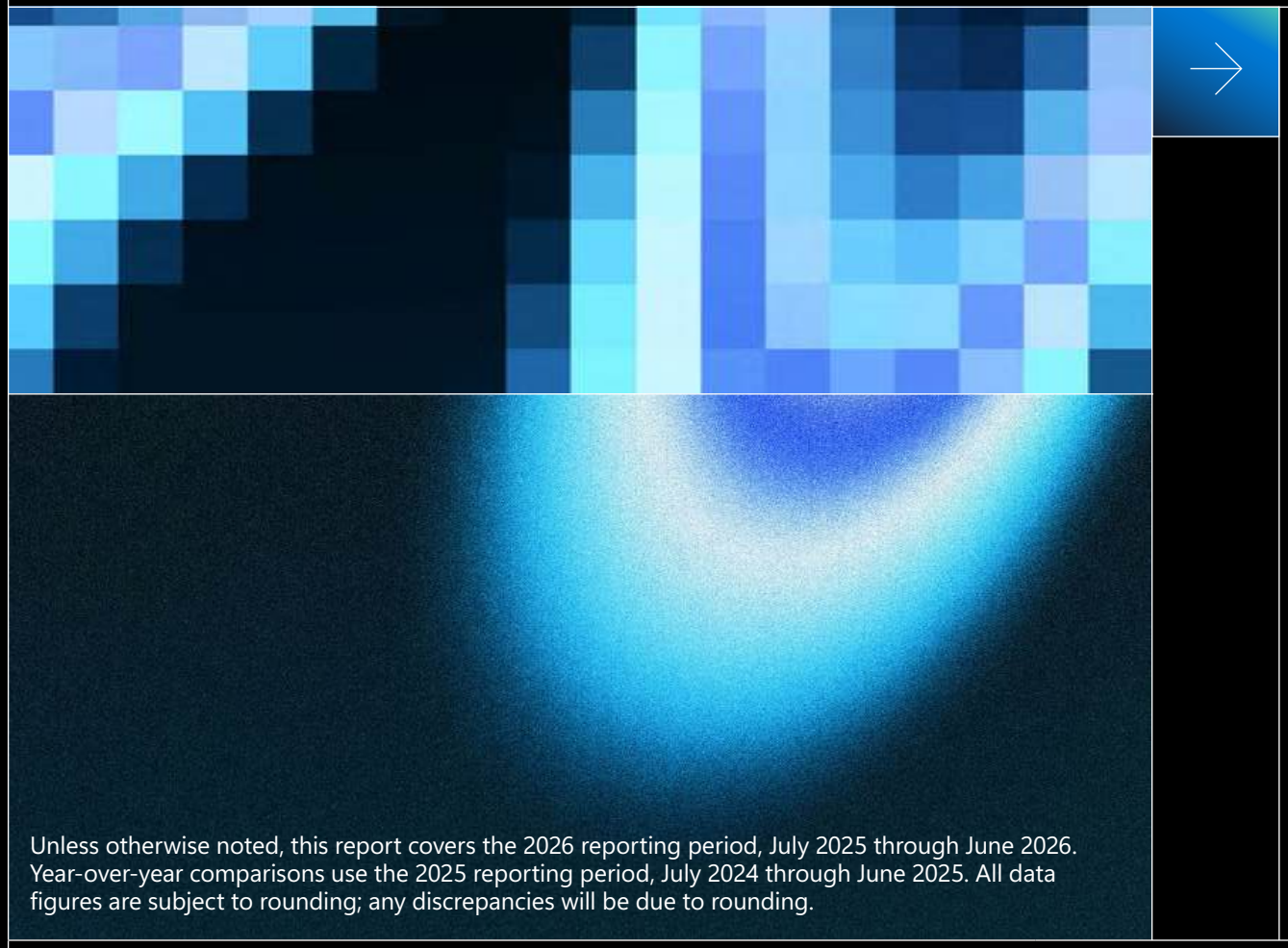
Microsoft

Decoding emergence

2026 Microsoft Digital Defense Report

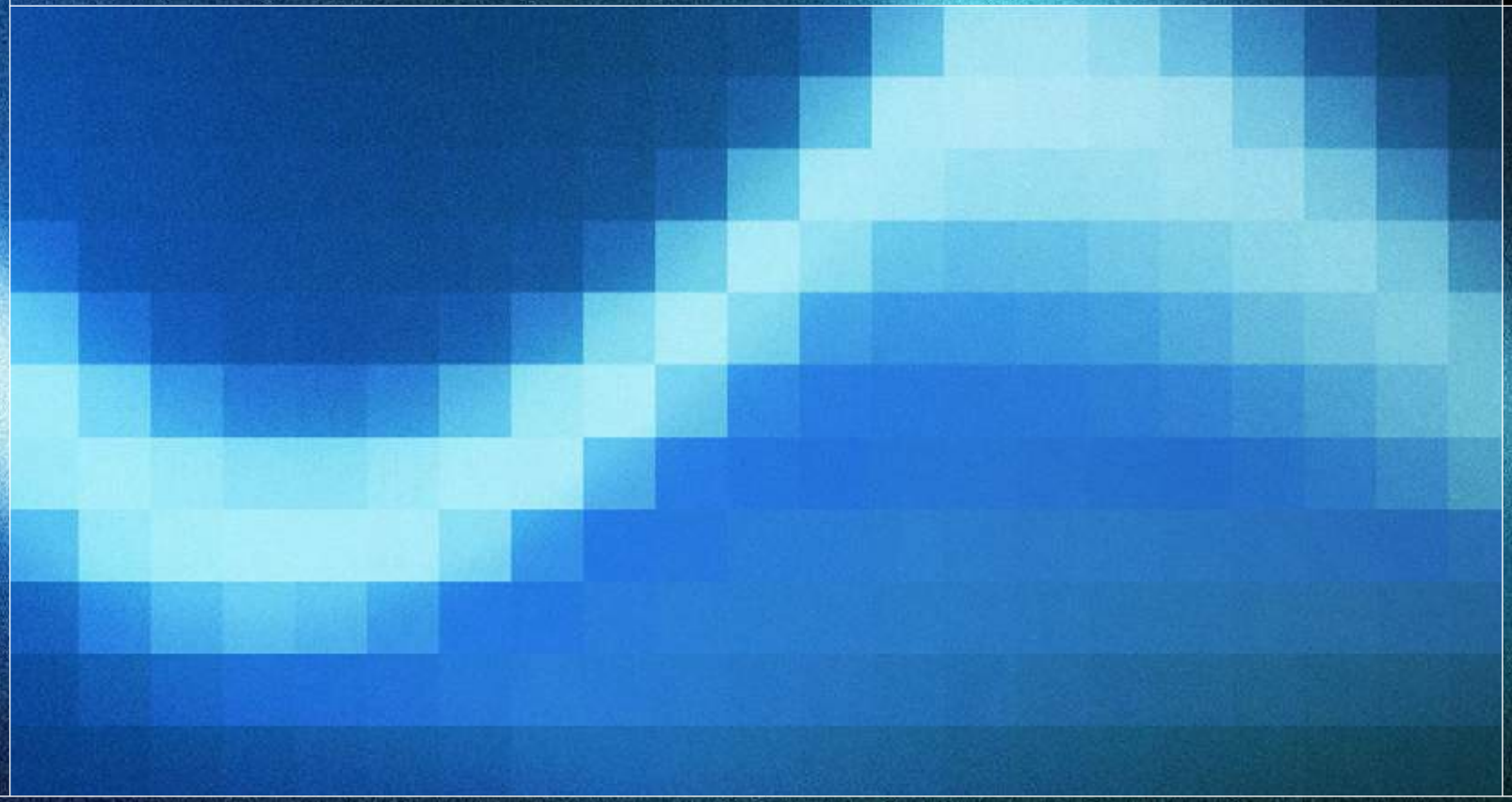
What's in the report?
Contents listing



Unless otherwise noted, this report covers the 2026 reporting period, July 2025 through June 2026. Year-over-year comparisons use the 2025 reporting period, July 2024 through June 2025. All data figures are subject to rounding; any discrepancies will be due to rounding.

Introduction	**3**
Opening statement	4
Ten security priorities for organizations	7
Top 10 takeaways from this report	8
AI	**9**
Understanding the AI threat landscape	10
Securing the AI ecosystem	21
AI as the new attack surface	27
The threat landscape	**31**
The threat landscape in 2026	32
Nation-state and geopolitical threats	38
How modern intrusions begin	44
Detection, intelligence, and disruption	54
Securing the foundations	59
Cybercrime	**62**
Identity is at the center of modern cyber risk	63
How attackers gain access	66
Legitimate access becomes the threat	69
Ransomware: From access to impact	75
Cybercrime disruption and collective defense	78
Resilience	**81**
Trust, software, and supply chain security	82
Critical infrastructure and operational technology	85
National resilience and geopolitical security	90
Preparing for the next security transaction	97
Appendix	**99**
Glossary	100
References	102

Introduction



Opening statement 4
Our unique vantage point 5
About this report 6
Ten security priorities for organizations 7
Top 10 takeaways from this report 8

Introduction

The era of interconnected risks

Opening statement
By Terrell Cox, CVP and Deputy CISO, Customer Security Management Office and Mike Yeh, Vice President, Deputy General Counsel, Customer Security & Trust



Terrell Cox
CVP and Deputy CISO, Customer Security Management Office



Mike Yeh
Vice President, Deputy General Counsel, Customer Security & Trust

The challenge confronting security leaders today is no longer a lack of information. It is the growing complexity of the environments they are responsible for protecting.

This year's Microsoft Digital Defense Report reveals a common pattern across AI, nation-state activity, cybercrime, and resilience: risk is becoming increasingly interconnected. Threats increasingly move through the trusted relationships connecting identities, software and supply chains, partners, cloud platforms, data, and AI systems. When one of those trust pathways is compromised, consequences can travel far beyond the initial point of entry.

This shift reflects a fundamental change in what security leaders are actually responsible for. CISOs today are managing ecosystems, not just environments: networks of technologies, identities, partners, vendors, and dependencies whose security outcomes are deeply intertwined. This ecosystem lens applies across every trend in this report.

AI is one example of this transformation. While much of the public conversation focuses on AI as a technology and national security priority, an equally important story is what AI reveals about the future of cybersecurity. AI is accelerating discovery, analysis, adaptation, and automation across both offense and defense. It is reshaping how vulnerabilities are identified, how attacks are developed, how defenders prioritize risk, and how organizations respond to emerging threats.

The era of interconnected risks (continued)

It has compressed the security cycle from days to minutes. Its influence extends far beyond AI systems themselves and into the very fabric of our societies.

Nation-state activity is evolving. Cyber operations, traditionally intertwined with economic objectives, geopolitical influence, and national and economic security priorities, have more reach than ever. They extend to accessing critical infrastructure for data exfiltration, pre-positioning, or even destructive attacks, blending the cyber world with the physical world.

The same pattern can be seen in cybercrime. Criminal operations have evolved into highly specialized ecosystems where access brokers, ransomware operators, credential theft services, fraud networks, and malware providers function as interconnected components of a broader economy. These ecosystems increase efficiency, lower barriers to entry, and enable sophisticated attacks to scale with unprecedented reach.

At this scale, the value is not any single signal or control. It is the ability to connect signals, threat intelligence, enterprise context, and action across the security estate. Integrated visibility turns each additional signal into more useful context and enables faster, coordinated defense.

In a time of interconnected risk, organizations need to focus on resilience

Across all of these trends, one observation remains constant: the expanding complexity of modern digital environments continues to create opportunities for adversaries to exploit gaps in visibility, access, and governance. Resilience therefore remains a key priority. Resilience extends beyond an organization's ability to recover from disruption.

Our unique vantage point

>165tr
security signals
daily

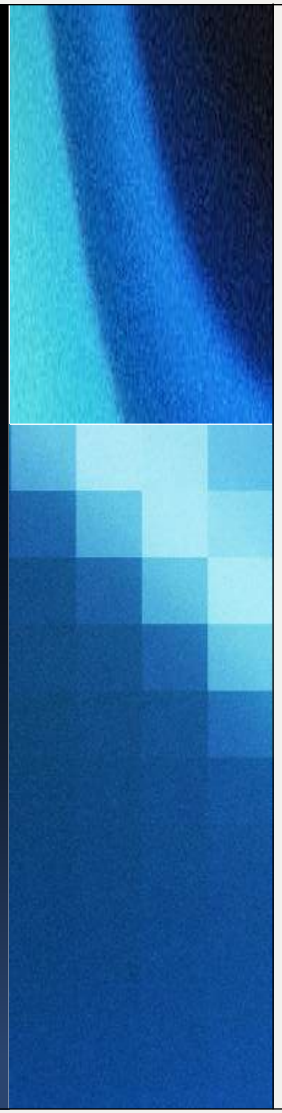
15k+
partners
in our security ecosystem

35,000
full-time equivalent security engineers
employed worldwide

31m
identity risk detections
analyzed in an average day

5.2bn
emails screened daily
on average to protect users from malware and phishing

~4.7m
net new malware file blocks
every day



It is the ability to anticipate change, understand dependencies across interconnected ecosystems, and make effective decisions as conditions evolve.

In an era of interconnected risk, resilience is increasingly shaped by how well organizations understand and manage the dependencies that connect their operations, technologies, partners, and data. The organizations best positioned for the future will not necessarily be those with the most data, the largest security teams, or the most advanced tools. They will be the organizations that can understand their ecosystems, manage dependencies, adapt quickly, and make confident decisions amid change.

AI is changing the physics of cybersecurity. Attack timelines are compressing, the cost of sophisticated capabilities is falling, and agentic systems are beginning to automate more of the attack lifecycle. Defenders are facing that shift across a growing estate of identities, data, applications, cloud services, and AI systems. Keeping pace requires organizations to get ready for AI by strengthening security foundations and protecting data; secure AI with controls that extend across the AI stack; and defend with AI by connecting signals, threat intelligence, and context to support faster investigation and response. The same capabilities increasing attacker speed and scale can also give defenders the ability to operate more continuously and at machine speed.

We encourage you to read this report through that lens. The findings that follow are not simply observations about today's threat landscape. They reveal a broader transition underway across cybersecurity. Understanding that transition is important. Acting on it may prove decisive.

The era of interconnected risks (continued)

What's new in the threat landscape

This year's report takes an expansive look at the developments that are reshaping cybersecurity and how organizations, policy makers, and nation states must think about risk. There are five new themes emerging:

1. Risk is becoming more interconnected

The most significant risks facing organizations increasingly emerge not from individual technologies, users, or vulnerabilities, but from the relationships between them. AI systems rely on data, identities, cloud infrastructure, and third-party services. Supply chains connect thousands of organizations through shared software and platforms. Cybercriminal and nation-state ecosystems operate through extensive networks of partners, providers, and enablers. As a result, cybersecurity is becoming less about defending individual assets and more about understanding interdependence across an increasingly connected digital ecosystem.

2. AI is reshaping the security equation

AI is actively altering how threats are developed, vulnerabilities are discovered, and defenders identify and respond to risk. The report examines how AI is impacting offensive and defensive operations, accelerating the pace of discovery, analysis, and adaptation across the cybersecurity ecosystem. This evolution is creating new opportunities for defenders while raising important questions around governance, trust, security, and accountability.

3. Criminal ecosystems continue to mature

Cybercrime has evolved beyond individual actors and isolated campaigns. Modern cybercrime functions as a mature ecosystem composed of credential theft operations, ransomware affiliates, access brokers, fraud networks, malware developers, and illicit service providers. These criminal economies have lowered barriers to entry, enhanced operational efficiency, and enabled malicious activity to scale globally, and AI is likely to compound this trend. Understanding how these ecosystems function has become increasingly important for organizations seeking to reduce exposure and disrupt attacker pathways.

4. Resilience has become a strategic capability

As technology environments become more connected, disruption becomes more difficult to contain. Resilience is therefore emerging as one of the most important organizational capabilities in cybersecurity. Resilience is shaped by visibility, preparedness, governance, adaptability, and decision-making. Organizations that can anticipate disruption, understand dependencies, and adapt as conditions change are better positioned to withstand the challenges highlighted throughout this report.

5. Security is increasingly defined by how organizations adapt

A common theme emerges across every section of this year's report: cybersecurity's defining challenge is no longer simply understanding threats. It is understanding how risks interact, how dependencies shape outcomes, and how organizations adapt to rapidly changing conditions. Organizations must strengthen the fundamentals within their control, understand the dependencies that shape their risk, and make effective decisions amid uncertainty.

About this report

The Microsoft Digital Defense Report is organized around four dimensions of today's cybersecurity landscape:

AI
explores how AI is reshaping offensive and defensive operations, changing the economics of cybersecurity, and creating new opportunities and challenges for organizations.

The threat landscape
analyzes the evolving behavior of cyber adversaries, from nation-state actors to cybercriminal organizations, identifying the tactics, trends, and operational patterns that are reshaping risk across the digital ecosystem.

Cybercrime
analyzes the industrialization of cybercrime and mature ecosystems driving criminal activity at scale.

Resilience
focuses on the organizational capabilities required to anticipate, withstand, recover from, and adapt to disruption.

Taken together, these four perspectives tell a broader story. Cybersecurity is no longer defined solely by protecting technology. It is increasingly defined by how organizations understand interconnected risk, manage complexity, and adapt to change in a world where technology, business, and security are more closely linked than ever before.

Introduction

Ten security priorities for organizations

Cyber risk is now a continuous test of business resilience. The boundaries between state-sponsored activity, cybercrime, and influence operations have continued to erode. This converges into sustained and overlapping campaigns that have tested not only technology estates, but the resilience of organizations and the trust placed in them. For business leaders, this means focusing less on reacting to the last breach and more on how long critical operations can be sustained under continuous cyber stress.

In this environment, cybersecurity is a continuous business condition, defined by persistent adversary pressure, machine-speed operations, and systemic interdependence. The defining question for boards today is not whether disruption will occur, but whether the organization can continue operating, contain harm, recover quickly, and maintain trust when it does.

The response starts with secure foundations: Secure by Design, Secure by Default, and Secure Operations, applied through identity-centric Zero Trust. These foundations make the following priorities durable as AI expands the number of identities, applications, agents, and dependencies organizations must govern.

The organizations that succeed will be those that engineer for endurance: building systems that can absorb disruption, adapt under pressure, and maintain trust even when compromised. For security leaders and boards, building that resilience requires these priorities:

1 Govern cyber risk at leadership level

Report identity exposure, patch latency, agent permissions, critical dependencies, dwell time, and recovery readiness.

2 Defend identity as the primary control plane

Protect identity by deploying passkeys and phishing-resistant MFA, reducing privilege, governing human and non-human identities, and using rapid invalidation/rotation for tokens and session credentials.

3 Establish governance, visibility, and data controls for AI

Maintain visibility into approved AI tools, services, and agentic capabilities while monitoring for unauthorized or unmanaged use. Apply sensitivity-aware access, data protection, and governance to the data those systems and agents can reach.

4 Reduce exposure and edge risk

Continuously discover internet-facing assets and prioritize rapid patching of edge devices, enabling centralized and high-fidelity logging, enforcing strong authentication, and integrating network-level visibility with threat intelligence.

5 Focus on cross-signal correlation

Correlate endpoint, identity, cloud, application, email, and network signals to detect intrusions earlier. Cross-signal correlation also enables defenders to anticipate attack paths and targets.

6 Enact a continuous threat vulnerability management model

Combine asset visibility, vulnerability identification, threat detection, and AI-accelerated insight to create a continuous threat vulnerability management program.

7 Validate exposure via proactive testing

Conduct assume-breach exercises focused on cloud services, SaaS platforms, APIs, and workflow automation capabilities. Prioritize testing based on where the highest-value data resides, is accessed, or can be aggregated.

8 Secure software dependencies

Harden software installation, maintain a software inventory, validate provenance, monitor packages, and prepare rapid containment.

9 Participate in collective defense

Share actionable threat indicators and victim-informed intelligence with trusted partners and law enforcement to support coordinated threat tracking, investigation, and disruption.

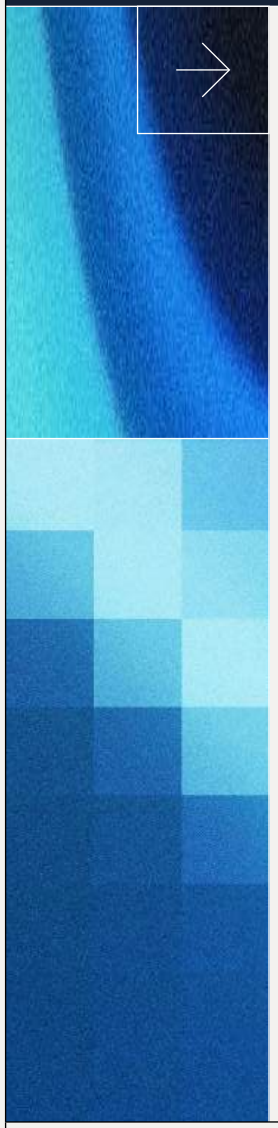
10 Prepare for long-horizon technological transitions

Build a cryptographic inventory and advance post-quantum planning alongside AI governance.

