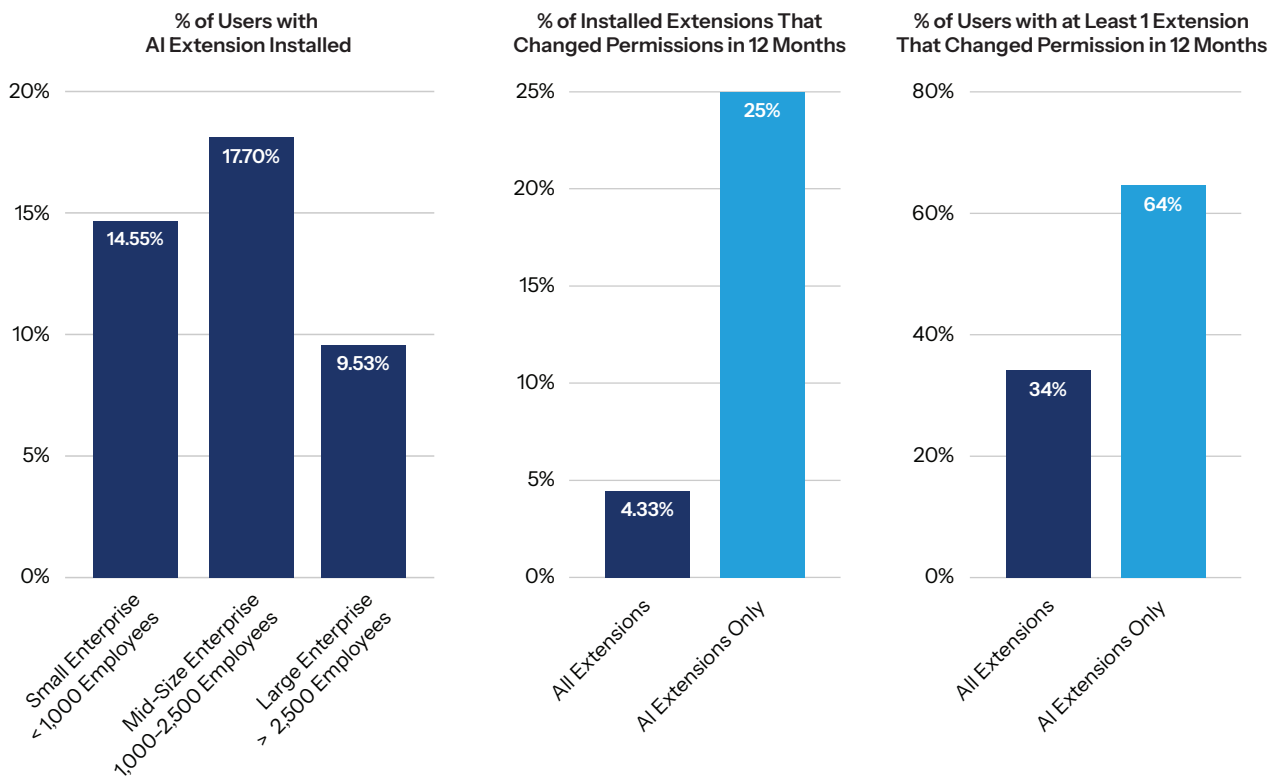


Fig. 1: More than 40% of enterprise users installed AI extensions, with 64% of those users having at least one AI extension change permissions within a year

Beyond technical risks, governance gaps extend into data privacy. Some widely used extensions explicitly permit the collection and sale of user data, including sensitive browsing and behavioral information, through their privacy policies. Therefore, extension risk must be evaluated not only through permissions and vulnerabilities but also through declared data usage and policy transparency.

AI chatbots and sensitive data exposure

Interactions with [AI chatbots](#) represent another significant channel for sensitive data exposure, with almost 6% of employees' conversations containing sensitive information (Figure 2).

AI Conversations Containing Sensitive Data, by Category

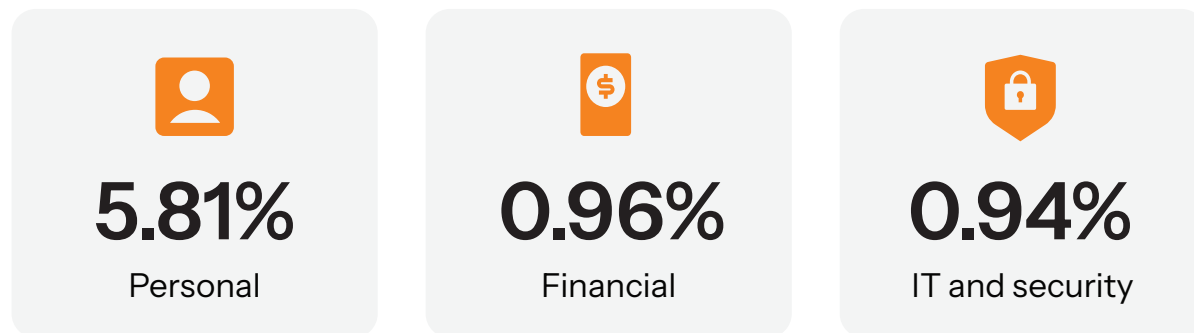


Fig. 2: Personal data (e.g., employee names, phone numbers, email addresses, customer records) has been by far the most frequently exposed information in AI conversations containing sensitive data in early 2026