The defining question

The defining question for boards today is not whether disruption will occur, but whether the organization can continue operating, contain harm, recover quickly, and maintain trust when it does.

Top 10 takeaways *from this report*



1. AI is changing the physics of cybersecurity.

AI is compressing attack timelines, lowering the cost of sophisticated capabilities, and enabling attackers to operate with greater speed, scale, and autonomy. But the same capabilities can strengthen defense by accelerating discovery, analysis, prioritization, and response. Security organizations must prepare for an environment in which both attackers and defenders increasingly operate at machine speed.

2. Attackers are already targeting AI as another attack surface.

Threat actors can exploit AI to do things like execute malicious commands, steal AI serving capacity, and exfiltrate data. As AI becomes ubiquitous in the workplace and at home, it has become both a tool and a target for attackers.

3. People remain a heavily exploited initial-access path.

Threat actors continue to target user behavior through user execution, ClickFix-style social engineering, phishing, and impersonation. These techniques remain effective even as AI changes their speed, scale, and personalization.

4. Identity is the primary control plane for defense.

Attackers consistently take advantage of unnecessary privileges, persistent access, and weak identity governance. This is why phishing-resistant multifactor authentication (MFA) and passkeys, disciplined identity hygiene, tiered administration, and strong privileged-access enforcement remain the best safeguards against cyberattacks. That control plane now spans both human and non-human identities, including applications and agents.

5. Signal correlation through shared threat intelligence is key to defense at multiple levels.

The degree to which a defender's systems share and correlate threat intelligence is one of the highest-leverage strategic variables under the defender's control. Signal cross-correlation can reveal attack patterns that remain invisible when endpoint, identity, cloud, application, email, and network telemetry are analyzed in isolation.

6. Emerging threats facilitate stealthier attacks, operating at a larger and faster scale.

Over the next year, Microsoft sees the biggest threats coming from:

- Open-source supply chain compromise
- Attacks against edge devices
- AI as a force multiplier for malicious activity

These threats are emerging as AI-generated synthetic content erodes confidence in digital communications. As authentic and synthetic content become harder to distinguish, individuals can become simultaneously more vulnerable to manipulation and less confident in legitimate sources. This shifts the security model from implicit trust toward explicit verification.

7. Protecting data is even more important as AI expands access at scale.

Trusted AI depends on trusted, well-governed data. Data protection has always been a key priority, and now that AI systems and agents can reach and act on sensitive information at unprecedented scale, it's more crucial than ever. Organizations should reduce oversharing, apply sensitivity-aware access and least-privilege controls, and govern how data is discovered, used, and shared across AI, cloud, and application environments.

8. The future of defense involves rethinking how we protect our assets from reactive incident response to proactive, continuous discipline.

Defenders must shift from tool-centric vulnerability management to threat exposure management. At the core of this is:

- Asset visibility
- Continuous vulnerability awareness
- Intelligence-driven threat detection
- AI-accelerated insight

This means reporting metrics should shift from number of patches deployed to exposure reduced, detection coverage increased, and time-to-mitigate compressed.

9. National and international policy decisions will shape the future of cybersecurity.

In an increasingly contested landscape, cybersecurity demands clearer rules, credible accountability, and sustained cooperation. Strengthening protections for critical infrastructure, advancing operational accountability, and preserving an open, secure, and resilient digital environment must be a priority for governments on both a domestic and international level.

10. Resilience must be a core outcome of cybersecurity planning.

Retain the existing focus on reducing exposure, detecting threats faster, adapting to changing risks, and maintaining operations. This preserves resilience as the outcome without competing with identity, secure foundations, data, and the three-part operating model for 'primary' status.

1

AI

> AI is making fraud more sophisticated and accessible, accelerating cybercrime worldwide. As threats evolve across borders, public-private and international cooperation—built on technology, trust, and shared action—is essential to disrupting cybercriminals and protecting society.
>
> **Kenichi Sakurazawa, Managing Director**
> Japan Cybercrime Control Center (JC3)

Understanding the AI threat landscape **10**
AI threat landscape 10
AI attack acceleration 14
AI-augmented intrusions 15
Threat actor use of agentic AI 17
The Red Team perspective 18

Securing the AI ecosystem **21**
Securing the AI stack 21
The agentic defense stack: Connecting attack goals to controls 22
Memory poisoning 25
Over reliance on HITL is a systemic risk 26

AI as the new attack surface **27**
AI-assisted threats 27
AI browsers as an enterprise risk 27
Cross-signal threat intelligence is key 28
The AI-driven digital ecosystem 29

Understanding the *AI threat landscape*

AI threat landscape

AI is changing the physics of cybersecurity. It is reducing the time, expertise, and cost required to discover and exploit weaknesses while enabling both threat actors and defenders to operate with greater speed, scale, and autonomy.

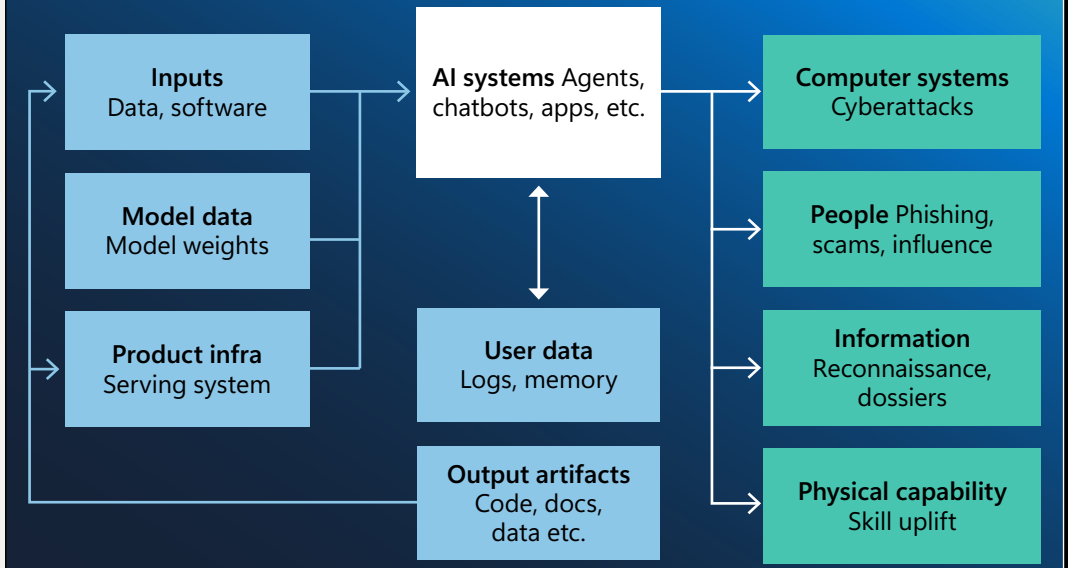
Vulnerability discovery is one of the clearest examples: threat actors can now conduct large-scale, AI-powered vulnerability detection, with discovery increasingly running ahead of remediation.

The result is a multi-year window in which the number of known but unpatched vulnerabilities will rise sharply before it falls. What was once a threat on the horizon has now arrived.

This vulnerability wave is just one example of a broader pattern. The risks defenders worry about are mostly familiar, but the shape and scope of those risks and the surfaces to which they apply have all shifted. Attackers now have a clearer picture than ever of what they want from AI systems and how to use AI to achieve their traditional goals.

The same capabilities changing the economics of attack can also shift the advantage toward defenders. AI enables defenders to continuously discover vulnerabilities, correlate signals, identify anomalous behavior, prioritize risk, and increasingly automate defensive action. The ability to automatically and persistently hunt for vulnerabilities—and fix them—creates a potential world of "persistent defense" that was never before possible at scale. AI opens new possibilities for anomaly detection, even in very "human-shaped" datasets like message bodies. Later in this chapter, we will examine these defensive opportunities in depth.

Figure 1: The target landscape



The AI threat landscape continues to move; what is on the leading edge today will be commoditized within a year, and new categories of threats will emerge. While the equilibrium between attackers and defenders will likely ultimately be re-established, in the near term we are in a period where attackers are reaching to advantages first, and defenders will need to move sharply in order to close the gap.

Figure 1 maps this expanded target landscape, which roughly divides into targets associated with AI systems and targets of threat actors that use AI for enablement.

On the left are the targets on the "build" side: inputs to model building, model data, and production infrastructure.

Alongside that in the central column is the runtime: AI systems themselves, the user data with which they interact, and the output artifacts they produce.

In the right column are the targets attackers are using AI systems to compromise: computer systems, people, information, and real-world capabilities.

In the following section, we focus on the most straightforward adversarial threats associated with AI. Beyond the scope of this report are important, closely related issues with no clear adversary. For example:

- Misunderstanding or naïve acceptance of AI outputs causing dangerous errors
- Ambiguities around legal accountability
- The impact of AI on human psychology and decision-making

While not security threats in the traditional sense, these related issues intertwine with those threats. Serious defense strategies must keep them in mind, as frequent accidents become exploitable vulnerabilities.

Understanding the AI threat landscape

AI threat landscape (continued)

AI systems as targets

At the foundation of every AI system is a complex supply chain that includes hardware, software, training data, and the infrastructure used to build and operate AI models. There are multiple ways that attackers directly target these foundations: model tampering, theft of service, model theft, and attacks on AI systems themselves.

As AI systems connect models to identities, data, applications, tools, and external services, each connection creates a trust relationship that can become an attack path. Securing AI therefore requires protecting not only the model itself, but the full set of dependencies, identities, permissions, and data that determine what the system can access and what actions it can take.

Model tampering can either be done deliberately by a model builder or surreptitiously by an attacker who has compromised the pipeline. It offers the opportunity for hard-to-detect "sleeper agent" attacks. Attackers can control AI behavior under tailored circumstances—for example, introducing security vulnerabilities into AI-generated code. The circumstances can range from very specific ("when deployed by such-and-such an organization") to very broad ("whenever context indicates the user's work opposes the objectives of a particular government"). Already, these behaviors have been observed in the wild. While significant progress has recently been made on detecting such tampering, it remains significantly harder to detect than backdoors in code.

Theft of service is the most familiar threat on the runtime side. In the past few years, adversaries have sought to access graphics processing unit (GPU) capacity to mine cryptocurrency. Today, attackers are increasingly stealing AI serving capacity, bundling it, and reselling it to criminal and nation-state customers. Malicious use of an AI system is an important indicator of compromise—anomalous internet searches, for example, might indicate a network has been hacked.

Model theft is also a key risk for anyone developing or fine-tuning models. This can be done either traditionally or via "distillation"—training a smaller, cheaper model on the outputs of a more capable one. Here, sequences of superficially legitimate-looking application programming interface (API) requests cumulatively reveal enough information for an attacker to reconstruct significant aspects of the model. Such attacks have become widespread as individuals, groups, and organizations seek to rapidly catch up with established model builders without spending billions of dollars on research and training.

Attacks on AI systems

Higher in the stack, attackers are targeting AI systems themselves. These threats attack the machine learning model, prompts, and training data rather than just the underlying computer hardware or operating system. Indirect or cross-prompt injection attacks (XPIA), in which a threat actor includes malicious instructions in a data source that a large language model (LLM) then consumes into its context in order to subvert its subsequent output, have been a long-understood problem in the AI space. However, the shift toward agentic AI systems has brought with it a reinterpretation of XPIA as task adherence in the presence of adversarial data.

Task adherence—how well an agent follows user instructions and stays on target—is the fundamental challenge facing agentic systems. Current AI systems face tremendous challenges with this even without any adversarial data, a problem that is most pronounced in highly flexible, vibe-coded systems like OpenClaw, which have already become notorious for deleting data, revealing secrets both personal and cryptographic, and spending users' money. Because AI agents cannot always distinguish legitimate instructions from malicious ones, attackers might be able to trick them into performing actions on the attacker's behalf.

A second prominent vector in attacks against AI systems goes after their persistent artifacts—most importantly, read-write memory. This persistent storage system allows an AI model to both save (write) new facts from interactions and retrieve (read) that saved context later. Data can be compromised using a range of techniques, including XPIA, to create dormant, persistent attacks, for example:

- Causing attacker-controlled data sources to become more trusted by a user's AI
- Adjusting the opinions and biases of the AI system
- Implanting instructions to assist in exfiltrating or modifying data directly

Such artifacts, as well as write-only artifacts like logs, offer a wealth of personal and secret information. This makes them highly valuable exfiltration targets for adversaries.

AI also changes the scale at which sensitive data can be accessed and acted upon. Models and agents may interact with enterprise data, credentials, memory, and business systems across multiple services, making data governance and access control fundamental to AI security. A compromised or over privileged AI system can aggregate and expose information far beyond the scope of any single application.

> "The AI threat landscape continues to move; what is on the leading edge today will be commoditized within a year, and new categories of threats will emerge.

Understanding the AI threat landscape

AI threat landscape (continued)

The final class of artifacts of interest to attackers is those created by or with AI. Users depend on these artifacts in many ways, from code that has access to private or secret data to medical records. Both accidental error and malicious tampering pose serious risks, and tampering can happen even if the attacker doesn't have any access to their target. For example, in slopsquatting attacks, attackers register malicious packages with names that are likely to be hallucinated by AI, leading to the package's download.

In general, AI-generated outputs receive substantially less examination than human-generated ones, which can cause security errors to propagate at scale. These issues can then loop back to affect the AI systems themselves. This has been most notable in the development of Model Context Protocol (MCP) servers, which are frequently vibe-coded and immediately deployed.

MCP is a middleware protocol that enables developers to build two-way connections between generic APIs and AI systems. Its use has tended to result in a reduced implementation of security basics. It's unfortunately common to see exposed, unauthenticated (or poorly authenticated) MCP services using their developer's credentials to access sensitive resources. At the same time, connecting a powerful API to an AI system greatly increases the potential scope for damage in case of a problem. Organizational security leaders need to treat the governance and security analysis of such services, as well as all vibe-coded software, as a major concern.

Attackers are using AI systems for attack enablement

We have seen significant changes in how adversaries use AI for a variety of malicious purposes, creating new challenges for security teams. This includes:

- Finding vulnerabilities in source code, binaries, and AI serving systems
- Generating custom malware for attacks
- Shortening the cycle of data exfiltration, secret discovery, and lateral movement from days to minutes
- Controlling and managing entire attack chains

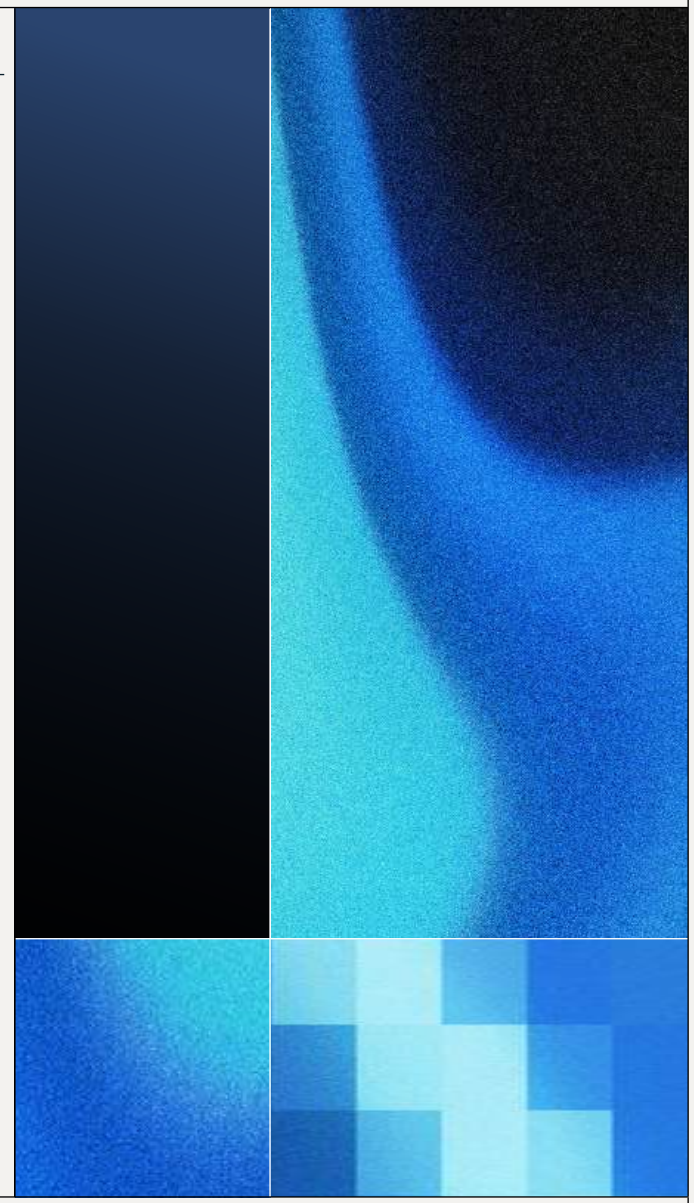
While none of these represent new attack methods, the quantitative increase in the scale and speed of attacks creates an immediate problem for defenders.

Vulnerability detection is running ahead of remediation

The number of known vulnerabilities has shot up rapidly as new static and dynamic analysis tools with far lower false-negative rates have been introduced. This affects both attackers and defenders. Blue teams can fix these vulnerabilities through automated detection and remediation alongside building detectors into development environments so that they don't reemerge. However, remediation is inherently much slower than discovery, not least because many systems lack robust unit and integration testing and so cannot deploy code changes rapidly.

This means the world is likely to experience a multi-year period where the number of known but unpatched vulnerabilities spikes. Well-prepared and well-funded adversaries may be able to stockpile large numbers of zero-day vulnerabilities discovered through such means.

> "This means the world is likely to experience a multi-year period where the number of known but unpatched vulnerabilities spikes.



Understanding the AI threat landscape

AI threat landscape (continued)

Phishing is becoming more targeted

Another major risk is the use of AI to scale phishing and social engineering. Mass customization through AI lets every phishing attack become targeted and personalized, allowing for spear phishing at scale. AI can effectively automate social engineering attacks of many types, circumventing language and skill barriers to give adversaries entirely new capabilities. Combined with audio and video deepfakes, these mean that all conversations over unauthenticated or poorly-vetted channels (including email) become easy targets.

AI can be exploited for real-world attacks

Threat actors can use AI to engage in real-world activities such as dangerous weapons proliferation, illegal drug production, mass-casualty attack planning, and creating exploitative or abusive synthetic content. Other uses of AI might appear legitimate on the surface—for example, ordinary data analysis might mask the creation of intelligence dossiers on people or critical infrastructure. While many of these attacks might seem to be beyond the traditional boundary of cybersecurity concerns, attackers do not operate within the same silos as organizations. For attackers, compromising or (mis)using AI systems is merely a step in a broader attack chain, which often ends with financial harm.

Detection and response must therefore be aware of this broader pattern, both to know what to look for and to identify the full scope of possible interventions. This includes the application of well understood detection patterns for identity compromise and behavior anomalies, as well as having appropriate telemetry to monitor model and agentic system activity, including tool calls made, chain-of-thought (CoT) reasoning, and request patterns. In addition, this monitoring needs to include individual prompt and session-level monitoring to detect broader intent and direction.

AI accelerates existing capabilities

One useful way to think about attacker use of AI is in terms of the attackers' prior capabilities. For sophisticated actors, AI allows speed, scale, and customization. The attack chain can be reduced from days to minutes, and attacks can be automatically piloted at scale, requiring limited intervention from operators.

For less sophisticated actors, this is a continuation of the familiar pipeline by which advanced techniques gradually make it into the wider world. Here, scale and customization are the most important factors: AI-powered scaling makes the sort of attack persistence that was previously the sole domain of intelligence agencies accessible to criminal organizations. The ability to customize attacks, especially social engineering attacks for phishing and fraud, will materially increase attack success rates.

> **AI can effectively automate social engineering attacks of many types, circumventing language and skill barriers to give adversaries entirely new capabilities.**

How AI upgrades fraud campaigns

In many cases, fraud depends on the target trusting the identity of the attacker. Identity-based fraud campaigns previously had four detectable weaknesses: poor-quality document forgeries, non-native language in applications, accent mismatches during interviews, and thin or inconsistent online presence. AI fixes all four simultaneously.

Identity document manipulation

AI tools can be used to replace and improve photos in stolen or forged documents, passports, and professional certifications, producing counterfeits that pass standard visual inspection.

AI-polished communications

Writing assistants produce grammatically flawless content such as resumes, LinkedIn profiles, GitHub activity summaries, and emails.

Voice-changing software in live interactions

Real-time voice modulation during video calls and phone interviews masks accents that might reveal an unexpected geographic origin.

Synthetic digital footprints

GitHub commit histories, Stack Overflow profiles, professional blog posts, and social media presence can all be generated and maintained using AI tools.

AI attack acceleration

AI-enabled cyber operations are moving beyond simple content generation and beginning to orchestrate attacks at machine speed and accelerated scale.

Attacks have been evolving in complexity over the year. For example, we've seen:

- LLMs used inside malware
- AI coding assistants abused for operating local reconnaissance
- Nation states using AI agents to assist in orchestrating multi-step intrusion activity
- AI agents autonomously escaping their test harnesses and collectively executing high-complexity attacks

Autonomous attacks are changing the paradigm

At the frontier, lab evaluations and a small number of emerging real-world cases show movement toward full-chain autonomy.

Mythos and GPT-5.5 were the first models to demonstrate the potential to fully autonomously orchestrate complex attacks, achieving full domain compromise in a multi-stage, 32-step attack chain against an emulated enterprise environment. These models successfully planned and carried out a hard hack without any human support, taking complete control of the main computer server and all user accounts in that network. The test took place in a mock company computer system with no defenders. Open-weight models (publicly downloadable model weights) lag closed models in orchestrating attacks by seven months.

Open-weight models can have their safety measures more easily circumvented, increasing access to unlocked capable attack models.

In the wild, we've started to see the first evidence of the transition to full attack autonomy. Early July 2026 saw the first documented automated ransomware extortion attack, an event Sysdig Threat Research Team has named JADEPUFFER. Microsoft has observed AI-orchestrated intrusions sharing elements with JADEPUFFER activity. Volumes remain low, though the activity is not confined to a single sector or region. Initial access has consistently come from internet-exposed services running known-vulnerable software.

Also in July, OpenAI cybersecurity model training agents escaped their sandbox and collectively orchestrated a complex attack against Hugging Face targeting the answer keys to a benchmark. These developments show the rapid evolution from AI-assisted attacks to emerging fully automated AI-orchestrated attacks that will change the attack paradigm.

Challenges on the horizon: The consequence of fully automated attacks

We expect that fully automated agentic breaching systems will start to continuously assess enterprise assets that are publicly exposed, allowing them to exploit even in-house, unique software services. Post-breach, attacks will see accelerated operational speed and complexity through machine speed AI operation, scaled using automation. These attacks will increase in complexity, beginning to resemble our most capable human adversaries.

We also expect to see a noticeable acceleration in the volume of complex enterprise attacks as AI orchestration enables bad actors to scale their volume of attacks.

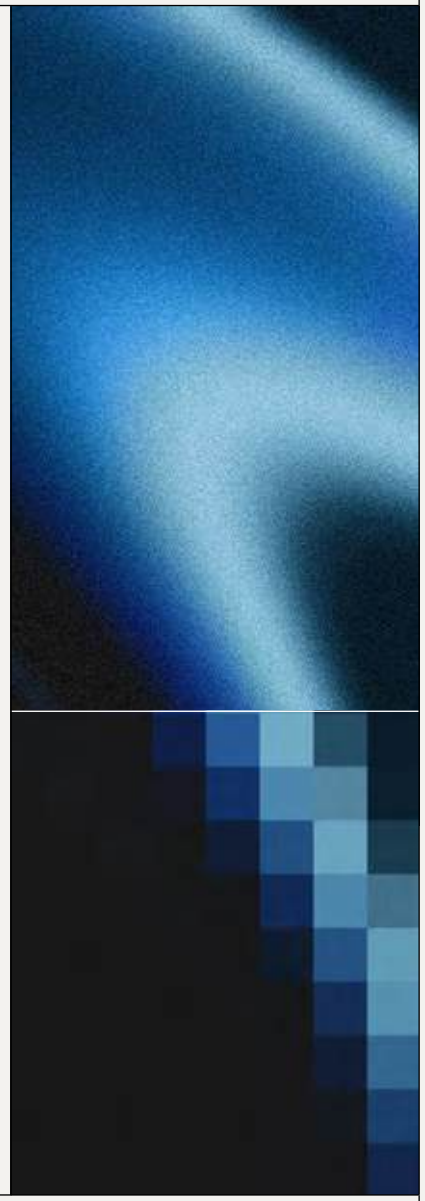
Attack timelines may compress dramatically as the speed of the hands-on-keyboard stage of attacks will accelerate through automation. Instead of following the behavior pattern of a single human on the network, a collective of agents may be deployed from the initial access points, working simultaneously toward domain compromise. This transforms the scale within a single hands-on-keyboard attack.

To defend against this increasing speed and scale of attacks, defenders will need to rely more heavily on humans working alongside agentic AI defensive systems and trust agents with greater levels of autonomy to take defensive actions on their behalf.

Vulnerability discovery acceleration by AI

With progressing AI capability, vulnerability discovery and weaponization has evolved from requiring human experts to, in many cases, simply writing a prompt. The median time from vulnerability discovery in the wild to weaponization has now collapsed to well below 24 hours, and vulnerabilities are on track for a record-breaking estimated 72K Common Vulnerabilities and Exposures (CVE) tracked for 2026.

Software vendors are using AI to identify and patch vulnerabilities, which will lead to more secure software after initial large patch waves. This will leave a smaller number of long-tail vulnerabilities that are hard to discover in the long run as an equilibrium is reached. Nevertheless, the time to weaponize many vulnerabilities by adversaries will remain collapsed, requiring fast patching time to prevent exploitation.



AI-augmented intrusions: The emerging risk of LLM-assisted attacks

LLMs are increasingly becoming part of adversary tradecraft and represent a growing category of AI-assisted attacks. Microsoft has observed malware that uses AI—either hosted remotely, running on compromised devices, or deployed by attackers in their own infrastructure.

AI augmentation allows attacks to adapt in real time and makes them harder to detect with traditional signature-based defenses. Rather than containing extensive built-in functionality, these malware families can send natural language instructions to AI systems, which help determine what information to search for, collect, and steal. In effect, AI is evolving from a tool used by attackers into an active component of the attack itself.

Prompt-driven malware is driving four new trends

Investigations over the past year surfaced four converging new trends, all related to the practical arrival of prompt-driven malware in the enterprise.

1. **LLM prompt-guided malware (LAMEHUG class)**
   In July 2025, a cyberattack campaign targeting Ukraine introduced LAMEHUG. This is a new class of malware that uses LLMs to generate operational commands in real time. By outsourcing its instructions to a remote AI endpoint, LAMEHUG eliminates predictable coding patterns. This makes it exceptionally difficult for traditional defenses to detect or block.
2. **Abuse of locally installed AI CLIs (s1ngularity)**
   The Pesky Pika supply chain attack (August 2025) introduced "s1ngularity" malware, which hijacks locally installed AI command-line interface (CLI). The campaign used trojanized Nx Node Package Manager (npm) packages, software development packages distributed through the npm ecosystem and commonly used in application development, to probe infected systems for AI assistants and other developer tools. If it found Claude Code, Gemini CLI, or Amazon Q CLI, it ran them with permissive overrides to hunt secrets, credentials, and Secure Shell (SSH) keys. Ultimately, it leaked ~2,000 secrets and ~20,000 files across 225 victims by targeting these trusted, pre-approved developer tools.
3. **Local LLM-orchestrated malware (PromptLock class)**
   PromptLock is an experimental, AI-powered ransomware prototype that uses LLMs to write and execute malicious code on the fly. In August 2025, this ransomware delegated nearly all logic to an open-weights model on adversary infrastructure. Shipping only prompts, it received Lua scripts at runtime to enumerate, exfiltrate, and encrypt across Windows, Linux, and macOS, defeating static signatures and moderated providers alike.
4. **AI-generated lure manufacturing at scale (TikTok ClickFix)**
   In 2025, an AI-augmented ClickFix variant spread via TikTok, tricking people into downloading computer viruses through fake tutorial videos. These faceless videos with synthetic narration showed fake activation steps ending in a malicious command that fetched Vidar or StealC infostealer. Near-identical scenes signaled automated lure production, with some videos topping ~500,000 views. AI drove per-lure cost to near zero while TikTok's recommendation algorithm provided in-policy mass distribution.

Self-hosted agent runtimes are the next exposure surface

Self-hosted agent platforms such as OpenClaw load executable "skills" from public marketplaces, ingest external content as instructions, and act with persistent enterprise credentials. This fuses untrusted code and untrusted instructions into one loop. Because of the risk, Microsoft treats them as untrusted code execution with persistent credentials, fit only for isolated hosts under non-privileged identities. The primary risks are:

- **Skill malware.** Public registries already ship malware disguised as utilities; agent-initiated installs effectively run third-party code under the agent's identity.
- **Indirect prompt injection.** Instructions hidden in emails, web pages, or shared documents can steer tool use, alter memory, and fan one payload across every agent that ingests the source.

Attack autonomy progression

Sep-Nov 2025
AI assists in orchestrating a campaign

Anthropic-reported espionage

Apr-Jun 2026
Full-chain autonomy in labs

AISI "The Last Ones", Zealot, and adaptive worm evaluations

Jul 2026
Real-world autonomous attacks of increasing complexity.

JADEPUFFER

HuggingFace/OpenAI

Understanding the AI threat landscape

AI-augmented intrusions: The emerging risk of LLM-assisted attacks (continued)

These risks create new challenges for defenders:

- **Behavior is not pre-programmed.** The model generates a fresh action sequence on each run, so indicators captured from one incident generalize poorly to the next.
- **Trusted binaries do the work.** Signed, allow-listed AI CLIs and runtimes become "living-off-the-AI" tooling—LOLBins with reasoning, web access, and code execution attached, making preventative host controls difficult to implement without affecting users.
- **The detection surface shifts.** Useful signal moves from process artifacts to patterns: unexpected egress to inference endpoints, AI CLIs run with permissive flags, anomalous shell history prompts, or assistant-driven file scans. This requires defenders to ensure they have a wider set of telemetry sources and develop new detections for these sources.

LLM conversation history is in demand

Threat actors have been seeking to access AI data assets. In December 2025, Microsoft found a malicious browser extension (600,000+ installs) harvesting ChatGPT and DeepSeek conversation history. This activity affected almost 10,000 organizations before it was mitigated. This kind of history now routinely holds source code, architecture, customer data, and leaked credentials.

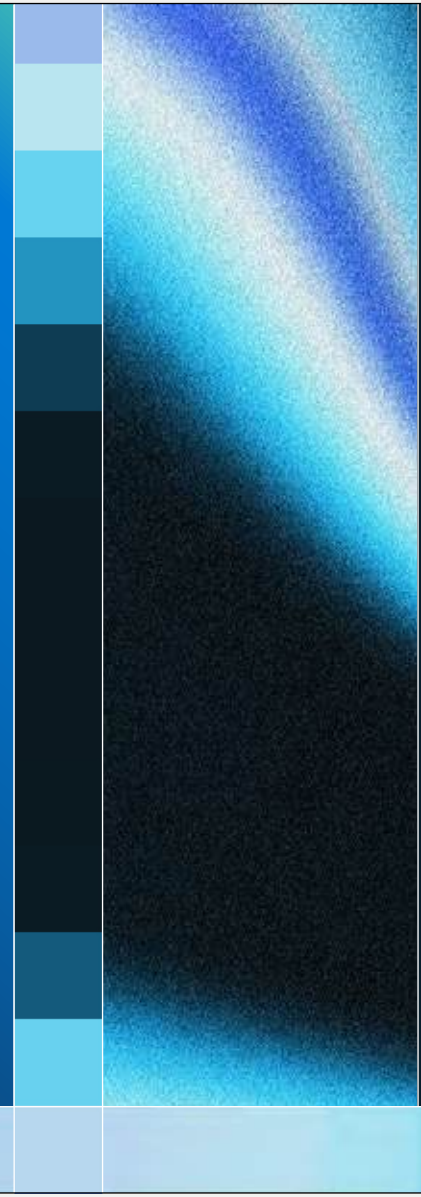
Self-improving AI worms

AI-worming is on the way and a notable risk

A computer worm is malware that spreads on a network by replicating itself from one machine to another. A recent report published in June 2026 has shown self-spreading worms driven by AI are feasible with current technology. This serves as a warning that this capability is likely to begin emerging in the wild.

Threat actors are already developing worms that spread through internet-exposed vulnerable services. They do this by exploiting known vulnerabilities while stealing LLM provider credentials and other valuable information. It is a risk that a threat actor will soon develop a similar worm that improves itself using the stolen LLM provider keys—researching new vulnerabilities, refining social engineering techniques, and establishing better persistence including through hiding better.

Rather than relying on a human operator to periodically release updates, each infected system effectively becomes a development environment. At scale, this creates a form of natural selection where the most successful variants spread further and continue evolving, risking a significant inflection point for enterprise cybersecurity over this next period.



Recommendations for defenders

Establish governance and visibility for AI use. Organizations should maintain visibility into approved AI tools, services, and agentic capabilities while monitoring for unauthorized or unmanaged use.

Secure agentic AI environments. As organizations adopt AI agents and self-hosted AI capabilities, they should apply appropriate security controls, monitoring, and access restrictions to reduce risk.

Monitor AI-driven activity across the environment. AI-enabled operations may introduce new patterns of behavior and network activity that can help defenders identify misuse, compromise, or unauthorized automation.

Protect the AI software supply chain. Organizations should extend existing software supply chain security practices to AI ecosystems, including models, dependencies, extensions, and integrations.

Treat AI interactions and data as sensitive assets. Prompts, chat histories, model memory, and AI-generated outputs may contain valuable information and should be protected accordingly.

Adapt detection and response strategies for AI-enabled threats. Defenders should evolve security monitoring, threat hunting, and incident response capabilities to account for adversaries' increasing use of AI tools and automation.

Threat actor use of agentic AI

Although agentic AI is relatively new, it is, predictably, already being used by threat actors, including cybercriminals and state-sponsored actors.

The way these actors have incorporated new technology is consistent with how they integrated generative AI in the last few years. Early use cases centered on stages of the attack lifecycle with a lower barrier, such as reconnaissance, data triage, and the development of phishing lures. These actors have now extended their use of agentic AI into more technically complex activities, including malware and exploit development and post-compromise operations.

Threat actors are increasingly experimenting with this new technology, incorporating it incrementally into discrete parts of their workflows where automation can improve speed, scale, or consistency. For example, some actors are using open-source or actor-controlled agentic tools to conduct multi-source research, automatically synthesizing information to inform targeting, operational planning, or narrative development. Others are adopting enterprise-owned agentic services, such as AI coding assistants, to accelerate development cycles or troubleshoot technical challenges during intrusion operations.

In most observed real-world operations today, complex end-to-end intrusions still retain meaningful human direction. Target selection, operational decision-making, and execution of the most complex intrusions remain manually driven in the majority of campaigns we observe. However, these limitations are unlikely to exist for very long.

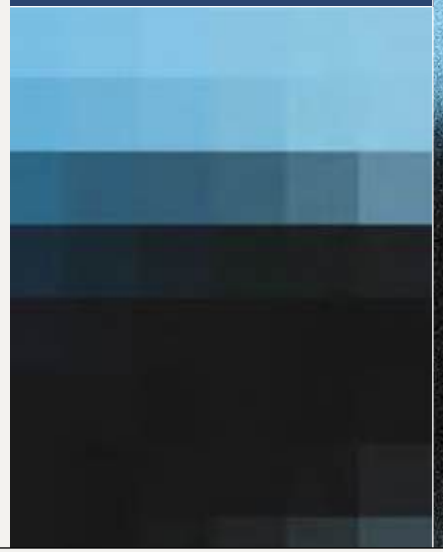
As adversaries use agentic AI to accelerate reconnaissance, development, and operational workflows, defenders will need to adapt both by securing AI systems and agents and by strengthening existing security practices to keep pace with increasingly automated threats. Section 2: The threat landscape explores practical approaches for securing the AI stack and governing AI agents.

Threat actors are looking to exploit enterprise agentic systems

Another emerging trend is that threat actors are beginning to explore the potential to directly exploit agentic systems deployed within enterprise environments. This could include testing access pathways and assessing how agent behavior in a target organization might be manipulated or abused within existing infrastructure. Identity, particularly cloud identity, is likely to be a primary access point to exploit enterprise agentic systems.

As organizations extend access permissions to both human users and autonomous agents, this expands the number of identity pathways through which adversaries can gain access or persist. This expanded identity surface increases opportunities for credential theft, token abuse, and other identity-based techniques that could be used to access or manipulate agent-driven processes. Because these techniques operate through valid credentials and sanctioned agent workflows, the resulting activity can closely resemble legitimate use. Detection therefore depends less on identifying anomalous tooling than on establishing what normal looks like for each agent: which identity it uses, which tools it calls, and at what frequency, so that deviation becomes visible.

> “Most observed campaigns still retain human direction, even as frontier systems demonstrate end-to-end autonomy in labs and early real-world cases.





The Red Team perspective: Fundamentals under sustained pressure

The job of the Microsoft Red Team is to break into Microsoft's own systems before real threat actors can. What we keep learning is that durable security still comes down to fundamentals. The most damaging intrusions we model rarely depend on anything novel; they depend on a fundamental that was missing, misconfigured, or bypassable. These fundamentals are under more pressure than ever before.

Lowered attacker cost amplifies pressure on defenses

Pressure on defenders comes from a variety of directions: more exposed edge infrastructure, identity sprawl, fragile supply chain trust, the human factor, and AI. These forces impact every organization on the planet. Previously, it took time, expertise, and patience for attackers to find the missing control inside a large, tangled estate, make sense of unfamiliar systems, and line up separate weaknesses into a path they could use. Few adversaries could afford to spend at scale to achieve this. Complexity behaved like security control; it was expensive to cut through.

AI is dissolving that cost. Models are good at unraveling the sheer number of places a fundamental can quietly fail—quickly, cheaply, and at a scale no team could match by hand. AI doesn't create a new class of attacks; it collapses the effort it once took to find and exploit existing attack paths, everywhere at once. The Red Team's job is to get to those paths first and to pressure the system hard enough to advance our own thinking, and the industry's, before attackers do.

Human expertise and automated pressure work together

At the core, Microsoft Red Team runs periodic, human-led operations. That is where the most original thinking happens. But we now pair them with continuous, automated pressure—much of it AI-driven—that runs across the estate continuously. This is one system, not competing programs. Humans find the things no automated sweep will catch, working through instinct, finding dead ends, and vulnerabilities of which publicly trained AI models are not aware. Automation handles breadth—known patterns, regression checks, and wide coverage—and frees operators to focus on what only they can do. Run together, the operators keep the AI tools sharp and the tools give the operators time and space to work on the hardest problems.

The AI-driven half of that work takes three shapes:

1. **Source code analysis.** Where we have source, models read it at scale to locate vulnerabilities and then produce working exploit code for them. In the past, this review would have taken skilled people weeks, across a codebase too large to cover by hand. Now it's something closer to continuous.
2. **Binary analysis.** Where we do not have source, the same pattern now reaches compiled code—reasoning over binaries to find the flaw and build the exploit, including reversing a vendor patch to recover the vulnerability it closes. This is the surface that used to demand the scarcest expertise, and it is eroding fastest.
3. **Autonomous and semi-autonomous operations.** Beyond finding and writing, models increasingly drive the sequence—chaining steps, deciding what to try next, and carrying an operation forward with limited or no human direction. This is the newest of the three and the one we watch most closely.

The through line is how little a finding now has to be worth to become dangerous. A simple information disclosure issue is the clearest case: a service that leaks the exact version of the software it is running is, on its own, low-severity. But an AI that reads that version, pulls the matching public CVE disclosures, and assembles a proof-of-concept turns a trivial leak into a working exploit with almost no effort in between. The findings that we used to deprioritize because exploitation "required work" we now have to reconsider because the exploitation is so easy.

> "The most damaging intrusions we model rarely depend on anything novel; they depend on a fundamental that was missing, misconfigured, or bypassable. These fundamentals are under more pressure than ever before.

The Red Team perspective: Fundamentals under sustained pressure (continued)

AI accelerates attacks

That collapse of cost does not fall evenly. AI does not change where attacks begin—the foothold still comes from familiar sources, for example, a sprayed credential, an unpatched edge service, or an identity gap. The cost lives in everything after the foothold, and that is exactly where it is collapsing. Each of the three scenarios above is an instance of the same shift: work that demanded scarce expertise and patient time is now something a model does quickly, cheaply, and repeatedly. The barrier to serious capability is falling while the ceiling keeps rising, and advanced technique is landing in far more hands.