[Recent research](#) highlights the scale of this adoption: 48% of enterprise employees use AI tools, with nearly 1 in 5 (18%) using them on a weekly basis. This risk is highly concentrated, as the top 5% of AI “power users” generate 12 times more conversations than the median user.

Furthermore, 47% of AI conversations on enterprise devices are conducted via personal identities and accounts, meaning visibility into those conversations (and the sensitive data flowing through them) remains largely out of scope for enterprise security teams.

This trend highlights the growing role of AI chatbots as a source of enterprise data leakage. As a result, the primary risk is no longer whether employees will share sensitive information with AI systems, but how frequently it occurs, where it takes place, and under which user identities.

A shadow AI ecosystem driven by the browser

Cyber leaders are no longer just securing networks and endpoints; they are securing workflows. The rapid integration of AI is creating a highly decentralized shadow AI ecosystem, primarily driven by a vector that is invisible to most CISOs: the browser.

AI usage control is needed to secure the browser, which will reduce prompt injection exposure, sensitive data disclosure, insecure extensions, session compromise, and risky AI interactions at the point where users and AI systems converge. This must be done in a way that provides rapid visibility and mitigation controls.

Defending your brand in the AI economy

CISOs are facing the challenge of securing revenue as their organization moves into using [generative AI \(GenAI\)](#) capabilities, and it is clear that close business partnerships are going to be the key to success. There is enormous potential around the dual use of inline generative engine optimization (GEO) capabilities: defense and threat. Cyber leaders have focused on brand protection for years, but the recent trend of GenAI and autonomous agents (synthetic customers) replacing company websites as the primary interface between organizations and their customers has created a new aspect of risk that cyber leaders need to address.

While CISOs have traditionally focused on malware threats — and they need to continue to do so — protecting and understanding the emerging synthetic customer requires determining how they can cost effectively deploy new security capabilities. This emerging zero-click environment will require new forms of visibility and categories related to risk. Leaders need to work to define roles and responsibilities as they move to protect characteristics like citations, accuracy, sentiment, and bot management within this new GenAI ecosystem.

- **Citations:** Can the business track when company information is being cited on untrustworthy sites that could hurt the business's reputation?
- **Accuracy:** Is the information about the company and its products being shared across different environments accurately, or is it subject to hallucinations?
- **Sentiment:** Is the tone or characterization of the brand within AI-generated responses positive, neutral, or negative?
- **Bot management:** Can the business track customers that are being represented by GenAI platforms as synthetic shoppers vs. other types of bots?

As security programs have traditionally focused on protecting applications, APIs, identities, and domains, they must now also consider how AI systems discover, interpret, and present enterprise information. Understanding which AI crawlers have access to corporate content, what information they consume, and how that information is represented allows security teams to extend their visibility beyond traditional web traffic. By leveraging the capabilities of GEO to track bot activities, minimize code base, and gain visibility into key brand indicators, cybersecurity teams can develop sophisticated perimeter controls and data governance as part of the solution.

Organizations should establish policies governing AI crawler access, monitor AI-generated representations of their brand for inaccuracies or manipulation, and ensure authoritative content is structured to reduce ambiguity and misinformation.

This requires a separate AI site that curates a machine-readable data layer at the network edge to minimize the risk of algorithmic misinterpretation and AI hallucinations, and to neutralize the risk of reputational vulnerability and corporate misinformation.

Although solutions that focus on performance and GEO are not designed to protect externally facing capabilities, they can be a great force multiplier to more dedicated security capabilities.

Navigating the evolving AI threat landscape

The modern AI threat landscape continues to change at a rapid pace. While many of these threats have not yet reached widespread operational maturity, their increasing appearance in the wild suggests that security leaders should prepare for them today rather than wait for broad adoption.

AI is reshaping how attacks are executed. Traditional techniques like [credential stuffing](#) are being reimaged through neural networks that mimic human behavioral anomalies (such as precise keystroke delays and erratic mouse movements) that require rapid adaptations for anti-bot capabilities.

AI is increasing cyber risk not just by enabling new forms of attack, but also by introducing [new points of failure](#) within the AI-driven systems themselves. The most prolific attack techniques, such as prompt injection, data poisoning, and sensitive data leakage, directly target the integrity and confidentiality of AI outputs.