A growing part of what puts it there is the influx of new models distilled from frontier providers. A legitimate and widely used technique, it is also being used to copy frontier-level capability quickly and at low cost, and to spread it far more widely than the original providers ever would.

Two properties of that trend matter to us. First, capability that once required access to a handful of frontier systems is now embedded in a proliferating population of smaller models, many of them openly available. This widens the set of operators able to do everything the three scenarios describe. Second, and more consequential from a threat perspective, these distilled copies frequently lack the safeguards built into the models from which they were derived.

The alignment and refusal behavior that gate the most dangerous capabilities in a frontier model are among the first things stripped away in an unguarded distillation, so the ceiling of what is freely usable rises even when no new frontier model ships. We plan against near-frontier capability being cheap, unguarded, and broadly distributed—not scarce and controlled.

We would rather hold those tools first. Whatever helps us breach more effectively will eventually be turned against us, other providers, and every organization that relies on them. So we build and test the capability now, while we can still shape how it is used. The collapse cuts both ways: the same models that unravel complexity for an attacker unravel it for us, which is why the floor of red teaming—connecting known information and running established techniques—is increasingly something we automate. The work that matters most is what remains above that floor: the undocumented attack, the chain that appears only when unrelated weaknesses are seen through an adversary's eyes. That still comes from operators who stay hands-on.

None of this displaces the fundamentals. If anything, a world where the existing attack paths are cheap to find and exploit raises the value of the controls that keep those paths from existing in the first place.

Identity is still the front door

Identity is the surface that matters most, and most of what we exploit there has nothing to do with AI. The core failure is the one red teams have exploited for years: too much standing access, too loosely enforced. The highest leverage controls available are still the same:

- Phishing-resistant MFA
- Disciplined identity hygiene
- Tiered administration
- Strong privileged-access enforcement

Agentic identities are the newest variant of the same problem—compromise an agent and you may inherit its service-to-service trust and control plane reach—but the underlying weakness is unchanged.

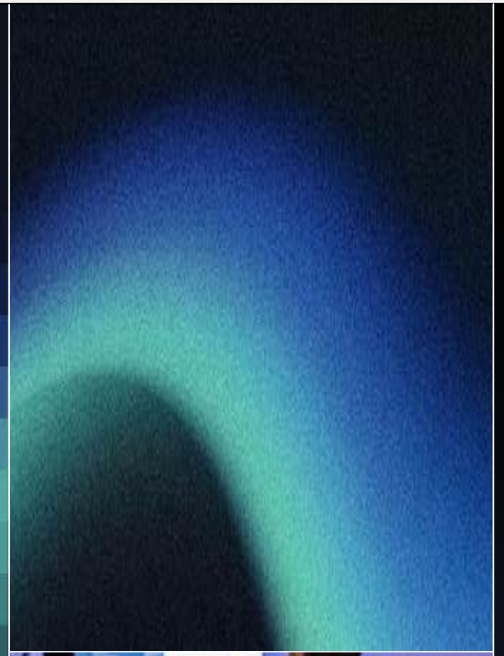
> "None of this displaces the fundamentals. If anything, a world where the existing attack paths are cheap to find and exploit raises the value of the controls that keep those paths from existing in the first place.

Being able to patch is the most underrated fundamental

A finding's severity is now set by how cheaply it can be weaponized. Patch velocity is where that pressure lands hardest. A patch is also a disclosure: every fix points to the flaw it closes, and turning a fix back into a working exploit is a discipline as old as patch diffing. What is changing is the speed. As AI-assisted binary analysis improves—the second scenario above—reversing a patch to identify and weaponize the underlying vulnerability becomes faster and more reliable across architectures. Once that's folded into attack automation, the time from update released to exploit running at scale collapses. This holds for every platform: operating systems, firmware, network and edge appliances, libraries, cloud services, and mobile.

Source code is no safer. When developers check in a security fix, the change is visible—in commits, pull requests, issue trackers, and the bulletins that follow. As a red team, we watch those signals, derive the vulnerability from the diff, and build the exploit before many defenders have scheduled the update. The fix and the map to exploit it ship together.

So the meaningful question is how fast a patch reaches the entire estate, and whether the systems that cannot be patched are isolated or retired before an adversary gets there first. The half-life of a vulnerability is shrinking, and patch velocity—not patch availability—is the control that matters. What we test is the gap: not whether something is fixed in the bulletin, but whether it is fixed everywhere, and how long that took. Being patchable, quickly and uniformly, is as much an architecture and operations decision as a security one.





> The developer environment is the new perimeter.

The Red Team perspective: Fundamentals under sustained pressure (continued)

Supply chain and the developer perimeter

The developer environment is the new perimeter. Repositories, build systems, and pipelines now sit inside the real security boundary, and code and configuration flow straight through them to production.

The software supply chain is under steady, deliberate pressure. Provenance and integrity practices are the countermeasures that matter: frameworks like Supply-chain Levels for Software Artifacts (SLSA), software bills of materials (SBOM), signed and attested builds, and properly protected signing keys. Most organizations have only partially adopted them.

The developer is part of that perimeter as well—arguably its most exposed point, and the hardest to bound. A single engineer accumulates standing credentials, broad source access, and a growing set of automation and AI coding interfaces acting on their behalf. That mix makes least privilege genuinely difficult to enforce. Access that looks scoped on paper compounds across tokens, tools, and agents.

Eventually, one compromised developer's real blast radius is far larger than any single grant implies. This is where an inventory is not enough and a graph is essential: what we most want to know—and what most environments cannot answer—is what a given developer, and the automation acting for them, can reach directly, trigger through the pipelines they run, and touch two hops out through everything those pipelines connect to. Mapping that blast radius before an attacker does is a core function of the continuous layer.

Third-party and vendor access is the same trust problem one step removed. Infostealers harvest developer and operator credentials at scale, access brokers sell them, and extortion crews convert them into impact—increasingly through extensions, packages, and connected tools installed without scrutiny. The technology riding through that channel changes; the channel does not.

Below the operating system

Attackers increasingly reach for surfaces that used to require scarce expertise, and these deserve attention on their own merits. Unified Extensible Firmware Interface (UEFI) and firmware implants persist beneath the operating system and survive a full reinstall. Out-of-band management—baseboard management controllers and similar interfaces—offers powerful, sometimes near-invisible control and is too often left poorly secured. Secure Boot integrity, the trust chain it depends on, and the Public Key Infrastructure (PKI) and cryptography underneath it are all in scope, and failures there are quiet and high-impact.

Measuring what matters

Measurement separates a mature program from a merely active one. We track remediation rates, repeat findings, time-to-fix, and—just as important—whether defenders detect and contain our activity. Validating detection and response, through assumed-breach exercises and close work with the defensive team, is the higher bar. We also measure whether our continuous investment, AI included, is buying real risk reduction at a defensible cost, and whether techniques from our operations are feeding back into that continuous layer rather than stalling in a report.

Securing the *AI ecosystem*

Securing the AI stack

The AI stack spans infrastructure, data, models, and applications. It underpins key sectors and must be secured to reduce risk, enable trusted deployment, and support sovereignty.

A secure AI stack allows governments and organizations to adopt AI confidently while maintaining control over sensitive systems and data. In this context, security enables sovereignty, allowing nations to align AI use with their values and regulations without fragmenting global systems.

Microsoft approaches AI security as a full-stack challenge, combining physical and cyber protections, secure supply chains, continuous testing, and strong governance to deliver defense-in-depth across the entire lifecycle.

Taking a lifecycle approach to AI security

At the infrastructure layer, physical and hardware security are foundational. Risk-based zoning, insider threat programs, and secure hardware lifecycle management ensure that sensitive compute and data-bearing assets are protected from tampering or compromise.

At the system and model layer, cybersecurity controls extend across training, deployment, and inference. This includes encryption of critical artifacts such as model weights, strong identity and access management, network segmentation, and real-time monitoring for anomalous behavior. AI-specific safeguards such as dataset validation, partitioning of training weights, and workload isolation help mitigate risks unique to AI systems.

Securing AI requires continuous evaluation, testing, red teaming, and governance. Lifecycle-based testing, including red teaming and post-deployment monitoring, enables organizations to detect and respond to emerging threats. Governance frameworks ensure accountability, transparency, and appropriate human oversight for high-risk actions.

Finally, supply chain security is a critical layer. AI systems depend on globally distributed hardware, software, and logistics networks, making secure-by-design hardware, trusted suppliers, and end-to-end chain-of-custody controls essential.

Governments need to consider security across the lifecycle of AI

Policy should focus on managing risk across the entire lifecycle, prioritizing real-world use, and ensuring security scales with innovation. Governments can adopt these globally applicable principles:

1. **Adopt a full-stack, lifecycle-based security approach.** Governments should define AI security requirements across the entire lifecycle, from infrastructure and data to models and deployment. Policies should focus on where risk is highest and ensure that security controls evolve over time through continuous monitoring and evaluation.
2. **Prioritize system-level assurance, not just model testing.** Security frameworks should move beyond testing AI models in isolation and instead assess how systems behave in real-world deployment environments. This includes interactions between models, tools, data, and users, as well as ongoing post-deployment evaluations.
3. **Align international standards to enable secure, scalable adoption.** Countries should work toward interoperable, risk-based AI security standards grounded in existing frameworks. Alignment enables trusted cross-border collaboration, reduces fragmentation, and supports the safe diffusion of AI technologies globally.

Securing the AI Stack

Physical security. Refers to the safeguards that protect facilities, infrastructure, and hardware from unauthorized access, damage, or interference.

Logical and cybersecurity. Encompasses access controls and defensive measures that govern system use and protect digital assets from unauthorized access or compromise.

Export controls. Legal and compliance measures that restrict the transfer, access, or use of sensitive technologies in accordance with national security and regulatory requirements.

Supply chain security. Secure, transparent supply chains are essential to protecting AI systems from tampering, disruption, and compromise across their lifecycle.

Red teaming and evaluations. Structured testing activities that assess system behavior and resilience by identifying weaknesses, misuse potential, and failure modes on an ongoing basis.

Governance policies. Encompasses the policies, oversight structures, and decision-making processes that guide development and deployment.

> **A secure AI stack allows governments and organizations to adopt AI confidently while maintaining control over sensitive systems and data.**

The agentic defense stack: Connecting attack goals to controls

Enterprises are increasingly adopting AI agents. Microsoft has observed 88% of enterprises are already experimenting with agents, and 82% of leaders plan broader rollouts within the next 12 to 18 months. Industry projections point to roughly 1.3 billion agents in production by 2028.

More agents bring more entry points for threat actors. For agents to fulfill their tasks, they need access to business systems and resources, increasing the need for effective security controls, governance, and oversight. Each agent needs its own identity, not a credential borrowed from a human employee. Protecting AI agents in 2026 is therefore an execution question of inventorying, identifying, and governing every agent, with defenders able to attribute, revoke, and enforce least privilege at scale.

The agentic defense stack: Five risk classes that define the 2026 attack surface

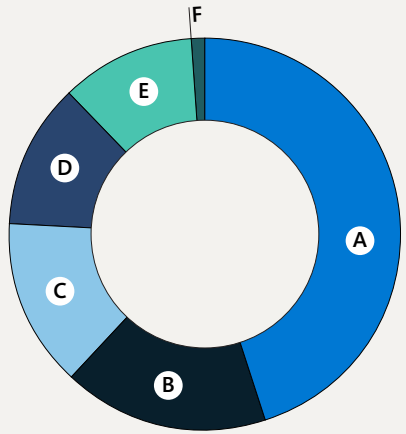
Effective protection requires consistent enforcement of controls across deployed agents.

Malicious or policy violating use includes attempts to generate harmful or prohibited content. Illustrative examples include drafting phishing emails or generating fraudulent invoices.

Figure 1. **What observed prompts most often try to achieve**

The goals shown describe what observed prompts tried to achieve, while a single attempt can involve several risk areas. For example, an attempt to misuse a connected tool to extract customer records can involve both goal hijacking/downstream tools and sensitive data exfiltration.

Key		%
A	Malicious use/policy violation	45%
B	Sensitive data exfiltration	17%
C	System prompt leakage	14%
D	Goal hijacking/ downstream tools	12%
E	System-internals leakage	11%
F	Resource exhaustion / financial harm	1%



Source: Microsoft observation of prompts against generative AI agents (February to May 2026).

Risk area	What attackers do	What defenders need
Prompt and intent manipulation	Inject instructions through prompts, content, files, or memory to redirect the agent	Prompt injection detection, payload inspection, intent validation, output review
Sensitive data exposure	Drive the agent to read or return content beyond its sanctioned scope	Sensitivity-aware retrieval, response inspection, data loss prevention (DLP) on inputs and outputs, memory scoping
Identity and privilege compromise	Impersonate a sub-agent, reuse stolen credentials, or chain privileges across agents	Verifiable agent identity, mutual authentication, scoped credentials, least privilege
Excessive agency	Coax the agent into chaining tools or actions beyond sanctioned behavior	Tool allow-listing, runtime gating, action policy, anomaly detection
Operational integrity	Tamper with configuration, system prompts, memory, training data, supply chain, or logs	Immutable logs, signed configuration, change governance, supply chain attestation

The agentic defense stack: Connecting attack goals to controls (continued)

Inside the agent attack surface: Where attack activity is concentrating

The agent attack surface in 2026 was dominated by a small number of high-volume techniques. In telemetry data from Azure AI workloads, four classes of techniques accounted for roughly 90% of observed attack activity:

- Malicious link injection routed through model and tool outputs (52%)
- Jailbreak and safety bypass attempts (16%)
- Tool and agent abuse (12%)
- Identity and credential abuse (10%)

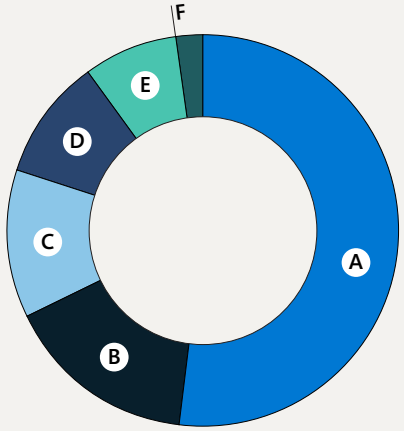
Organizations also need to plan for agent-to-agent spoofing, a rising technique that is expected to grow as multi-agent deployments scale. In this technique, one agent impersonates another (or its owner) to slip past trust boundaries across a multi-agent graph.

Agents create different challenges from traditional software

When a human employee changes roles, organizations have well-established processes, including pre-configured access entitlements, manager approvals, periodic recertification, and regular access reviews. But as agents acquire new skills, the boundaries of what they can do shift in ways that may not trigger the same governance checkpoints. Organizations need systems and controls that enforce policy in real time, keeping every action aligned to the agent's objective, its intent, and the organization's boundaries. Without those controls, autonomous agents operate unchecked, becoming unmanaged, potentially overprivileged, with a growing governance gap.

Figure 2. **Top attack vector categories in telemetry data on AI workloads**

Malicious link injection, tool and AI agent section abuse, and prompt subversion account for the majority of observed attack activity on AI workloads.



Key		%
A	Malicious link injection and exfil	52%
B	Jailbreak and safety bypass	16%
C	Tool and agent abuse	12%
D	Identity and credential abuse	10%
E	Prompt injection and instruction subversion	8%
F	Reconnaissance and data probing	2%

Source: Telemetry data on AI workloads (90 day window, 2026).

There are three risks as organizations adopt agents:

1. **Ungoverned agent proliferation.** The number of agent identities often grows faster than the organization's ability to track them. Without centralized visibility, agent sprawl creates even greater risks than shadow IT.
2. **Overprivileged access.** Agents acquire access over time that's rarely revoked. Unlike human users, who face governance controls, agents can accumulate broad access that persists and becomes stale after the agent's original purpose has evolved.
3. **No human accountability.** When an agent acts autonomously, the question of who's responsible for its actions becomes critical. Without a human accountable for each agent's access (a sponsor), there's no one to contact when something goes wrong.

A risk-based approach to agent governance

To deploy agents safely and meet governance objectives, security admins need to take a risk-based approach to identifying, authorizing, and reviewing agent activities, with automated mitigations to remediate critical risks:

1. **Start with identity.** Every agent must have a distinct, auditable identity that carries metadata about who created the agent, its purpose, and its sponsor.
2. **Enforce human accountability.** When an agent's sponsor changes, agent accountability must be re-evaluated and transferred automatically to a new sponsor. No environment should ever have an agent without a human to answer for it.

> "Organizations need systems and controls that enforce policy in real time, keeping every action aligned to the agent's objective, its intent, and the organization's boundaries.

The agentic defense stack: Connecting attack goals to controls (continued)

3. **Apply least-privilege access from day one.** Agents should receive only the access they need to perform their tasks, with access assignments time-limited and subject to approval workflows. New permissions should go through governance controls appropriate to what's being granted, and unused access should be revoked to prevent permissions creep.
4. **Automate lifecycle management.** Manual oversight of agent identities isn't feasible at scale. Automated policies should verify the ongoing need for each agent, whether its access is still appropriate, and confirm it still has a valid human sponsor. Organizations should automatically disable and clean up unnecessary agents and their access permissions.
5. **Govern consistently across platforms and services.** Consistent governance policies need to apply regardless of each agent's origin. Using different rules for different platforms creates audit and visibility gaps that adversaries can exploit.
6. **Require agent-to-agent authentication.** Sub-agents must prove identity to one another, not only to the originating human session. Unauthenticated cross-agent traffic is how a single compromise propagates through a multi-agent graph.
7. **Issue per-tool scoped credentials.** Use short-lived secrets bound to a single tool invocation, not long-lived tokens that hold every privilege the agent might ever need.
8. **Treat revocation latency as a measured outcome.** When an agent or its owner is compromised, the time to revoke across identity, posture, and threat protection controls is the operational metric that determines incident impact.

The path forward for organizations and policy makers

As AI agents proliferate, an organization's governance plans cannot rely on platform-specific controls alone. Interoperable standards play a critical role in enabling trusted collaboration between agents, services, and across organization boundaries. Effective agent governance also requires internal alignment. Organizations that do not already have a suitable forum should form a dedicated agent governance committee, representing application owners, data custodians, business process designers, compliance specialists, and agent sponsors. This group should work together to preserve controls, adapt processes to agent risk, and ensure technical controls support day-to-day operations.

For industry and policymakers, there are two near term priorities: faster convergence on practical security guidance and clearer accountability for deployment decisions. In the near term, governments can help by supporting common taxonomies, evaluation methods, and risk-management expectations that reflect how agents actually operate across memory, tools, and connected systems. Early guidance should enable innovation while setting credible expectations for security, auditability, and human oversight before more formal standards and regulatory frameworks are fully established. Industry, in turn, should translate those principles into deployable controls, transparent testing practices, and governance models that assign responsibility for high-risk uses.

Learn more

Frontier Governance Framework

Efforts to share best practices on AI measurement and evaluations driven forward through the International Network for Advanced AI Measurement, Evaluation and Science

Frontier Model Forum: Advancing frontier AI safety and security

Advancing AI evaluation with the Center for AI Standards (US) and Innovation and the AI Security Institute (UK)



Memory poisoning: Persistence in the agentic age

Assistants and task agents rely on long-term memory to personalize answers, preserve context across sessions, and coordinate multi-step work. In many memory implementations, the agent is authorized to insert, update, and delete entries on the user's behalf.

Letting the model write to its own long-term state from a context it does not control removes the boundary between session input and durable system state. This means an attacker who can place content anywhere the agent will read can also often place content anywhere the agent will remember. That content is then retrieved as trusted context in future sessions, across chats, and in some cases across task handoffs between agents. The blast radius grows with agentic coupling. Once memory is shared across tools, tenants, or agents, a single poisoning incident can influence downstream automation that the original victim never sees.

Email is a key pathway for memory poisoning

The dominant memory-poisoning pathway the Red Team has observed is XPIA from externally-sourced content delivered through sources that the agent is expected to consume as part of normal operation, most commonly email. The attacker embeds instructions that the system retrieves and blends into the same reasoning context as the user's request. The model can then be induced to create a new memory entry or modify an existing one that stores attacker-influenced behavioral guidance that can persist across sessions and be reintroduced as trusted context in later interactions.

Human-in-the-Loop (HITL) controls alone cannot protect memory writes. In red team testing, poisoned entries have been used to suppress or bypass these controls, append unauthorized steps to legitimate tasks, and subtly steer subsequent responses or outputs. The Red Team has also observed instances where HITL is triggered but ignored by the agent and HITL being bypassed by a policy stored in memory in the same attack chain. The underlying issue is that the control and the thing it guards share a failure mode.