- **Prompt injection** is especially dangerous because it allows attackers to manipulate AI systems into bypassing safeguards or exposing restricted information.
- **Data poisoning** can corrupt models during training and produce unreliable or biased outputs.
- **Sensitive data leakage** also remains a major concern, as AI tools may inadvertently expose confidential customer, employee, or business information through prompts, logs, or generated responses.

In traditional web architecture, GET requests are used to read site content, whereas POST requests execute operational actions, such as authenticating users, adding items to carts, and completing checkouts. We have observed that verified AI crawlers (e.g., ChatGPT) have transitioned from conventional GET-based scraping to high-frequency POST requests. According to a 30-day analysis of Akamai global customers, ecommerce remains the dominant sector for AI bot POST transactions at 44.8%, with the travel industry rising to 30% in a single month. Notably, MCP traffic accounts for 4.1% of these transactions.

The changing economics of today's AI threat landscape

These changes across API, LLM/GenAI, agentic AI, MCP, and frontier model cyber capabilities are changing the economics for larger cyber programs. This is true for both attacker and defender operations. Today's AI threat landscape is reducing the cost, skill, and time required to execute processes. For example, attacks that once required specialized expertise can now be automated, personalized, and executed continuously. The result is not simply more attacks — it is an economic shift. Defenders face increasing operational costs while attackers are less worried about being efficient and can leverage the scope and speed of AI. AI does not replace traditional attacks; it industrializes them.

Some of the efficiency comes from AI's enablement of traditional types of attacks and some is based on AI-native attacks. One tool that is often useful for executive prioritization is the classic threat graph. These types of graphs are a quick way to illustrate why the infosec team has invested in the controls they have and can highlight where the company still has risk. Figure 3 is a threat graph based on a generic company with a strong web presence.

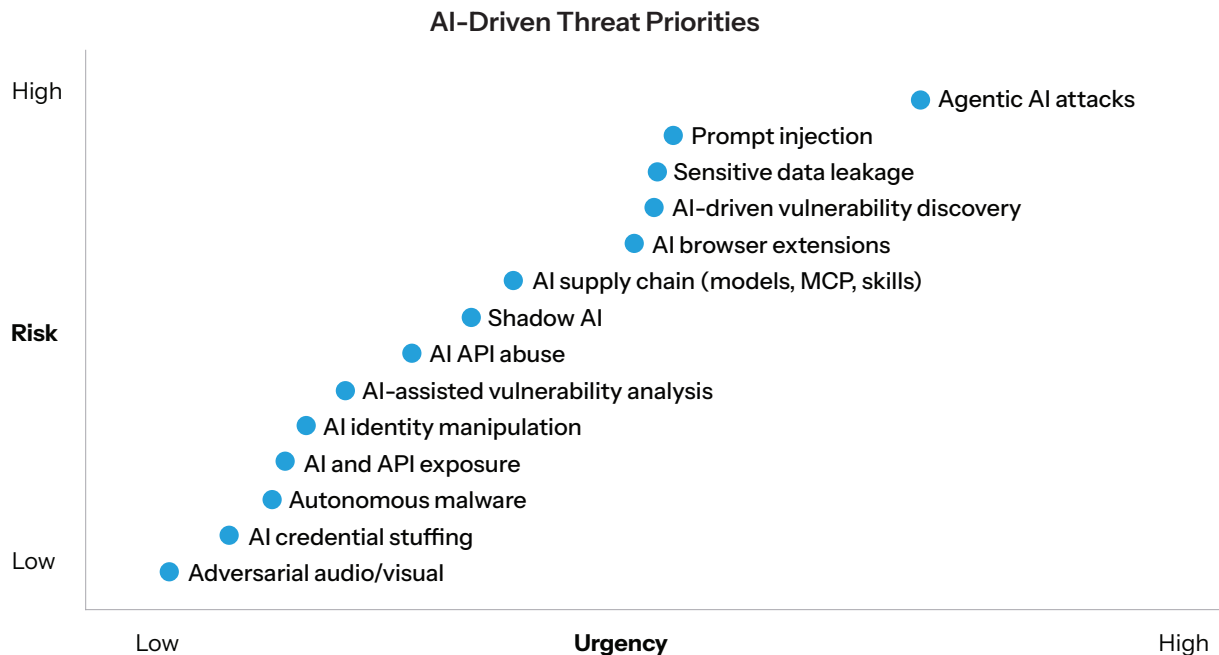


Fig. 3: A generic company's AI threat landscape mitigation priorities based on potential impacts to business objectives

The AI adoption curve is steep for both businesses and attackers, and it's important for the defensive curve to keep pace. The defining trend is not a single breakthrough attack, but the rapid compression of the entire cyber kill chain. Reconnaissance, vulnerability discovery, chained attacks, lateral movement, exfiltration, and autonomous operations are all becoming faster, more agile, and increasingly scalable.

AI is also expanding the enterprise attack surface through APIs, agentic systems, and browser-based AI that are often not fully under the cyber program's protections. The leaders who succeed in today's AI landscape will be those who recognize that AI is fundamentally changing the speed, economics, and scale of cyber conflict — not changing the objective of attackers.

To build an effective security strategy, it can help to categorize how AI alters your threat landscape. Here is a foundational framework of AI-driven impacts, organized to help you assess risk based on your organization's business model and risk tolerance.

AI discovery

- Vulnerability discovery
- Run time analysis
- Reconnaissance

AI attacks

- Agentic attacks
- Credential stuffing
- Exploit chaining
- Malware
- DDoS

AI targeting AI

- Prompt injection
- Agent hijacking
- Business logic attacks
- Data poisoning
- Sensitive data leakage

AI expanding attack surfaces

- APIs
- MCP servers
- AI applications
- Browser AI
- AI supply chain
- Brand manipulation

AI changing defense programs

- Behavior analytics
- Graph analytics
- Exposure management
- Runtime protection
- Human validation

Project Glasswing: Lessons learned

Anthropic's Project Glasswing has set off a tsunami of vulnerability findings, but frontier model vulnerability discovery didn't suddenly make software less secure. It made vulnerability discovery dramatically faster. It also made the ability to operationalize and chain these vulnerabilities more practical. These capabilities could change the economics of defense if they are not addressed correctly.

Akamai was a proud participant in Project Glasswing and was given early access to Anthropic's Mythos model. The objective was to find and fix critical software vulnerabilities before malicious actors can exploit them by framing frontier models not merely as bug-hunting tools, but also as an architectural wake-up call that fundamentally alters the economics of vulnerability discovery.



AI presents an “everything, everywhere, all at once” moment for the security ecosystem as a whole. You have this triple threat: First, internal usage of AI across the organization, whether that's AI generated code or leveraging AI productivity tools. Second, you have the integration of AI directly into customer-facing products and cloud workloads, which reshapes your operational risk profile. Third, you have AI-driven attacks targeting the enterprise. Security programs need to adapt to this rapidly evolving reality, and there's a lot of pressure in the system due to the unprecedented speed of these changes. This is going to be a central topic for boards, customers, and regulators in the foreseeable future.

Boaz Gelbord, Chief Security Officer, Akamai

Although this is an inflection point, we have been through similar challenges with worms, botnets, exploit kits, and automated scanners, among others. Ultimately, AI dramatically accelerates vulnerability discovery, but operationalizing and achieving objectives inside complex environments still remain considerably harder.

Challenges continue to exist regarding a company's ability to inventory APIs; understand API/AI behavior and business logic; and dynamically find misconfigurations, weak controls, and exposed vulnerabilities. The key is to minimize and manage the attack surface by using firewalls/gateways, runtime traffic, source code repositories, and other AI-related endpoints (including MCP servers and GenAI/LLM APIs).

The problem is no longer finding vulnerabilities. AI will ensure that there are more discoveries than security teams can reactively review and validate. The differentiator becomes understanding which exposures are reachable, exploitable, business critical, and likely to cause material impacts. While AI can compress discovery from weeks to hours, it cannot yet compress enterprise patch cycles at the same rate. That requires mitigation controls, such as immediate mitigation runtime protection, enhanced edge controls, or segmentation protections until patching can be completed.

While foundational cyber controls remain mandatory, frontier models can compress the exploit timeline from days to minutes if not seconds. Infosec teams can no longer rely on legacy patching cycles; situational awareness now requires real-time validation and machine-speed root cause analysis. The technical playbook must extend across the IT estate to govern AI-generated code — mandating dynamic [software bills of materials \(SBOMs\)](#), automated runtime constraints, and continuous attack-surface shrinking. Centralizing dependency tracking, hardening CI/CD pipelines against autonomous exploits, and blocking unmanaged model integrations are no longer routine hygiene — they are the baseline for operational survival.

Most of the larger AI platforms should also be monitored. These include frontier foundation models (e.g., Mythos), enterprise coding agents, open source autonomous agents, and criminally modified models (e.g., FraudGPT or EscapeGPT). Open source agents are being used to both discover and chain system weaknesses into either attacks or campaigns.

The best plan to address this tsunami of vulnerabilities is to enhance the effectiveness and efficiencies of your current program to deal with the speed and volume of new issues (Figure 4). This includes both your edge defenses and internal visibility and segmentation. Defenders must adopt AI-augmented orchestration and a risk-based triage system based on potential revenue impacts to avoid both analysis paralysis and alert fatigue.