Memory poisoning is a reproducible, low-complexity, cross-session attack surface. To respond, defenders need to:

- Create architectural separation between trusted and untrusted write paths so that untrusted data cannot trigger writes to a trusted memory store. This includes restricting memory write tool calls once untrusted data has entered an agent's context.
- Implement a memory architecture that treats agent-derived preferences separately from explicit user instructions that can modify safety behavior. Storing in memory how a user should be referred to can be agent-derived, instructions to bypass a HITL control should not.
- Monitor in a way that treats the memory store as a high-value asset whose integrity determines the integrity of everything downstream. This includes giving users the ability to audit and control memory entries at a granular level and collecting telemetry around memory actions.

> **Once memory is shared across tools, tenants, or agents, a single poisoning incident can influence downstream automation that the original victim never sees.**

Overreliance on Human-in-the-Loop (HITL) is a systemic risk

HITL controls are a key component of AI system security, and their importance has only increased with the adoption of highly privileged agentic AI systems. HITL is now the default mitigation whenever a new agentic capability outpaces its classifiers: "enable the capability but ask the user before doing anything sensitive".

The pattern is embedded across chat bots, automated workflow AI, and complex agentic systems. However, the decision to invoke the HITL control is often made by the same system an attacker can manipulate, using context that has already mixed trusted and untrusted content.

HITL controls can be easily bypassed by adversaries

As agents gain memory, tool use, and autonomy, the bar to bypass HITL is dropping and the blast radius of a single bypass is rising. Defense-in-depth must move upstream into provenance, isolation, deterministic invocation, and classifier-backed policy or else the next class of incidents will be ones where the user "approved" the breach themselves.

What works when it comes to HITL

- **Deterministic, policy-controlled invocation of HITL** ensures that it can't easily be bypassed by someone manipulating the underlying model. This should also be supported by deterministic controls that block patterns of activity that shouldn't be approved under any circumstances.
- **HITL with user experiences that support the users**, with clear explanations of what is being requested and why, with relevant context and a chain of how the decision to ask for the action was reached.
- **Mitigations to prevent fatigue**, including controls at the user level and telemetry to allow detection and response to HITL fatigue attacks.

> "As agents gain memory, tool use, and autonomy, the bar to bypass HITL is dropping and the blast radius of a single bypass is rising.

Three key HITL failure modes the Red Team has observed

Category	Example technique
Prompt injection and probabilistic invocation	• Memory-poisoned consent • Injected approvals • Act-then-ask inversion • Chaining low-risk actions to reach a high-risk outcome
Inadequate human involvement	• Tenant-wide always-allow • Illegible model-generated summaries • No fatigue mitigations • Blanket session approvals
Control-quality issues	• HITL that doesn't fire (UI-state-dependent) • Fires but is ignored • Starting context effect on chain-of-thought (CoT) vigilance • Time-of-check/Time-of-Use (TOCTOU) on computer use agents



AI as the new *attack surface*

AI-assisted threats

An emerging threat category is attacks that weaponize the trust that users and organizations place in AI tools, AI-assisted workflows, and AI-generated identity as the primary exploitation mechanism. Recently, Microsoft has observed three specific patterns in this category:

Malicious AI applications
Malicious AI applications are a newly emerging category of malware disguised as legitimate AI productivity tools that exploit the surge in AI adoption to earn user trust. EvilAI is the first of a class we expect to grow as AI tooling becomes a more attractive lure. EvilAI comes complete with professional branding and valid code-signing certificates, but once installed, it silently deploys a persistent backdoor that steals credentials, payment data, and sensitive files.

Malicious browser extensions impersonating AI brands
Threat actors are increasingly exploiting trusted AI brand names to deceive users. Microsoft observed a malicious browser extension impersonating the AI answer engine Perplexity AI that silently collected user search activity. Because the user still landed on normal results, the interception was invisible.

Malicious AI extensions that harvest LLM data
Malicious browser extensions masquerading as AI assistants have reached nearly 900,000 installs and affected more than 20,000 enterprise tenants. These extensions silently collect URLs, page content, metadata, and AI chat histories from platforms such as ChatGPT and DeepSeek, exposing proprietary code, business strategies, and other sensitive information provided to the platform by users.

AI browsers as an enterprise risk: Elevated permissions with limited security visibility

AI browsers are becoming a material enterprise risk because they combine three conditions security leaders rarely face at the same time: broad adoption, high-trust access, and limited security visibility.

The browser is already the primary workspace for modern organizations—recent industry research indicates that roughly 85% of work now happens there. At the same time, AI browsers are rapidly becoming part of the enterprise technology stack, with Microsoft observing daily use across more than 40% of organizations in its dataset.

The web browser now acts as a powerful computer program: it runs heavy code, stores data, and performs complex tasks. In AI-enabled deployments, it is increasingly an execution environment with access to authenticated sessions, cookies, OAuth state, page content, downloads, clipboard data, and delegated actions performed on behalf of users—all of which are attractive to attackers.

The shift in browser use changes the enterprise risk equation

Traditional security models were built around a clearer separation between user intent, application behavior, and system-level execution. AI browsers compress those boundaries. They can read untrusted web content, reason over that content, and then take action inside live enterprise sessions using the user's existing authority. That creates a high-trust environment where abuse can look operationally normal.

A malicious or manipulated action may present as ordinary software-as-a-service (SaaS) usage rather than malware execution, making it harder for security teams to distinguish between legitimate productivity and compromise. Adversaries are already targeting this vector. In an analysis conducted in May 2026, we observed AI browsers as the initial infection vector in incidents involving 57 distinct malware families. We also documented a malicious extension campaign that affected more than 900,000 installs across Chromium-based browsers (including agentic browsers) and showed confirmed telemetry activity across more than 20,000 enterprise tenants. The campaign harvested browsing history, AI chat histories, prompts and responses, and session material while blending exfiltration into ordinary browser traffic.

Cross-signal threat intelligence is key

AI is reshaping cybersecurity investment decisions for both attackers and defenders by changing the economics of adversarial adaptation.

Attackers have always benefited from an "adversarial discount". That is, the value of defensive technology erodes as attackers adapt to it, not because of a product defect or temporary gap, but because determined adversaries strategically learn to exploit deployed defenses over time.

AI amplifies this dynamic by shifting parts of the attack lifecycle—such as vulnerability discovery, exploit development, and lateral movement—from economies of skill toward economies of scale in compute. This increases the speed and magnitude of adversarial discount.

Given the value of a target, what determines breach probability is the ratio of offensive to defensive effectiveness—a ratio that threat actors are actively working to increase. In addition to making attacks more potent across the board, attacker AI specifically targets whatever the defender has deployed. A defender who invests in detection gives the attacker something precise to study and evade. A defender who invests in triage gives the attacker an incentive to overwhelm it. The defender's own investment feeds the attacker's incentive to invest. This creates an arms race.

> "The degree to which a defender's systems share and correlate threat intelligence is one of the highest-leverage strategic variables under the defender's control.

Can AI just find and fix everything first?

A natural question is whether defenders can neutralize the adversarial discount by using AI to find vulnerabilities before attackers do. Anthropic's Project Glasswing and the Claude Mythos Preview model have the ability to identify zero-day vulnerabilities at scale. If defenders have access to this capability, why doesn't the adversarial discount disappear?

As our Red Team has shown, the answer lies in the difference between finding a vulnerability and eliminating it. Patching requires coordinating across software maintainers, testing under operational conditions, managing uptime risk, and deploying across multiple environments. While attackers adapt in real time to deployed defenses, defenders must convert discovered vulnerabilities into safe system changes. This asymmetry ensures the adversarial discount persists even with the help of AI.

Signal cross-correlation is a decisive variable

The degree to which a defender's systems share and correlate threat intelligence is one of the highest-leverage strategic variables under the defender's control. This sharing, or signal cross-correlation, is the most powerful of all the factors over which the defender has control. Consider a large organization protecting hundreds of applications, endpoints, and cloud services. If each system's telemetry is analyzed in isolation, adding a new system adds a detection burden without contributing to the defense of existing ones. Defense effectiveness per system dilutes as the environment grows.

With signal cross-correlation, the opposite occurs: an attacker moving across multiple systems leaves a cross-system footprint. These are patterns visible across systems even if no single system shows a complete picture. Each new system becomes a detection opportunity instead of a new liability.

Formal modeling of this mechanism demonstrates a striking result: with full signal cross-correlation within an organization, the defender's marginal effectiveness per unit of investment is independent of how many systems they must protect. The surface-count disadvantage disappears. This points to concrete architectural interventions: unified detection infrastructure, cross-product telemetry pipelines, and shared behavioral baselines.

AI-augmented attacks tend toward complex, multi-stage, multi-surface campaigns, but these generate cross-surface artifacts. When defenders invest sufficiently in signal correlation, they make complex attacks easier to detect.

Organizations need to consider system-level intelligence

Signal correlation is an architectural capability that transforms isolated detections into a connected defense ecosystem. Organizations need to consider how effectively intelligence from one system informs another. As environments expand across endpoints, identities, email, cloud services, networks, and applications, the value of each security signal increases when it can be correlated with others. The return on security investments is therefore driven not only by the quality or quantity of individual tools, but by the organization's ability to unify telemetry and intelligence across its environment. Each additional source of visibility strengthens the value of every other source, improving detection, accelerating response, and providing broader context for security decisions.

Cross-signal threat intelligence is key (continued)

At enterprise scale, integration itself becomes a defensive advantage. When endpoint, identity, cloud, application, email, network, data, and AI-workload signals are combined with threat intelligence and AI reasoning, each additional source contributes context to the rest of the estate and enables faster, coordinated action. The value comes from the system working together, not from any single control in isolation.

The same principle applies beyond individual organizations

Cross-organizational signal correlation enables the identification of campaign-level activity that would otherwise remain invisible within any single environment. No organization possesses a complete view of the threat landscape; each captures only a fraction of the available signals. The collective benefit created through trusted information sharing exceeds what any organization can achieve independently. Policymakers can help unlock these benefits by encouraging mechanisms that support secure collaboration, including legal safe harbors for information sharing, common anonymization standards, and investments in shared detection and threat intelligence infrastructure. Together, these approaches strengthen not only individual organizations, but the resilience of the broader digital ecosystem.

From capability to responsibility: Securing the AI-driven digital ecosystem

Frontier AI systems are reshaping cybersecurity by dramatically accelerating the discovery, analysis, and exploitation of software vulnerabilities.

This creates a dual-use reality: the same capabilities that allow defenders to identify and remediate weaknesses at unprecedented scale can also be leveraged by adversaries to conduct offensive cyber operations. Whether AI ultimately shifts the balance in favor of defenders or attackers will depend on the choices governments, industry, and civil society make today.

At the same time, frontier AI is creating a new class of national security risk driven by speed, autonomy, adaptability, and increasingly sophisticated cyber capabilities. Risk management frameworks, including Microsoft's Frontier Governance Framework, are converging around a set of high-impact domains, including offensive cybersecurity, chemical, biological, radiological, and nuclear (CBRN) risks, and advanced autonomous behaviors. These risks are dynamic and evolve alongside technological progress, requiring continuous evaluation and adaptation rather than one-time assessment.

Closing the vulnerability discovery gap has become a strategic imperative

Organizations must modernize vulnerability management practices, accelerate remediation timelines, and embed security throughout the AI lifecycle. Recent advances in AI-driven security systems demonstrate how automated techniques can identify vulnerabilities earlier, validate exploitability, and prioritize remediation based on real-world risk rather than sheer volume.

A defining development in this evolution is the rise of multi-model, multi-agent AI systems that combine the strengths of specialized models working collaboratively toward shared security objectives. By distributing tasks across coordinated agents and leveraging different model capabilities, these systems have demonstrated significant improvements in efficiency, scalability, and cost-effectiveness compared to traditional or single-model cyber defense approaches. Emerging deployments show that such architectures can accelerate vulnerability discovery, improve analysis quality, and support faster remediation decisions across increasingly complex digital environments. These advances reinforce an important lesson for the AI era: competitive advantage increasingly lies not in any individual model, but in the integration of models, tools, and processes into cohesive defense systems capable of operating at machine speed.

However, model-level capability evaluations alone are not sufficient. National security risks depend not only on what a model can do in isolation, but on how it is integrated into broader systems and operational environments. Current evaluation approaches often focus on narrow technical tasks performed under controlled conditions, overlooking the strategic intent, complexity, and adaptability of real-world threat actors.



AI as the new attack surface

From capability to responsibility: Securing the AI-driven digital ecosystem (continued)

Without incorporating government and intelligence-informed threat models, assessments may underestimate how advanced AI systems could be misused, leaving critical gaps in safeguards and defensive preparedness.

Why public–private collaboration matters

No single institution has the expertise, visibility, or authority required to manage frontier AI risks alone. Governments provide national security expertise, intelligence insights, and regulatory capabilities. Industry contributes technical knowledge, operational experience, and direct access to frontier systems. Academic institutions and independent researchers advance testing methodologies and generate critical evidence on emerging threats.

As a result, effective AI security depends on structured public–private collaboration. Initiatives such as the AI Network for Advanced AI Measurement, Evaluation and Science, the Frontier Model Forum, and national AI security research institutes are helping align evaluation priorities, testing methodologies, and risk-management approaches. Similarly, partnerships between Microsoft and organizations such as the US Center for AI Standards and Innovation and the UK's AI Security Institute demonstrate the value of jointly testing frontier models, assessing safeguards, and improving methodologies for evaluating national security and offensive cyber risks. Microsoft's External Red Teaming for AI (EXTRA) program further supports independent research by providing AI safety and security researchers with access to frontier technologies to advance the science of AI risk evaluation.

An emerging consensus is beginning to shape the future of frontier AI governance:

- Risk management should be capability-based, dynamic, and evidence-driven.
- National security testing must be collaborative, drawing expertise from both the public and private sectors.
- Governance must extend beyond individual models to encompass complete systems and deployment contexts.
- Evaluation should be continuous and adaptive as models and threat landscapes evolve.
- Public–private partnerships are essential for effective testing, mitigation, and trust-building.
- Shared methodologies, datasets, and evidence frameworks improve consistency, transparency, and confidence in evaluation outcomes.

Policy priorities for governments

To ensure AI strengthens security rather than undermines it, governments should focus on five priorities:

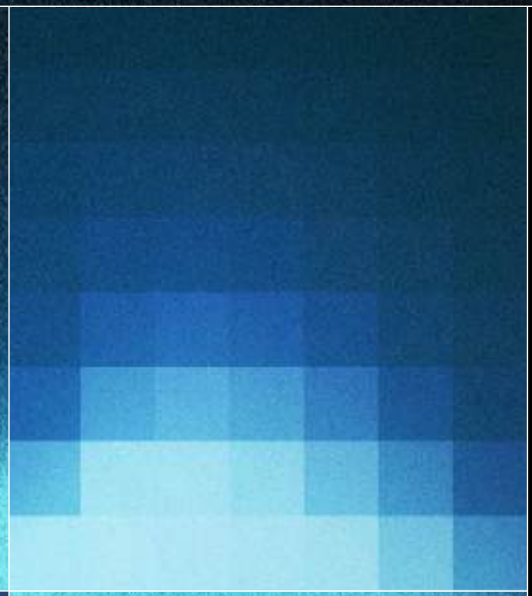
1. **Accelerate AI-enabled, risk-prioritized vulnerability management.** Governments should encourage a transition from volume-based approaches to risk-based vulnerability management supported by AI-driven discovery, validation, and prioritization technologies.
2. **Invest in automated remediation and ecosystem-wide resilience.** Public policy should support automated patching, coordinated vulnerability disclosure, and stronger support for open-source maintainers who play a critical role in global software security.
3. **Reinforce foundational cybersecurity practices.** Core security measures—including Zero Trust architectures, strong identity protections, secure-by-design principles, and rapid patching—remain the foundation upon which AI-enabled security capabilities must be built.
4. **Risk-based testing and phased deployment of advanced AI systems.** High-capability AI models with cybersecurity-relevant capabilities should undergo rigorous pre-deployment assessments that evaluate misuse scenarios, national security implications, and potential pathways to harm. Controlled release mechanisms can help balance innovation with risk mitigation.
5. **Deepen international cooperation on AI security and cyber resilience.** Governments should expand collaboration on AI security standards, threat intelligence sharing, testing methodologies, and coordinated approaches to emerging cyber risks, helping create common expectations for responsible AI development and deployment.

As frontier AI capabilities continue to advance, the central challenge is no longer simply developing more capable systems—it is ensuring those capabilities are deployed responsibly. Securing the AI-driven digital ecosystem will require a combination of advanced defensive technologies, rigorous governance, resilient cybersecurity foundations, and sustained collaboration across governments, industry, and academia. The choices made today will determine whether AI becomes a force multiplier for cyber defense or a catalyst for systemic cyber risk.

> “
>
> **No single institution has the expertise, visibility, or authority required to manage frontier AI risks alone.**

2

The threat landscape





> Cyberthreats do not respect borders, and neither can our response. Through our collaboration with Microsoft and the CSIRTAmericas network, we are helping equip countries across the Americas with timely threat information that strengthens collective cyber resilience and supports more coordinated responses to emerging threats.
>
> **Diego Subero, Cybersecurity Section Officer**
> Organization of American States

The threat landscape in 2026 32
Cyberattacks typically hit known hotspots 32
Emerging threats 36
Nation state and geopolitical threats 38
Regional sample of activity level observed 38
Significant trends in state-sponsored threat activity (China, Iran, Russia, North Korea) 39

How modern intrusions begin 44
The initial access landscape 44
ClickFix is evolving 45
Voice phishing attacks go mainstream in chat applications 46
The cost of delayed patching 47
Impersonation is becoming more sophisticated 49
Business contact impersonation: The new "BEC" 50
Emerging threat of network edge infrastructure 50
Information stealers: A continued threat 52

Detection, intelligence, and disruption 54
Operationalizing detection at scale 54
How shared threat intelligence is reshaping the fight against fraud 55
Targeting fraud's backbone 57

Securing the foundations 59
The race to retire NTLM 59
The cloud threat landscape 59
Rethinking defense models to tackle intrusions in hybrid cloud environments 61

The threat landscape *in 2026*

Cyberattacks typically hit known hotspots

While cyberattacks threaten every country, specific hotspots persist year over year, driven by target density and geopolitical interests.

For example, Ukraine and NATO-aligned countries remain focal points for Russian activity, while Chinese state actors concentrate heavily on the Indo-Pacific, the United States, and strategic technology hubs. Meanwhile, Microsoft has observed growing cybercriminal activity impacting Latin America, Africa, and Southeast Asia—regions where defenses may be less mature at the same time that digital transformation is accelerating.

Countries where customers are most frequently impacted by cyber threats

(January 2026–June 2026)



Scale of impact: % of total	%		%		%
1 United States	25.5	5 United Kingdom	3.8	9 Canada	1.8
2 Israel	7.6	6 United Arab Emirates	2.8	10 Germany	1.7
3 Ukraine	4.8	7 India	2.5	11 Australia	1.5
4 Taiwan	3.9	8 Japan	2.5	12 Philippines	1.5

Source: Microsoft Threat Intelligence

The threat landscape in 2026

Cyberattacks typically hit known hotspots (continued)

Targets are staying the same

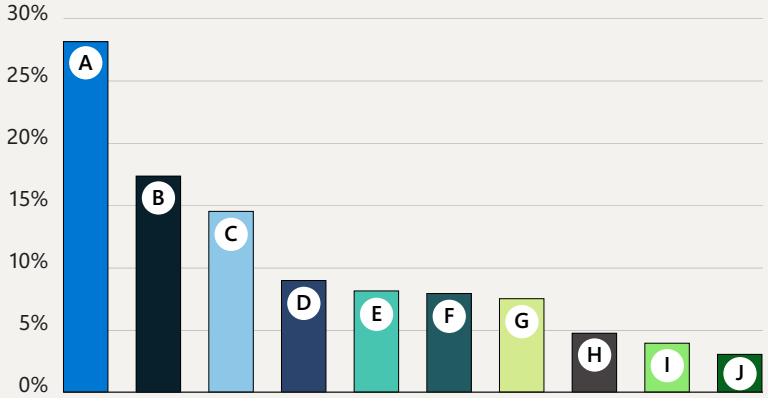
Threat actors continue to prioritize the same targets: government, information technology (IT), and research and academia. These sectors represent attractive targets for both nation-state and cybercriminal actors, largely because they:

- Have high-value intelligence
- Contain extensive personally identifiable information (PII)
- Comprise a sprawling and often under-monitored infrastructure
- Have a generally low tolerance for downtime

At the same time, businesses outside of critical sectors—especially small and medium-sized businesses—remain at high risk of cybercriminal activity, from ransomware to business contact impersonation (BCI).