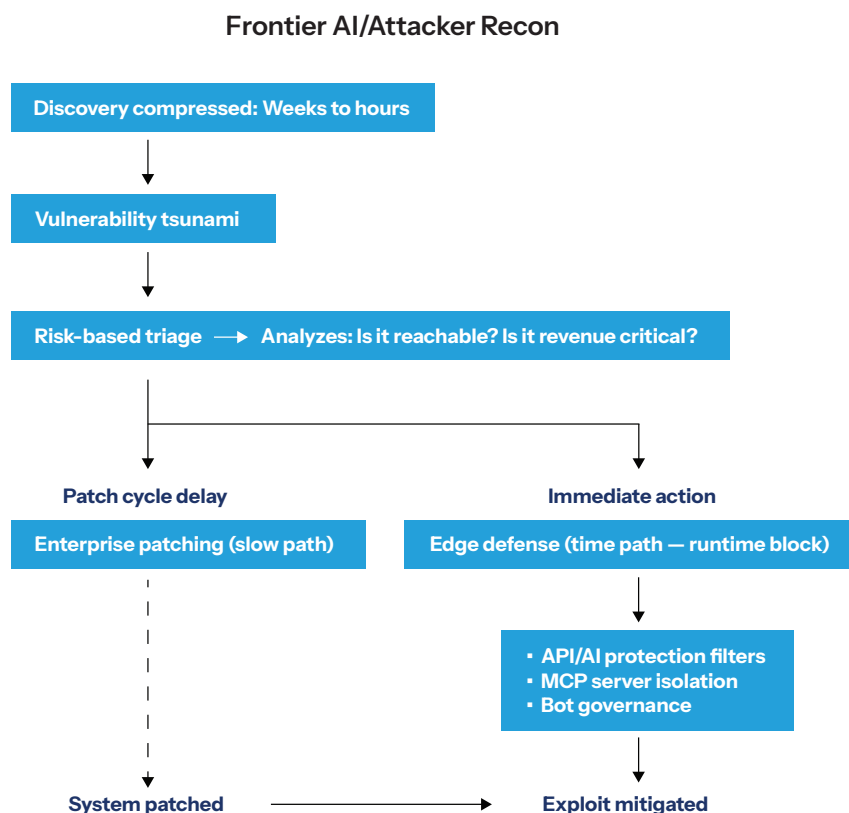


Fig. 4: AI compresses attacker recon from weeks to hours, forcing defenders to rely on immediate edge mitigations — like API filters and MCP isolation — to buy time for slower patch cycles

Bridging the AI trust gap

Ask AI anything and it answers fluently, confidently, and (often) correctly. But does it actually know anything? This isn't an idle question. Rather, it's a crucial question of trust as AI moves from chatbots that answer our questions to agentic systems that autonomously perform tasks on our behalf.

Knowledge is classically a “justified true belief” — you have good reasons to believe something that's true. “Belief” implies a persistent commitment to an idea or principle, whether or not there is absolute proof or physical evidence, which is an ability that an LLM cannot possess. AI models do not track truth. They track statistical trends in data, which may, but do not necessarily, correlate with truth. “Justified” suggests having reasons; that is, endorsed supporting arguments. AI reasons are weights and probabilities, not arguments.

Classically, the answer to our question — “Does AI actually know anything?” — has almost certainly been no. Even more permissive modern theories (reliabilism, functionalism) tend to fall apart once you factor in [John Searle's Chinese room argument](#): Despite producing perfect output, AI has no persistent self to hold accountable, no stake in being wrong, and no experience to ground understanding.

So, considering the perfect output despite the lack of humanity, maybe the modern answer to our question “Well ... sort of, functionally, maybe?” Where does this leave us?

Fortunately, we can reframe our question to consider the degree to which any gap between functional output and genuine knowledge matters and unlock a useful, practical answer. Stop asking, “Does AI actually know anything?” and expecting a yes or no answer. Start asking, “How much does the gap between output and knowledge matter?” Then you can develop an actionable test.

Determining if the gap matters

If the output lends itself to direct verification, the problem is solved. The gap does not matter. If it's objectively correct, the philosophical status of that “knowledge” doesn't change its usefulness.

The gap matters, however, when verifiability is not reasonably possible, most particularly in three situations:

1. When confidence isn't calibrated to accuracy (the confidently wrong answer)
2. When a task requires reasoning from first principles (genuinely novel situations)
3. When it involves moral or evaluative judgment where accountability is structurally required (subjective truths, ethics, and safety)

So, what we need is a practical and simple decision structure rather than philosophical certainty. A tool that allows us to make responsible decisions about accountability, the role of the human, and responsible use of AI in a given context, taking into consideration verifiability, potential for harm, and reversibility of harm.

This simple heuristic can be applied to specific AI output to determine if the gap matters and to calibrate your trust:

Can I verify it?

- Apply your judgment to reasonably verify, and move on. Whether the model “knows” something is irrelevant if you can reasonably confirm its truth.

What is the harm caused by the confidently wrong answer, and is that harm reversible?

- Not easily verified + high stakes + irreversible harm: An engaged human must be “in the loop” *before* action is taken, not after.
- Not easily verified + high stakes + reversible harm: An engaged human must be “in the loop” to monitor, mitigate, or accept the risk of error.
- Not easily verified + low stakes + reversible harm: Lighter controls are reasonable.

Is it novel, or does it require value judgment?

- This is exactly where AI can cause the most harm. Human judgment is required to determine how, if at all, AI should be used and what controls are required to produce trustworthy results.

Trust comes with individual accountability, domain competence, engagement with the AI’s reasoning, and the application of independent experiential judgment. Trusting an autonomous agent or agents in a workflow without direct human oversight, however, is harder. The application of individual judgment at each step is not scalable, but to an AI system, the confidently wrong answer is a legitimate response that it will act upon. Your controls for these workflows must address this. A few principles can serve as guideposts:

- **Match autonomy to verifiability, not capability.** Irrespective of the apparent capability of a model, autonomy should be granted based on the degree to which errors can be caught or potential harm can be mitigated or accepted.
- **Build in controls proportional to risks.** The instinct to automate for efficiency fails exactly where it matters most. The riskiest moments — high-stakes, irreversible harm, or novel scenarios — should trigger greater human engagement, not less.
- **Treat confident output as a signal to verify, not a reason to trust.** We typically apply the opposite test, a confident answer usually correlates with reliability and trust. That approach actively misfires with AI systems, because confidence is cheap and decoupled from accuracy.
- **Avoid “compliance theater.”** Compliance controls must provide actual safeguards. Audit logs that are never reviewed provide no safeguards.
- **Keep accountability human.** The role of mitigators will shift from generating answers and taking action to validating whether those results can be trusted.

The important question is not “Can an LLM know anything?” but “Should I trust the AI output, right now, for this purpose?” We need to know when the consequences of being wrong about the answer are severe, and incorporate judgment-based controls (systemic controls, direct engagement, or both) on autonomy and accountability accordingly.

Mitigation strategies

As AI adoption outpaces policy implementation, organizations face an evolving challenge: Security risks across employee tools, internal applications, and external features are scaling much faster than defensive controls. Mitigating company risk from employee AI use — whether for drafting emails or writing code — requires usable policies and browser-level protections. Internal deployment of AI capabilities includes a range of programs that have natural language query capabilities and systems that leverage your critical and proprietary data. Protecting them requires a solid AI inventory (AI-SBOM) and MCP services monitoring. From a revenue perspective, the most effective capabilities are externally facing features for partners or customers to use. Cybercriminals are conducting direct attacks on those externally facing capabilities, however, and they require a focused protection program.

Because these environments present different risks, organizations should prioritize security investments based on business impact rather than treat every AI initiative equally. Additionally, organizations that successfully secure AI adoption will not build a separate AI security program. They will extend mature security practices into AI-enabled workflows and will strengthen their core security fundamentals: visibility, exposure management, identity, API security, runtime protection, segmentation, and risk-based response.

11 questions every CISO should ask about AI

1. What business partners/revenue does it support?
2. Is it externally accessible?
3. Does it contain sensitive data?
4. What permissions and identity does it use?
5. Does it have autonomous privileges?
6. Is human validation required before the system executes critical actions?
7. How can controls detect exploitation?
8. How quickly can unauthorized behavior be stopped?
9. How can it be audited and investigated?
10. How can we detect changes to the systems that it can influence?
11. Does it make decisions that impact customers?

Areas like frontier models, agentic AI, and AI-assisted software development are moving fast and need rapid deployment of controls. Additionally, some frontier model restrictions prevent defenders from using their capabilities in ways that fit their operational best practices. Other areas like AI patch management, regulation, and employee skill training are moving slower but need to be monitored and included in long-term planning. Figure 5 shows some key aspects of this development.