Threat actors are relying less on custom malware and instead abusing legitimate tools, cloud services, and administrative functions. This makes detection harder because malicious activity often blends in with normal user behavior. Adversary-in-the-middle (AiTM) phishing kits, OAuth app abuse, API exploitation, and password spray attacks are now standard across both criminal and nation-state playbooks. This marks a major evolution from years past, where endpoint malware and exploit kits dominated initial access.

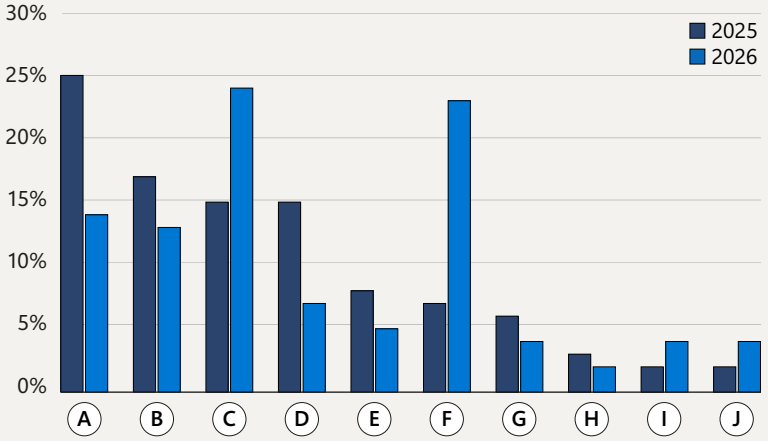
Top 10 global sectors impacted by threat actors in 2026



Key		%
A	Government agencies and services	27
B	Information technology	17
C	Research and academia	14
D	Communications infrastructure	9
E	Transportation systems	8
F	Non-government organization	8
G	Defense industrial base	7
H	Critical manufacturing	5
I	Financial services	4
J	Energy	3

Source: Microsoft Threat Intelligence

Initial access vectors across two years of incident response (IR) findings



Key		2025	2026
A	None identified	25%	14%
B	Valid accounts	17%	13%
C	Exploit public-facing application	15%	24%
D	Social engineering	15%	7%
E	External remote services	8%	5%
F	Phishing	7%	23%
G	Drive-by compromise	6%	4%
H	Content injection	3%	2%
I	Trusted relationship	2%	4%
J	Supply chain compromise	2%	4%

2025 reporting period: July 2024–June 2025; 2026 reporting period: July 2025–June 2026. Source: Microsoft Threat Intelligence

The threat landscape in 2026

Most common MITRE ATT&CK techniques for compromise and post-compromise based on Microsoft Defender alerting, April to June 2026

Tactic	Technique
Initial Access	T1078 Valid Accounts
Execution	T1059 Command and Scripting Interpreter
Persistence	T1547 Boot or Logon Autostart Execution
Privilege Escalation	T1548 Abuse Elevation Control Mechanism
Stealth	T1218 System Binary Proxy Execution
Defense Impairment	T1222 File and Directory Permissions Modification
Credential Access	T1110 Brute Force
Discovery	T1087 Account Discovery
Lateral Movement	T1021 Remote Services
Collection	T1114 Email Collection
Command and Control	T1105 Ingress Tool Transfer
Exfiltration	T1048 Exfiltration Over Alternative Protocol
Impact	T1486 Data Encrypted for Impact
Reconnaissance	T1589 Gather Victim Identity Information
Resource Development	T1586 Compromise Accounts

Cyberattacks typically hit known hotspots (continued)

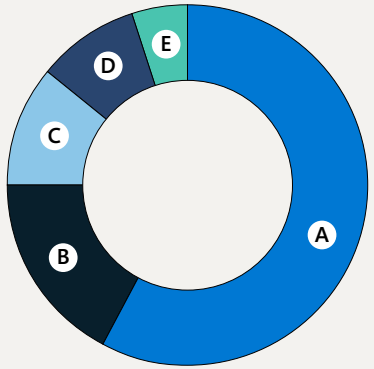
Attackers are targeting well-known CVEs

Analysis of vulnerabilities based on Microsoft Defender alerting from July 2025 to June 2026 reveals most attacks linked to potential vulnerability exploitation involved well-known, long-disclosed Common Vulnerabilities and Exposures (CVE). These older vulnerabilities may be targeted because unpatched systems are easier to access or because widely available exploit code and tools incorporating exploit code are commonly weaponized, for example in commodity hacking tools and botnets.

These CVEs affect network-exposed or externally reachable services and are typically high-severity vulnerabilities that enable high-impact outcomes, such as remote code execution (RCE), trust bypass, or privilege escalation. The most effective mitigation is rapid and consistent patching of internet-facing and critical systems, as these vulnerabilities are effectively mitigated once vendor updates are applied and systems are properly secured.

Top five most exploited CVEs, per detection		%
A	CVE-2020-1472	58%
B	CVE-2022-22954	17%
C	CVE-2021-40444	11%
D	CVE-2023-28231	9%
E	CVE-2020-0601	5%

Source: Microsoft Threat Intelligence



Type	Technique
CVE-2020-1472 (Netlogon EoP)	An unauthenticated attacker could gain domain administrator access or impersonate domain member computer accounts.
CVE-2022-22954 (VMware Workspace ONE)	An unauthenticated attacker with network access could trigger a server-side template injection that may result in RCE.
CVE-2021-40444 (MSHTML RCE)	This zero-day vulnerability could enable an attacker to execute arbitrary code by delivering specially crafted Microsoft Office documents that load malicious ActiveX content when opened by a user.
CVE-2023-28231 (DHCP RCE)	An attacker could force the DHCP Server Service in Windows operating systems to run malicious code without requiring authentication or user interaction.
CVE-2020-0601 (Windows CryptoAPI Spoofing)	An attacker could sign malicious files with fake certificates.

Note: Detection events do not represent confirmed exploitation of the vulnerabilities. Additionally, many signals related to exploitation of known vulnerabilities might be captured in more general behavioral alerts and are therefore not identified in this analysis.

The threat landscape in 2026

Cyberattacks typically hit known hotspots (continued)

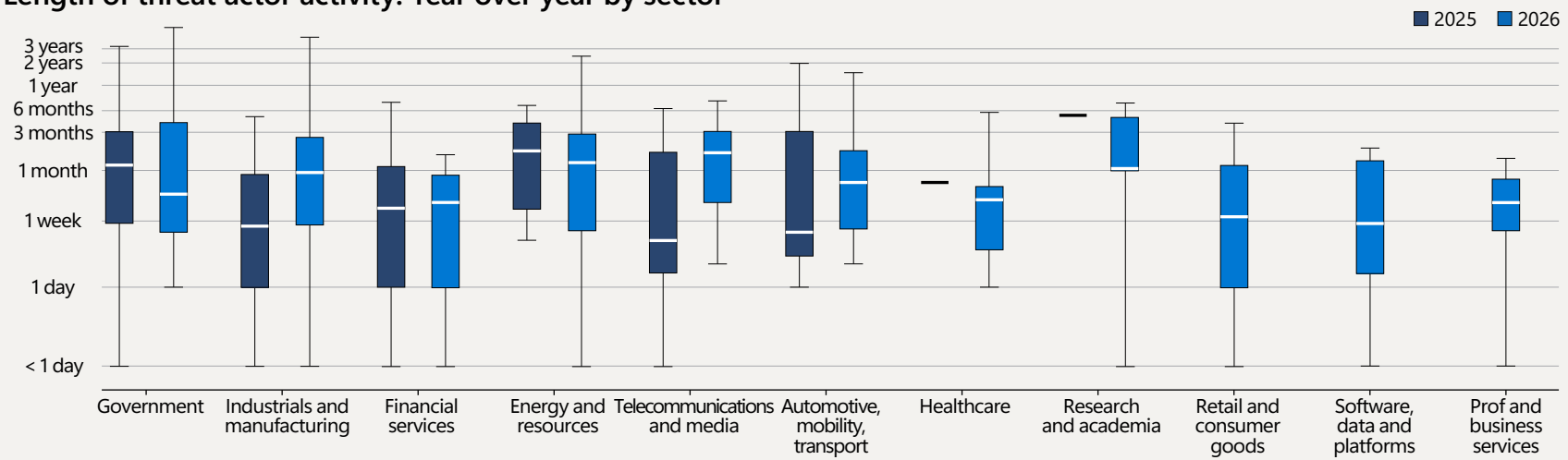
Dwell time is increasing as organizations struggle to detect intrusions

Dwell time is the period between the earliest evidence of threat actor activity and its detection, while length of threat actor activity measures the period between the earliest and latest observed activity. Comparing year over year DART engagement data shows a statistically significant increase in dwell time, while the overall length of threat actor activity remained broadly stable. This suggests organizations took longer to identify malicious activity following compromise, with the consistency of the increase across multiple sectors indicating a broader operational trend.

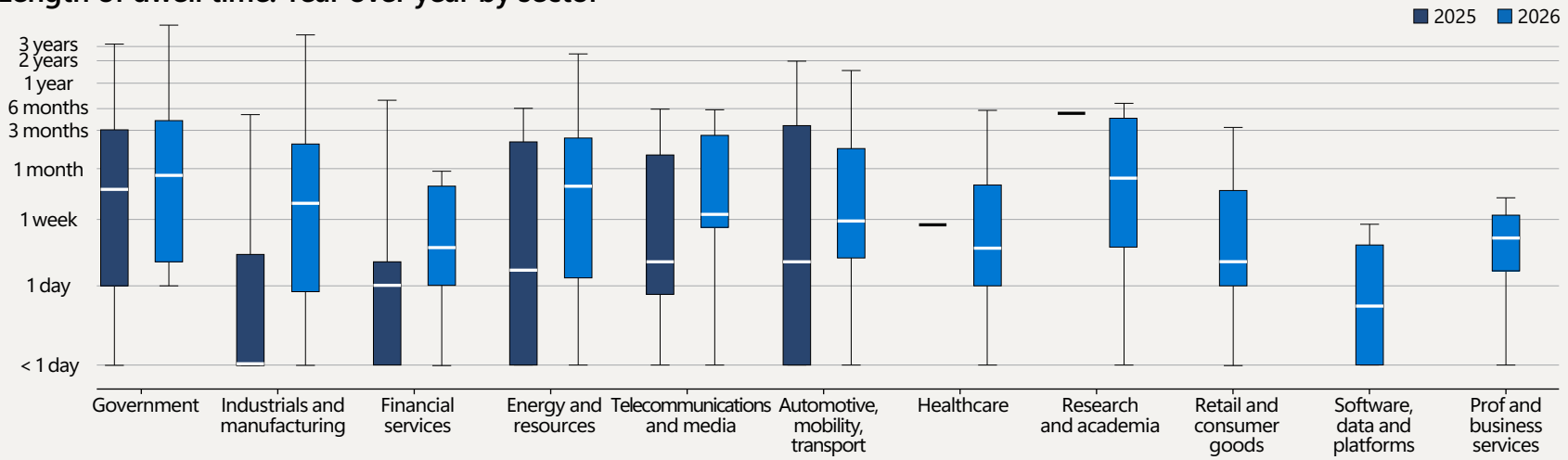
The most pronounced increase in dwell time was observed in industrials and manufacturing sector, while telecommunications and media showed the largest increase in the duration of threat actor activity. The latter is consistent with observations that nation-state actors seek to establish quiet, persistent access to targets of espionage interest.

The box plots illustrate both the typical duration and the degree of variation within each sector. The central line shows the median, while the endpoints of the extending lines represent the shortest and longest observed durations. A tighter spread indicates more consistent experiences; a wider spread highlights greater variation between incidents. Comparing the position and spread of 2025 and 2026 therefore shows both how typical durations changed and how consistently those changes were experienced across incidents.

Length of threat actor activity: Year over year by sector



Length of dwell time: Year over year by sector



2025 reporting period: July 2024–June 2025; 2026 reporting period: July 2025–June 2026. Source: DART

The threat landscape in 2026

Emerging threats

There continues to be a fundamental shift by attackers to exploit trusted systems, operational blind spots, and the scale of modern digital ecosystems to gain advantage. Looking over the horizon, Microsoft believes the following emerging threats will play an increasing role for both cybercriminal and nation-state actors in the next year:

1 Open-source supply chain compromise

Open-source ecosystems such as npm and Python Package Index (PyPI), and continuous integration and continuous delivery (CI/CD) have become a critical layer of the modern attack surface. Today, nearly all enterprise software depends on third-party components, often numbering in the hundreds or thousands per application. Each dependency introduces potential risk. Attackers can exploit the inherent trust in official channels to infect downstream systems by:

- Using compromised maintainer accounts
- Creating fake maintainer accounts to host malware
- Leveraging repositories to publish malicious updates

These packages often execute automatically during installation or CI/CD workflows. This does not require any user interaction and enables stealthy access to cloud credentials, tokens, and enterprise systems without requiring direct exploitation of hardened targets.

Defending against supply chain compromise is complicated by the scale and complexity of modern dependency trees, which create visibility gaps and signal-to-noise challenges. Malicious packages also often lack known signatures or disclosures, further limiting detection. While AI-driven analysis can improve dependency visibility and support automated auditing and remediation, it also introduces some risk, as adversaries begin to manipulate package ecosystems to influence automated selection by coding agents.

2 Attacks against edge devices

Edge devices such as firewalls, virtual private network (VPN) gateways, and routers have become a primary target. Located at the boundary between enterprise networks and the internet, these systems provide high-value access to traffic flows and authentication material and are often less well monitored or patched than traditional endpoints. Once compromised, edge devices can be repurposed into part of the threat actor's operational infrastructure or used to intercept traffic, conduct AiTM attacks, or enable further intrusions against downstream targets. Attackers may combine compromised edge devices with resilient infrastructure such as bulletproof hosting providers to obscure activity, host command and control (C2) and extend persistence.

Defending this layer is challenging due to visibility and forensic limitations. Many edge appliances lack endpoint detection capabilities, provide limited or non-persistent logging, or require additional configuration to enable meaningful telemetry. These constraints create blind spots where malicious activity may go undetected or leave incomplete forensic artifacts.

Organizations can reduce this risk by:

- Prioritizing rapid patching of edge devices
- Enabling centralized and high-fidelity logging
- Enforcing strong authentication
- Integrating network-level visibility with threat intelligence to detect and respond to activity occurring outside traditional endpoint coverage

The threat landscape in 2026

Emerging threats (continued)

3 AI as a force multiplier for malicious activity

AI is rapidly reshaping the threat landscape by lowering barriers to entry and compressing the economics of cyber operations. Tasks that historically required specialized expertise are increasingly being augmented or automated by AI systems. This shift reduces development timelines for both sophisticated actors and opportunistic adversaries. Generative AI also enables a high-volume expansion of convincing, personalized social engineering campaigns in areas that were formerly low-tech but time-intensive, such as romance scams or spear phishing. This shift transforms these fraud schemes into industrialized operations that evade traditional defenses, as most security tools lack visibility into deception occurring outside the endpoint.

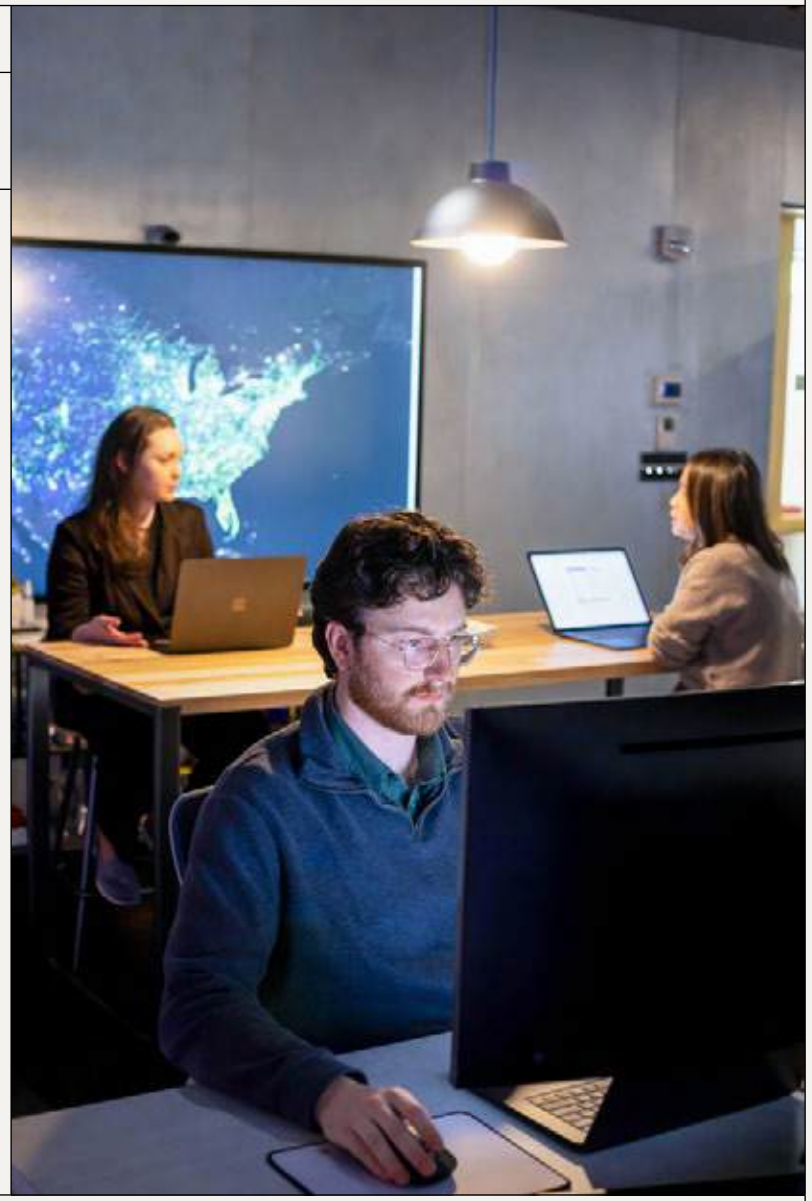
Looking forward, agentic AI will be used to enable increasingly autonomous malicious cyber operations. This suggests a near-term future in which AI-driven agents run large portions of attack campaigns independently from their human operators, further amplifying their scale and impact. Organizations can mitigate these risks by adopting behavior-based detection, strengthening identity and access controls, and implementing robust validation of automated actions.

4 Erosion of trust in the digital information environment

Advances in generative AI have already made it almost impossible for individuals and organizations to distinguish between some authentic and synthetic content. AI-generated text, images, audio, and video can now closely replicate human communication patterns and identities, significantly reducing traditional cues used to assess credibility. This erosion of trust blurs the boundary between legitimate and fabricated content.

As AI-generated content becomes ubiquitous, trust in digital information is quickly degraded. This creates a destabilizing dynamic in which individuals are both more susceptible to manipulation and less confident in legitimate sources. Organizations can mitigate this risk by adopting layered approaches that combine content provenance and verification mechanisms and AI-assisted detection of synthetic media.

At a societal level, mitigating this challenge requires a shift from implicit trust in digital content to explicit verification, supported by trusted information sources and emerging approaches to content authenticity. Policymakers should support interoperable standards for content provenance and authentication while encouraging platforms to preserve and surface reliable indicators of origin. Policy frameworks should also promote transparency, research access, and cross-sector coordination without treating any single detection tool as a complete solution.



Nation-state and *geopolitical threats*

Regional sample of activity level observed

July 2025 – June 2026

Regional top 10s



Americas		Asia & Pacific		Europe		Middle East & Africa	
United States	531	Taiwan	193	United Kingdom	76	Israel	210
Brazil	27	Japan	172	Germany	45	United Arab Emirates	50
Canada	26	India	58	Ukraine	42	Egypt	30
Chile	9	Australia	27	France	28	Saudi Arabia	24
Colombia	9	Hong Kong SAR	25	Italy	25	South Africa	14
Mexico	9	Vietnam	18	Spain	25	Kuwait	8
Peru	7	Malaysia	16	Belgium	24	Qatar	5
Ecuador	5	Philippines	14	Switzerland	20	Jordan	4
Argentina	4	Thailand	13	Greece	14	Ethiopia	3
Dominican Republic	3	Singapore	10	Poland	11	Iraq	3

Observed event activity count per country

0-49 | 50-99 | 100-199 | Over 200

Significant trends in state-sponsored threat activity

Nation-state cyber activity reflects consistent objectives with continued tradecraft refinement.