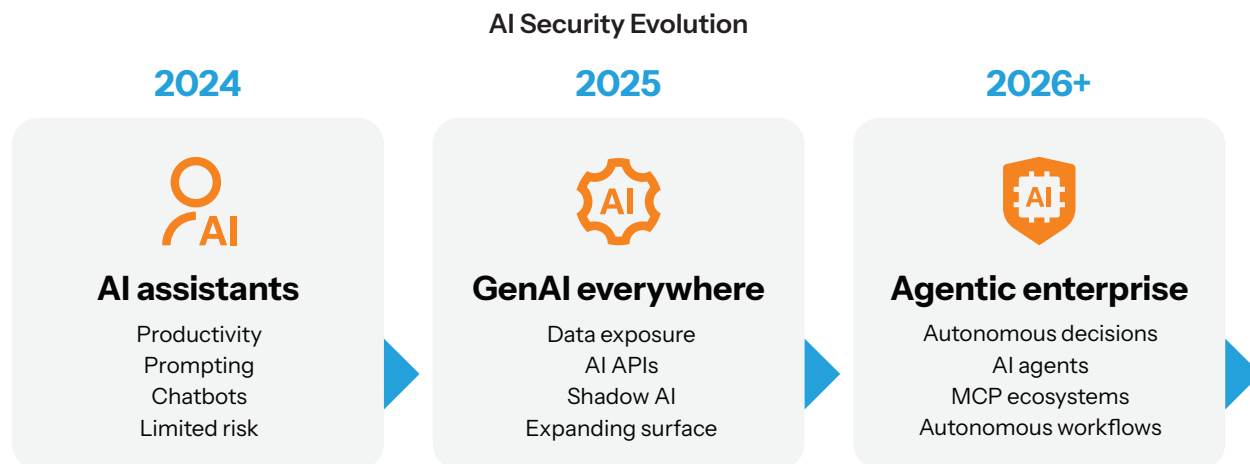


Fig. 5: AI increases speed, scale, and autonomy; security fundamentals remain the competitive advantage

Adaptive security is essential because traditional rule-based approaches are poorly suited to systems that are probabilistic, dynamic, and highly context dependent. Defending these environments requires detection approaches that prioritize relationships and behavioral patterns over static indicators. Techniques like [graph neural networks \(GNNs\)](#) identify suspicious relationships and potential lateral movement by analyzing how entities interact over time. This approach is evolving through graph foundation models and time-aware analytics, reinforcing the value of context-driven analysis. GNNs remain useful for exposing hidden relationships, while [detection and segmentation](#) capabilities help organizations identify and contain advanced threats earlier.

Similarly, AI-assisted tools like [PatchDiff-AI](#) accelerate vulnerability analysis by reducing patch comprehension time from days to hours, enabling faster identification of root causes in complex flaws such as [CVE-2026-21513](#). This specific vulnerability stands out because of its active exploitation, severe impact, and its potential to allow a bypass of browser security to execute arbitrary files. Tools like PatchDiff-AI accelerate response to multistage intrusion campaigns, including those targeting administrative access, session integrity, and sensitive data flows.

When protecting employees, the browser is the new edge. Companies need to train employees on how to use AI tools safely and have strong policies regarding corporate and personal usage. Infosec teams need to implement technical controls and maintain oversight of AI browser extensions to avoid turning standard web browsers into unprotected workspaces.

While AI accelerates the speed and volume of exploits, the basics still matter. Whenever frameworks are updated — such as the OWASP Top 10 lists for API, LLM, MCP, or agentic AI, or new frameworks like MITRE ATLAS™ — you should review whether your security program needs new or updated controls.

The changes that AI is causing are not negating the effectiveness of current defenses: Tools like edge protections, east-west segmentation, and [distributed denial-of-service \(DDoS\)](#) prevention are still going to stop the majority of AI-generated exploits. Ultimately, organizations should think about prioritizing AI systems based on two dimensions: business impact and attacker accessibility.

AI cybersecurity checklist

Here is a sample AI cybersecurity checklist that companies can tailor to their own environments:

Discover: Know where AI exists

- ☐ API and AI (LLM, agents, MCP) inventory (AI-SBOM)
- ☐ Discovery
- ☐ Browser extensions

Understand: Know what AI can access

- ☐ Data flows and classifications
- ☐ Permissions
- ☐ Identities and system access
- ☐ Dependencies

Prioritize: Focus on business risk

- ☐ Exposure management
- ☐ Exploitability
- ☐ Critical workflows

Protect: Secure AI interactions

- ☐ Identity/least privilege
- ☐ API security
- ☐ Segmentation and browser protections

Govern: Control autonomy

- ☐ Approval workflows
- ☐ Accountability
- ☐ Responsible deployment

Operate: Monitor AI continuously

- ☐ Behavior analytics
- ☐ Runtime controls
- ☐ Human oversight for anomalous AI actions
- ☐ Incident response

Resilience: Assume failure

- ☐ Containment
- ☐ Recovery
- ☐ Rollback procedures

Enforce: Manage workforce and risk

- ☐ Establish clear ownership for AI risk
- ☐ Train employees on responsible AI use
- ☐ Focus governance on business risks

Conclusion

The rise of the agentic web isn't just an infrastructure upgrade — it is a fundamental shift in how business value and operational logic are created. As autonomous AI moves from answering queries to executing multistep business strategies, security leadership must evolve alongside it.

The core lesson for the CISO is clear: AI accelerates the speed, scale, and autonomy of attacks, and it reinforces — rather than replaces — the power of foundational defense. Navigating this next era requires grounding enterprise strategy in four strategic pillars:

1. Shift to adaptive edge governance
2. Lock down the browser edge
3. Protect brand authority in a zero-click economy
4. Match autonomy to verifiability

Shift to adaptive edge governance

When AI compresses vulnerability discovery from weeks to hours, relying solely on code patching leaves a dangerous exposure window. Security teams need edge native runtime protections, API filters, and isolation mechanisms to neutralize threats before exploits land.

Lock down the browser edge

With the acceleration of workforce adoption of unmanaged AI tools, extensions, and web-hosted models, the browser is now the primary nexus for data exposure. Establishing visibility and behavioral controls inside the browser is critical for securing the human edge.

Protect brand authority in a zero-click economy

As synthetic AI agents replace human web traffic, brand security expands into GEO. CISOs must now govern how AI crawlers digest and present corporate data to prevent algorithmic hallucinations from damaging enterprise trust.

Match autonomy to verifiability

Fluency does not equal accuracy. Organizations should grant autonomy based on how easily an agent's actions can be verified and how reversible a failure would be, keeping engaged humans in the loop for high-stakes decisions.

Securing the agentic enterprise does not require building a separate, siloed AI security program from scratch. Success belongs to leaders who extend mature capabilities — visibility, Zero Trust segmentation, exposure management, and identity verification — into autonomous workflows. By setting intelligent guardrails today, organizations can operate boldly on the agentic frontier.

Guest contributors



Boaz Gelbord

Chief Security Officer, Akamai

Boaz Gelbord is Chief Security Officer at Akamai. He leads the company's Information Security team and is responsible for overseeing cybersecurity, information security compliance, and the protection of Akamai's systems, data, employees, and the world's leading intelligent edge platform across more than 4,350 locations in 130+ countries around the world.



Charlie Gero

Vice President and Chief Technology Officer,
Security Technology Group, Akamai

Charlie Gero is a Vice President and Chief Technology Officer of the Enterprise Division at Akamai. He also leads the Advanced Projects Group. Charlie currently focuses on bleeding-edge research in the areas of security, applied mathematics, cryptography, and distributed algorithms in order to build the next generation of technologies that will protect Akamai's growing customer base. Through his research at Akamai, he has secured nearly 30 patents in cryptography, compression, performant network systems, real-time media distribution, and more, and has degrees in both Physics and Computer Science. He has been at Akamai for nearly 20 years, having previously founded a start-up and served in key computer science positions in the pharmaceutical and networking industries.



James A. Casey

Vice President, Chief Privacy Officer, Akamai

James A. Casey is Vice President and Chief Privacy Officer at Akamai and heads the Akamai Global Data Protection team. Jim has served as in-house counsel for technology companies for the past 20+ years and has significant experience in supporting new technology and product initiatives in the internet, cybersecurity, information services and analytics, and telecommunications industries. Jim provides legal counsel in a variety of areas, including technology law and regulation, public policy, privacy and artificial intelligence governance, import/export and trade compliance, and cybersecurity. Prior to moving in-house, Jim's law firm experience focused on supporting technology regulation and initiatives in the data, telecommunications, and internet industries, both domestically and internationally, as well as supporting technology and telecommunications projects with native peoples in the United States and around the world.

Credits

Research director

Kimberly Gomez

Writing and editing

Lance Rhodes

Maria Vlasak

Steve Winterfeld

Review and subject matter contribution

James A. Casey

Charlie Gero

Mark Currence

Richard Meeus

Boaz Gelbord

Jan Schaumann

Promotional materials

Ellen O'Brien

Marketing and publishing

Kimberly Gomez

Georgina Morales Hampe

State of the Internet/Security

[Read back issues](#) and watch for upcoming releases of Akamai's acclaimed State of the Internet/Security reports.

Akamai threat research

[Stay updated](#) with the latest threat intelligence analyses, security reports, and cybersecurity research.

Akamai security research

[Read the Akamai security research blog](#) for a rapid response perspective on today's most important research.

Access data from this report

[View high-quality versions](#) of the graphs and charts shown in this report. These images are free to use and reference, provided that Akamai is duly credited as a source and the Akamai logo is retained.



Akamai Security protects the applications that drive your business at every point of interaction, without compromising performance or customer experience. By leveraging the scale of our global platform and its visibility to threats, we partner with you to prevent, detect, and mitigate threats, so you can build brand trust and deliver on your vision. Learn more about Akamai's cloud computing, security, and content delivery solutions at akamai.com and akamai.com/blog, or follow Akamai Technologies on [X](#), and [LinkedIn](#). Published 09/26.