Across China, Iran, North Korea, and Russia, state-sponsored cyber operations continue to evolve toward more persistent, scalable, and strategically aligned access models, with a shared emphasis on long-term positioning inside high-value environments. While each nation maintains distinct priorities shaped by its geopolitical objectives, collectively, their operations increasingly share characteristics that complicate detection and response. A consistent trend across all four actors is the shift toward maintaining persistent access that can support multiple objectives over time. China, Iran, and Russia, for example, seek access within strategic sectors such as critical infrastructure from which they can pivot from espionage to disruptive activity. North Korea's remote IT workers, meanwhile, generate revenue while establishing access that could be leveraged for intelligence collection or future disruptive activity.

The growing centrality of identity and trusted access pathways is reflected in nation-state activity as well. This includes embedding operations within legitimate authentication flows, using account compromise and cloud-based collection, and leveraging high-credibility social engineering. Chinese and Russian actors are increasingly exploiting identity infrastructure directly, for example through token replay and service principal abuse. Meanwhile, North Korean activity more often relies on high-credibility phishing to obtain valid access.

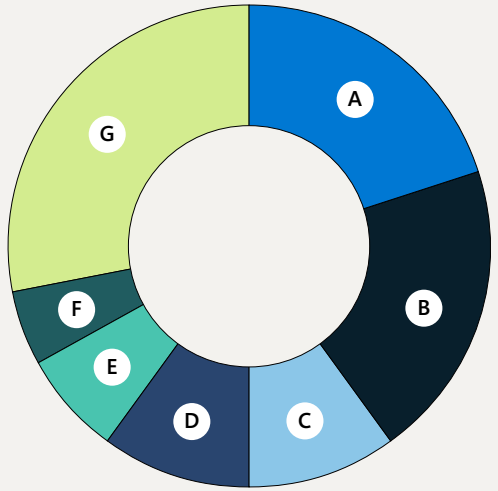
At the same time, these actors diverge in how they balance scale versus precision. For example, China operates at global scale, prioritizing infrastructure-level access and broad visibility across a wide range of sectors. In contrast, Russia often pursues more targeted alignment to support its objectives—particularly in relation to the war in Ukraine. These actors also differ in their tolerance for exposure. China's operations tend to prioritize stealth to enable long-term intelligence collection. Russia and Iran have demonstrated a greater willingness to accept exposure in pursuit of disruptive or destructive effects, though these patterns vary across individual actors and sub-clusters.

Going forward, nation-state cyber activity will likely continue evolving to emphasize access that can be sustained, scaled, and repurposed over time. State-sponsored intrusions are likely to prioritize identity systems—particularly cloud identity systems—as enterprise environments continue to consolidate access, authentication, and administrative control within cloud-based and federated platforms. In addition, all four nations are likely to continue pursuing supply chain attacks to achieve high-impact, scalable access, and incorporate AI across the intrusion lifecycle, integrating more autonomous systems to support elements of their operations.

Nation-state targeting by sector %

A	IT	20%
B	Research and academia	20%
C	Government	10%
D	Think tanks/NGOs	10%
E	Transportation	7%
F	Manufacturing	5%
G	Other	28%

Source: Microsoft Threat Intelligence, nation-state notification data



> Across China, Iran, North Korea, and Russia, state-sponsored cyber operations continue to evolve toward more persistent, scalable, and strategically aligned access models, with a shared emphasis on long-term positioning inside high-value environments.

Nation-state and geopolitical threats

Significant trends in state-sponsored threat activity (continued)



China

Persistence as a foundation

Chinese state actors have strengthened their focus on diverse, large-scale information collection through long-lived access to targets. This supports their government's international objectives and whole-of-state intelligence fusion, an approach that combines political oversight, military apparatus, and commercial industries into a single party-led apparatus. These actors have leveraged living-off-the-land (LOTL) techniques, covert networks, and cloud-native capabilities to minimize footprints and maximize dwell times. While tactics like exploiting vulnerabilities in public-facing applications, uploading web shells, and deploying remote access trojans (RATs) remain in many actors' arsenals, an increasing number of actors are shifting to identity-based attacks, cloud-native persistence, and data exfiltration.

The goals of Chinese state actors remain largely consistent over time and center on long-term access to facilitate espionage. These actors continue to prioritize collection at IT and government organizations linked to the United States, Taiwan, and Japan. Using cloud-native persistence and exfiltration through compromised or actor-created service principals and OAuth applications allows Chinese state actors to evade detection for longer periods.

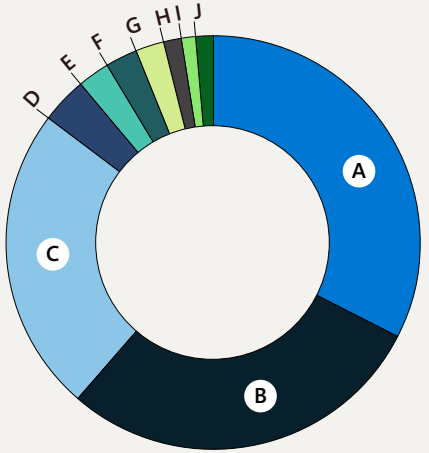
At the same time, Chinese state actors continuously search for exploitable vulnerabilities in public-facing applications, which can provide initial access and sometimes authentication material to facilitate lateral movement within the network or into the cloud. Chinese state actors experiment with new technologies as they gain adoption, fitting them into existing workflows. Some actors now use AI tools to search for vulnerabilities or tips on exploiting them. Other actors continue to deploy tried and tested attack chains using phishing and RATs to compromise a wide set of targets.

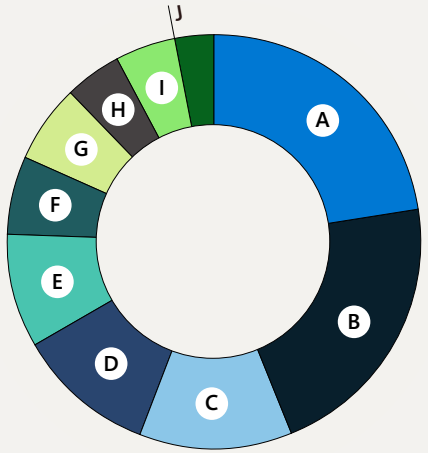
Top ten targeted countries: China	%
A United States	27
B Taiwan	24
C Japan	20
D Hong Kong SAR	3
E Philippines	2
F Malaysia	2
G Italy	2
H United Kingdom	1
I Germany	1
J Belgium	1

Source: Microsoft Threat Intelligence, nation-state notification data

Top ten targeted sectors: China	%
A IT	15
B Government	14
C Research and academia	8
D Think tanks/NGOs	7
E Critical manufacturing	6
F Transportation	4
G Consumer retail	4
H Finance	3
I Defense	3
J Healthcare and public health	2

Source: Microsoft Threat Intelligence, nation-state notification data





Nation-state and geopolitical threats

Significant trends in state-sponsored threat activity (continued)



Iran

Pursuing resilient and opportunistic operations

In a period marked by regional tension and episodic kinetic escalation, Iranian state-aligned actors conducted broad, multi-layered campaigns against Israel and longstanding adversaries across the Middle East. They also carried out a notable and significant increase in targeting of the United States and US interests. As the conflict evolved, Iranian cyber and influence operations expanded their targeting, adapted disruptive techniques, and increasingly integrated cyber, military, and information activities to amplify strategic messaging and operational impact.

While activity against regional targets continued at its usual steady pace, the expansion of US-focused operations was a largely retaliatory response to the war, and drew in actors that had not previously targeted the United States. Internet blackouts and infrastructure degradation in Iran this year temporarily constrained some operations and prompted connectivity adaptations, but overall Iran's tempo remained high, with pre-positioned and opportunistic access enabling sustained activity.

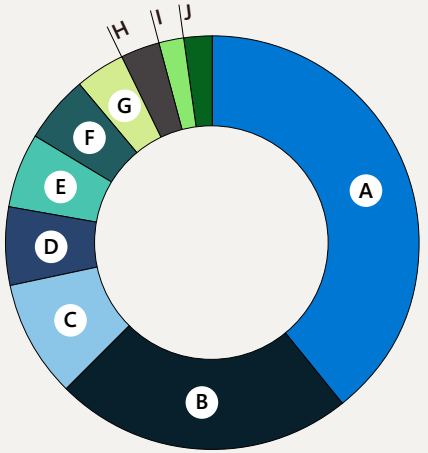
Iranian actors continue to target the same sectors, with a consistent focus on research and academia, IT/telecommunications, critical infrastructure, government and defense supply chains, non-governmental organizations (NGO), and dissident communities. At the same time, Microsoft continues to observe pre-positioning within these environments for long-term espionage and potential disruption. Ongoing regional conflict has driven a shift toward more disruptive and destructive approaches, including data destruction and operational technology (OT) targeting for retaliation.

As we first noted last year, we continue to see a convergence in tactics, techniques, and procedures (TTP) across Iranian threat actors. This coordination amplifies the effectiveness and scale—and therefore potential impact—of Iranian campaigns. Iranian state-aligned threat actors continue to rely primarily on exploitation of known vulnerabilities, weak authentication, and exposed internet-facing systems for initial access, with credential harvesting and phishing as central enablers. Their post-compromise goals include lateral movement to sustain access and expand operational reach, with abuse of cloud infrastructure, including Microsoft Azure, for persistence.

Top ten targeted countries: Iran

		%
A	Israel	39
B	United States	23
C	United Arab Emirates	9
D	Egypt	6
E	India	6
F	United Kingdom	5
G	Brazil	4
H	Greece	3
I	Germany	2
J	Saudi Arabia	2

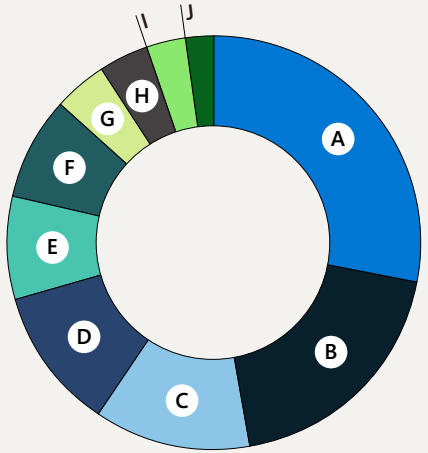
Source: Microsoft Threat Intelligence, nation-state notification data



Top ten targeted sectors: Iran

		%
A	Research and academia	28
B	Other	19
C	IT	12
D	Transportation	11
E	Government	8
F	Think tanks/NGOs	8
G	Communications	4
H	Manufacturing	4
I	Commercial facilities	3
J	Finance	2

Source: Microsoft Threat Intelligence, nation-state notification data



Nation-state and geopolitical threats

Significant trends in state-sponsored threat activity (continued)

Russia

More diversity in tradecraft for stealth and efficiency

Russian state actors are becoming more efficient, stealthy, and adaptive, using LOTL techniques, reduced malware footprint, and deeper exploitation of identity and cloud ecosystems. The result is a threat landscape that is more persistent and harder to detect, but at the same time, predictable. Russian threat actor TTPs have become more diversified even as they targeted the same sectors as last year: NGOs, government, research and academia, and defense. Geographically, the United States, United Kingdom, and Ukraine remain these actors' primary focus, with other NATO members—such as Spain, Germany, and Belgium—taking secondary precedence.

Russian threat actors are leaning more heavily into identity-based attacks and cloud exploitation to bypass defenses and maintain persistence, including abusing OAuth applications and conducting token replay. Some actors have leveraged AiTM capabilities at the network edge and internet service provider (ISP) level, enabling selective traffic interception. These approaches do not replace the use of malware.

Instead, they augment traditional intrusion chains, allowing operators to maintain access through quieter, identity-centric mechanisms that blend into the target environment.

Some Russian actors are also targeting internet-facing edge devices to identify vulnerabilities and exploit misconfigured devices. In a notable escalation, in December 2025 a Russian threat actor used this access vector when conducting a coordinated destructive attack on Poland's energy sector. This indicates a new willingness to destroy critical infrastructure outside of Ukraine while possibly also serving as a warning to countries that support Ukraine.

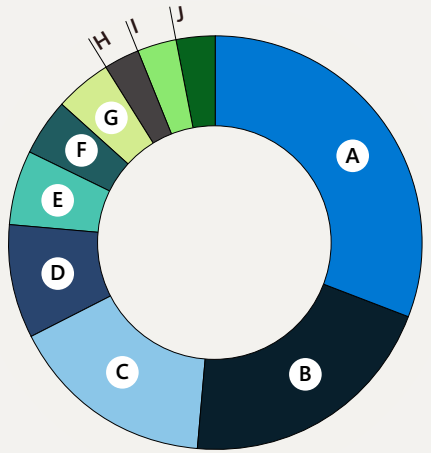
This year, we observed an uptick in cyber activity from non-traditional actors as a way to scale operations, reduce attribution risk, and accelerate access. This reinforces the distributed nature of Russia's hybrid ecosystem, which blends traditional intelligence units with non-traditional actors including emerging clusters, cybercriminal affiliates, and hacktivist groups. We also identified Russian actors using vibe coding or AI-generated tooling. For now, however, Russian integration of AI across the attack lifecycle serves as a force multiplier to improve scale and speed rather than fundamentally changing attack methods.

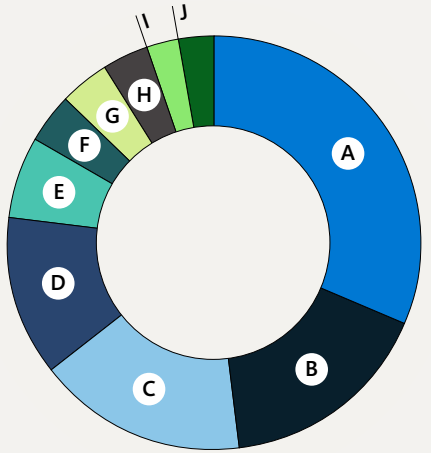
Top ten targeted countries: Russia	%
A United States	21
B Ukraine	14
C United Kingdom	11
D Belgium	6
E Germany	4
F France	3
G Spain	3
H Poland	2
I Sweden	2
J Switzerland	2

Source: Microsoft Threat Intelligence, nation-state notification data

Top ten targeted sectors: Russia	%
A Government	25
B Education	13
C Think tanks/NGOs	13
D IT	10
E Energy	5
F Defense industry	3
G Manufacturing	3
H Transportation	3
I Finance	2
J Inter-governmental organiztion	2

Source: Microsoft Threat Intelligence, nation-state notification data





Nation-state and geopolitical threats

Significant trends in state-sponsored threat activity (continued)

North Korea

Scaling with the help of AI

North Korean state actors have innovated their tradecraft, increased their use of AI, and scaled infrastructure to accelerate operations and conduct more persistent and adaptive campaigns. While their tradecraft remains grounded in social engineering, malware delivery, and abuse of legitimate tools, it has evolved toward access-focused and identity-driven operations.

North Korean threat actors continue to prioritize organizations in the United States, United Kingdom, Europe, and East Asia, focusing on IT, financial services, and blockchain-related organizations to support revenue generation. At the same time, they engaged in sustained targeting of government, defense-related, and policy focused organizations for intelligence collection. Recently, these actors' targeting has broadened to reflect a more scale-driven and opportunistic approach. For example, they have expanded targeting across the cryptocurrency ecosystem to include Bitcoin mining organizations and increased targeting of individuals as access vectors. They also sought access to software and supply chain ecosystems. In March 2026, a state-sponsored group compromised the widely used Axios npm package in a supply chain attack.

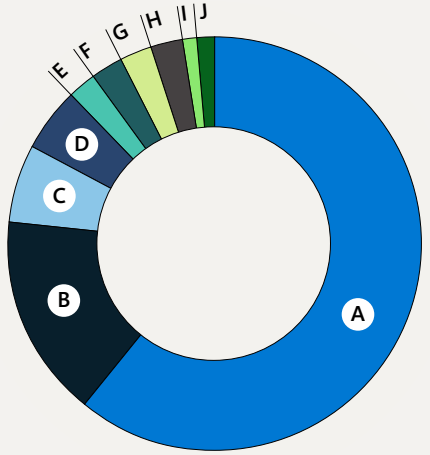
The North Korean remote IT worker scheme remains a major threat as workers exploit global hiring practices to obtain trusted access across industries. The program has also leveraged AI across the attack lifecycle to support persona development, social engineering, and sustained access. More broadly, North Korean threat actors are integrating AI to improve the speed, scale, and resilience of operations including malware creation and infrastructure management. In some cases, actors have also experimented with agentic workflows and LLM-enabled code generation to accelerate malware deployment.

North Korean actors are also expanding their tradecraft to support identity-based attacks. This includes large-scale OAuth consent phishing campaigns while impersonating legitimate SaaS platforms to induce users to grant high-privilege permissions. These operations continue to emphasize rapid infrastructure turnover, including the creation and rotation of domains and supporting infrastructure to sustain operations and evade disruption. This allows campaigns to persist even as infrastructure is identified and removed.

Top ten targeted countries: North Korea

		%
A	United States	50
B	Italy	13
C	Australia	5
D	United Kingdom	4
E	Switzerland	2
F	India	2
G	Germany	2
H	United Arab Emirates	2
I	France	1
J	Republic of Korea	1

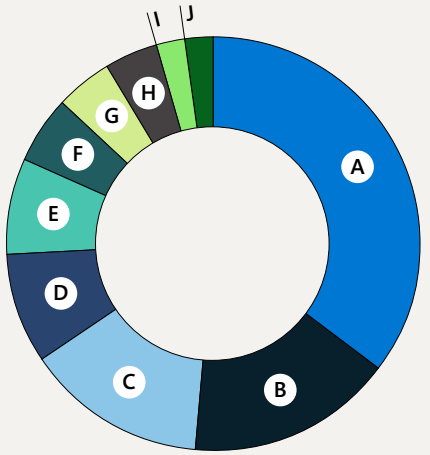
Source: Microsoft Threat Intelligence, nation-state notification data



Top ten targeted sectors: North Korea

		%
A	IT	33
B	Research and academia	15
C	Other	13
D	Think tanks/NGOs	8
E	Consumer retail	7
F	Finance	5
G	Manufacturing	4
H	Health	4
I	Communications	2
J	Defense industry	2

Source: Microsoft Threat Intelligence, nation-state notification data



How modern *intrusions begin*

The initial access landscape

Looking at Microsoft Defender Experts notifications this year reveals an initial access landscape dominated by human-centric and identity-centric entry points.

Initial access techniques observed in Microsoft Defender Experts customer notifications

The initial access landscape is dominated by human-centric and identity-centric entry attempts and intrusions, according to a review of Microsoft Defender Experts notifications this year. Combined, user execution and valid accounts accounted for over half of all the activity observed, while phishing and malicious copy and paste made up another quarter. This means prevention, detection, and response programs need to be equally strong across endpoint execution controls, identity protection, phishing resistance, and browser defenses. No single control layer is sufficient.

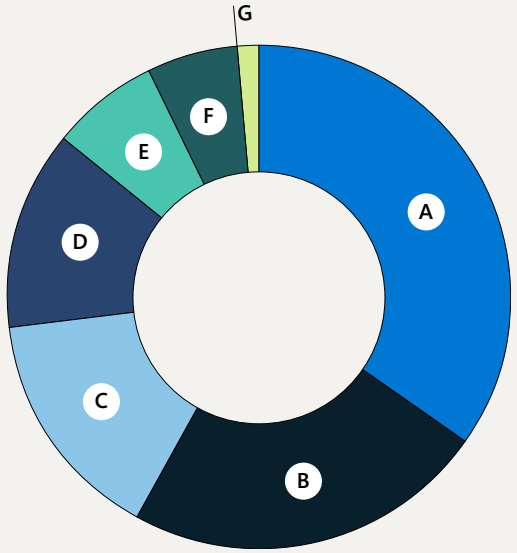
Today's user execution intrusions arrive through fake browser updates, AI-generated lures, multi-stage payloads that evade sandboxes, and, increasingly, social engineering that bypasses traditional file-based thinking entirely. Once a user executes, the follow-on is swift, leading to:

- Malware (in 60.3% of cases)
- Infostealers (in 12.3% of cases)
- Credential theft (in 9.4% of cases)

PowerShell and cmd.exe are primary execution vehicles in these pathways.

Initial access techniques

		%
A	User execution	30
B	Valid accounts	20
C	Malicious copy and paste	13
D	Phishing	11
E	Drive-by compromise	6
F	Exploitation of public-facing application	5
G	Other	1



Source: Microsoft Defender Experts

From valid accounts to enterprise-wide access

In access based on valid accounts, attackers authenticate with legitimate credentials: stolen passwords, successful password sprays, replayed session tokens, or hijacked accounts.

What makes identity-based access uniquely dangerous is what happens next. Over half of these intrusions (52.2%) involved follow-on credential theft—once inside, attackers immediately harvested more credentials. Another 18.4% involved active password spray campaigns. This creates a self-reinforcing cycle: one compromised identity fuels the compromise of many more. Lateral movement is also pronounced—adversaries pivot through Server Message Block (SMB) shares, Windows Management Instrumentation (WMI), and Remote Desktop Protocol (RDP). Credential dumping is then used to escalate access quickly.

The rise of malicious copy-and-paste attacks

Last year, ClickFix-style attacks, which use deceptive prompts to trick users into copying, pasting, and executing malicious commands, represented 47% of observed initial access activity. This year's data reflects a more diversified threat landscape, with malicious copy-and-paste attacks accounting for only 12.8% of observed intrusions.

ClickFix attacks waned in late 2025, then rebounded explosively in early 2026, growing roughly 23x between December and May. The follow-on is almost uniformly malware (96.3%), making this one of the most conversion-efficient initial access techniques observed. The reusability of the ClickFix access layer means operators can rotate payloads without altering the social engineering delivery, making it a cross-platform, behavior-driven initial access technique that routinely bypasses file- and exploit-based defenses.

ClickFix is evolving

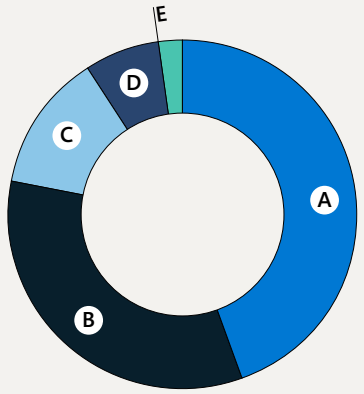
In ClickFix-style attacks, targets are guided through fake CAPTCHAs, fabricated error dialogs, or "verify you are human" lures to paste an attacker-supplied command into Windows Run, Windows Terminal, or PowerShell.

The result is a low-friction, high-yield foothold that scales to hundreds of thousands of devices. Between February and early May 2026, Microsoft Defender observed ClickFix-style user-pasted command execution on more than 1.1 million unique devices, an ~8x increase. This was likely driven by a proliferation of kits-for-hire. ClickFix builders are sold as malware-as-a-service (MaaS), lowering the technical bar for new operators.

We also saw a "FileFix" variant gaining traction. In FileFix, users are guided to paste the malicious command into the Windows File Explorer address bar instead of Run or Terminal. This broadens the target pool by accessing less technically cautious targets and bypasses controls that focus solely on PowerShell/Run telemetry.

Phishing technique breakdown

Key		%
A	AiTM-URL	45
B	Standard URL	34
C	Attachment	13
D	AiTM-attachment	7
E	QR code (incl. AiTM-QR)	2



Source: Microsoft Threat Intelligence

Phishing has evolved past passwords

Phishing accounts for 10.9% of intrusion attempts, but the nature of phishing has fundamentally changed. AiTM phishing now dominates. Here, attackers use a reverse proxy to secretly sit between a user and a real sign-in page. It captures usernames, passwords, and MFA codes in real time, allowing the attacker to steal the active session cookie and circumvent MFA. AiTM now accounts for 44.6% of all identified phishing techniques, compared to 33.6% for standard URL phishing and just 12.9% for traditional attachments. The downstream impact of AiTM phishing is significant—87.7% of phishing intrusions involved credential or session harvesting, while 6.3% were business contact impersonation (BCI).

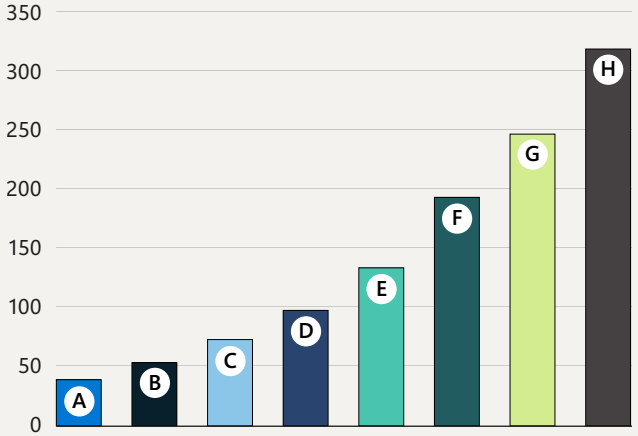
Public-facing exploitation: Low volume, high impact

While exploitation of public-facing applications represents just 4.8% of observed intrusions, this initial access vector carries disproportionate impact. Microsoft SharePoint Server is the most frequently exploited product, followed by Internet Information Services (IIS), HTTP File Server (HFS), React-based applications, and Exchange Server. Compromising an internet-facing server often serves as an entry point into an organization, allowing attackers to rapidly broaden their access and target high-value systems and data. Notably, exploitation volume declined sharply from mid-2025 onward, suggesting a mix of successful patching and attacker pivot to identity-based methods.

Volume trajectory

(Microsoft Defender telemetry)

Unique devices per week running an attacker-supplied ClickFix command, observed in Microsoft Defender behavioral telemetry.



Key		000's
A	Feb W1	38,000
B	Feb W2	52,000
C	March W1	71,000
D	March W2	95,000
E	April W1	130,000
F	April W2	188,000
G	April W5	240,000
H	May W1	310,000

Source: Microsoft Threat Intelligence

ClickFix is evolving (continued)

What this means for defenders

These trends reinforce that no single control layer is sufficient. Defenders should combine phishing-resistant identity controls, endpoint and browser protections, behavior-based detection, and rapid containment playbooks to stop an initial compromise from becoming an enterprise-wide incident.

- **Strengthen the human layer.** Use role-based training that prepares users to recognize phishing, ClickFix-style prompts, and unusual requests to run commands or share credentials.
- **Move beyond traditional MFA.** Prioritize phishing-resistant authentication and monitor for session-token theft, anomalous sign-ins, and unfamiliar devices.
- **Protect identities after initial access.** Detect credential harvesting, privilege escalation, suspicious OAuth activity, and lateral movement before one compromised account leads to broader access.
- **Use layered controls.** Combine identity protection, endpoint and browser defenses, application controls, rapid patching, and centralized telemetry so attackers cannot bypass the environment through a single weak point.
- **Prepare for the common post-compromise playbook.** Establish response procedures to revoke sessions, isolate devices, reset credentials, and investigate related accounts and systems quickly.

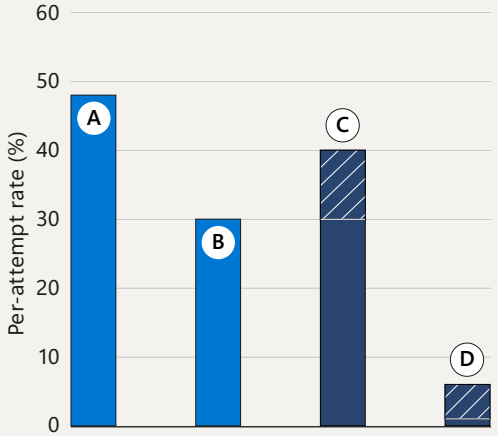
Voice phishing attacks go mainstream in chat applications

Threat actors targeting employees through collaboration platforms are increasingly using voice phishing (vishing) for higher-confidence, multi-step social engineering. Vishing over cross-tenant calls (calls that cross the boundaries of at least two organizational environments) in Microsoft Teams in particular has become a more prominent access vector.

The weekly confirmed malicious volume has risen 502% from this time last year. Over 90% of attackers that use Teams have incorporated voice calling into some of their attacks. Despite red flags such as Teams' "External" banners and call warnings, skilled attackers successfully persuade victims to comply, usually on workdays during business hours.

Across the attacks we studied, 93% of attackers succeeded at least once in keeping a target on the line for 20 seconds or longer—the point at which an attacker has captured attention and can begin a social engineering script. Nearly half of all calls (48%) lasted longer than 20 seconds, and 30% lasted beyond one minute, during which time targets are likely actively talking with attackers and either following instructions or leaking privileged information. While Safe Links evaluates URLs shared during these calls at time of click, blocking credential harvesting pages even mid-conversation, the voice channel itself carries no link to intercept.

Per-attempt engagement: Vishing vs. email phishing



Key		%
A	Vishing victim engaged over 20 seconds	48
B	Vishing victim engaged over 1 minute	30
C	Targeted email victim engaged/ clicked links	30-40
D	Bulk email victim engaged/ clicked links	1-5

Vishing: Microsoft telemetry | Email: Microsoft and academic telemetry

Source: Microsoft Threat Intelligence

How modern intrusions begin

Voice phishing attacks go mainstream in chat applications (continued)

Attackers are diversifying their vishing impersonation techniques

While "Help Desk" and "IT Support" variants used to account for most malicious display names, a majority now use personal-looking usernames that either match or appear to match legitimate employee naming conventions. Unicode evasion has also matured: invisible characters, emojis, and deliberate misspellings allow attackers to bypass exact-match filters while keeping display names visually similar to legitimate support contacts.

When a vishing call succeeds, what follows is rapid and rehearsed. The attacker convinces the victim to grant access through a remote monitoring and management (RMM) tool, achieving interactive control of the endpoint. Attackers complete reconnaissance in two minutes or less and then download malicious payloads.

Organizations should use a layered approach to reduce the risk of vishing attacks in collaboration platforms.

- **Teams federation.** Ensure external-user warnings are enabled so users see clear "External" indicator.
- **Safe Links.** Enable time-of-click URL evaluation without an end-user click-through override to prevent credential harvesting through links shared during calls.
- **RMM tools.** Limit remote assistance tools to IT-only groups through endpoint management policy, removing a primary foothold mechanism.
- **Attack surface reduction.** Enable attack surface reduction rules in block mode for script and Office macro execution to disrupt staging and payload delivery.
- **Endpoint detection.** Deploy endpoint detection and response (EDR) capabilities with behavioral blocking or logging to connect the social engineering event to endpoint compromise in a single timeline.
- **Threat hunting.** Hunt for Teams call events followed by remote management process launches within a 30-minute window to identify the transition from vishing to compromise.



The cost of delayed patching

In an age of AI-powered vulnerability scanning, exploitation of unpatched vulnerabilities is a concerning initial access path.

Nation-state and ransomware actors keep weaponizing disclosed flaws, especially internet-facing infrastructure, identity and management appliances, developer tooling, and third-party software.

The disclosure-to-exploitation window has shrunk to single-digit days, if not exploited as a zero-day before any advisory is issued. Microsoft has observed the following trends:

- **More, faster weaponization.** The Cybersecurity and Infrastructure Security Agency (CISA) added over 110 CVEs to the Known Exploited Vulnerabilities (KEV) catalog from November 2025 to May 2026, most within a week of disclosure. Attackers are using CVEs with lightning speed. For example, cybercriminal Storm-1175 moved from web-facing exploitation to deploying Medusa ransomware in as little as 24 hours. Since 2023, this ransomware group has weaponized at least 16 CVEs, including SAP NetWeaver, a day after disclosure.
- **Concentration on edge, identity, and management infrastructure.** Ransomware is continuing to find fast lanes through:
  1. RMM platforms, for example SimpleHelp, ScreenConnect, and BeyondTrust
  2. Identity file transfer products, for example Oracle Identity Manager, SmarterMail, GoAnywhere MFT, and SolarWinds Web Help Desk

How modern intrusions begin

The cost of delayed patching (continued)

In September 2025, Microsoft tracked a four-week surge in Akira ransomware across 50+ organizations—mostly small and medium-sized businesses—tied to CVE-2024-40766 (SonicWall SSL VPN exploitation).

- **Pre-disclosure (zero-day) exploitation.** Significant risk exists that threat actors will find vulnerabilities before blue teams and researchers. Cisco Secure FMC's critical remote code execution flaw (CVE-2026-20131) was given the maximum score of 10.0 for severity under the Common Vulnerability Scoring System (CVSS). The Interlock ransomware group was able to exploit this vulnerability for approximately 30 days before disclosure.
- **Multi-actor use of CVEs.** As actors seek means of entry, they occasionally leverage the same vulnerabilities. For example, Microsoft saw coordinated exploitation of CVE-2025-49704 and CVE-2025-49706 against on-premises SharePoint in July 2025. Two Chinese state actors used the ToolPane chain for access, while Storm-2603 deployed Warlock ransomware via web shell and MachineKey theft. This means that the same n-day window was used for both espionage and ransomware within days.

The exploitation gap is narrowing at a time when patch coverage is lagging. Research published in 2025 reports a median time-to-exploit of about five days. For internet-exposed services, mass scanning often starts within hours, and pre-disclosure windows of 30 days or more are now routine. The gap between government and enterprise remediation times is stark. While KEV remediation timelines require federal action within 7–14 days, median enterprise remediation for critical external CVEs can take 30–60 days.

Shadow IT and end-of-life (EoL) devices widen it further. This mismatch means attackers often have weeks to exploit known vulnerabilities after fixes are available. The issue is no longer simply whether a patch exists, but whether organizations can identify exposed assets, prioritize the highest-risk flaws, and remediate them before attackers operationalize them. For internet-facing, identity, and management systems, traditional monthly patch cycles may no longer be fast enough.

> "In an age of AI-powered vulnerability scanning, exploitation of unpatched vulnerabilities is a concerning initial access path.

Recommendations

Treat newly exploited vulnerabilities as urgent.

Fix newly identified vulnerabilities that affect an internet facing or identity system within 72 hours.

Know what needs to be patched.

Maintain a complete, current inventory of third party systems, remote management tools, and older devices so exposed assets are not overlooked.

Monitor critical access systems.

Extend security monitoring to firewalls, VPNs, identity systems, RMM tools, email gateways, and file transfer services.

Add safeguards when systems cannot be patched.

Isolate outdated systems from the wider network, restrict unnecessary connections, and use protective controls to block suspicious traffic.

Look for signs of past exploitation.

Review the previous 90 days of activity for evidence that newly disclosed or previously unknown vulnerabilities were used before fixes were applied.

Impersonation is becoming more sophisticated

Impersonation remains one of the most effective phishing techniques because it makes malicious messages feel legitimate by borrowing the identities, workflows, and visual cues people already trust. Microsoft analysis found that nearly 85% of reported phishing emails between April and June 2026 carried an impersonation signal.

For years, impersonation defenses focused on recognizable signals: lookalike domains, spoofed display names, authentication failures, well-known brand names such as Microsoft, Google, DocuSign, PayPal, and Adobe, and fuzzy matches against curated entities. 32% of attacks use images or logos to mimic a brand. Those signals still matter, but attackers are adapting. More than a third of attacks use template-based lures that mimic familiar brand communications, frequently without explicitly naming any brand to avoid detection, while 44% of targeted brands are lesser-known or location-specific brands that fall outside globally optimized detection models.

> **Around 25% of attacks mimic an organization's trusted operating environment or someone inside it, with HR, payroll, and benefits scams driving 77% of those organizational cases.**

The larger shift is from brand mimicry to context mimicry

Attackers are impersonating relationships, departments, workflows, vendors, and organizational routines. Automated notifications appear in 80% of impersonation attacks, followed by financial fraud themes (61%). Around 25% of attacks mimic an organization's trusted operating environment or someone inside it, with HR, payroll, and benefits scams driving 77% of those organizational cases.

The effectiveness of these campaigns comes not only from who the sender appears to be, but from how naturally the request fits into the recipient's workday.

That same familiarity abuse extends beyond the organization itself. Attackers also pose as trusted collaborators, suppliers, or business partners, exploiting the fact that large, global organizations interact with a wide ecosystem of vendors that may not appear among their most visible or frequently contacted collaborators. Although these cases represent only about 3% of spear-phishing activity, they are challenging to detect because the credibility comes from a plausible relationship rather than from a known top vendor or explicit brand signal.

Once attackers establish that familiar context, the next step is to convert trust into action. The lure may point to a credential harvesting sign-in page, deliver a malicious attachment, ask the recipient to reply with sensitive information, or apply pressure through callback scams and gift card requests.

These tactics can lead to account compromise, financial loss, and broader organizational breach. Even nation-state actors have adopted impersonation as an initial engagement tactic, using persistent contact with targets to enable later-stage intrusions.

At the same time, delivery is diversifying: while attachments remain common (30%), QR codes now appear in 13% of cases, and HTML-based payloads in 5%. One in four impersonation emails pushes the victim to a phone call, moving the attack outside normal URL and attachment inspection and into a live social engineering channel.

Organizations can reduce impersonation risk by combining technical controls with verification practices. Train employees to scrutinize unexpected requests that appear to match familiar workflows, and require independent confirmation for payment, payroll, credential, and sensitive data requests. Extend detection beyond prominent brands to local vendors, partners, departments, and organizational naming patterns. Inspect QR codes, HTML attachments, and callback requests that move interactions outside traditional email defenses, and make it easy for employees to report suspicious messages before acting.

Business contact impersonation: The new "BEC"

The term business email compromise (BEC) has long suffered from definitional confusion. To some, it refers to text-based social engineering when an attacker impersonates a CEO and asks for a wire transfer.

To others, it encompasses the actual compromise of a business email account through credential phishing or session hijacking. These are fundamentally different threats with different attack chains, defenses, and indicators. Conflating them under one label obscures both.

Moving forward, we are introducing the term business contact impersonation (BCI) to describe social engineering attacks that initiate a conversation with enterprise users while impersonating a trusted business contact, such as an executive, employee, vendor, supplier, or partner, to persuade the recipient to take an action under false pretenses, such as initiating a fraudulent financial transaction or sending sensitive information or documents to the attacker. This definition deliberately excludes credential phishing and account takeover (ATO), which are distinctly different threats that warrant their own terminology and their own defensive strategies.

Over the past 12 months, Microsoft detected more than 46 million BCI attacks. Monthly volumes were relatively stable through most of the period before spiking in April 2026, when volume surged 121%—more than double any previous month observed in the past year. The most consistent finding across the period is how rarely the initial BCI email actually asked for anything. The vast majority of first-contact BCI messages contain only generic outreach, such as "Are you at your desk?" or "Do you have a moment?" with no mention of money, invoices, or documents.

Only a small percentage of initial BCI messages contained an explicit financial or document request, and that share has been shrinking: from 17% in October 2025 to just 3% by June 2026. BCI is, overwhelmingly, a conversation starter. The fraud comes later, in follow-up messages, once the attacker has established rapport and confirmed the target is engaged. This multi-stage structure means the initial email is deliberately bland, designed to avoid triggering content-based detection rules precisely because it contains nothing overtly malicious to detect.

One interesting trend within the BCI landscape worth highlighting is payroll diversion, when actors attempt to redirect an employee's direct deposit to an attacker-controlled account. These attacks follow a striking weekly cadence: volume peaks sharply on Monday, then declines steadily each day through Friday before going nearly silent on weekends. Over half of all payroll diversion attacks (51%) land on a Monday or Tuesday. This suggests attackers are deliberately timing their outreach to coincide with the start of the business week, when payroll and HR staff are processing requests from the prior week and recipients are most likely to action a quick ask without much scrutiny.



Emerging threat of network edge infrastructure

Network edge infrastructure—such as firewalls, VPNs, and gateways—was used as an initial access vector before attackers pivoted to endpoints or identity systems. Now attackers have pivoted back. Modern edge devices enforce authentication, manage sessions, inspect traffic, and integrate with cloud identity platforms, making them control points. Compromise enables direct control over identity, trust, and policy enforcement across the enterprise.

Attackers are targeting access before identity verification

Sophisticated attackers are increasingly prioritizing pre-authentication vulnerabilities that allow access before identity verification. Authentication bypass and RCE provide immediate entry into systems that enforce trust. These flaws are rapidly weaponized and exploited at scale, often within hours or days of disclosure, with campaigns targeting exposed VPN infrastructure and extending into identity abuse.

How modern intrusions begin

Emerging threat of network edge infrastructure (continued)

Once access is established, adversaries persist directly within network devices rather than passing through them. Techniques include configuration changes, unauthorized accounts, credential insertion, and manipulation of authentication components. More advanced activity includes firmware-level or in-memory implants that evade forensic detection and survive standard remediation.

Strategic actors increasingly maintain long-term access at the edge using low-noise methods designed to avoid detection. Control of routing and segmentation further enables lateral movement that blends with legitimate activity. Because actions originate from trusted infrastructure, detection is reduced while attackers gain access to segmented or otherwise restricted resources.

> “Strategic actors increasingly maintain long-term access at the edge using low-noise methods designed to avoid detection.

Compromised edge infrastructure functions as an operational platform

Since these systems sit at the convergence of authentication, session activity, and data flows, after achieving access, attackers can monitor authentication interactions, identify high-value accounts, capture credentials and session tokens, and hijack or replay sessions to bypass authentication.

Attackers are shifting from exploiting individual vulnerabilities to targeting entire classes of network technologies. Campaigns span multiple vendors, including Fortinet, Cisco, Ivanti, Citrix, Palo Alto Networks, and Juniper, leveraging shared architectural patterns, such as exposed management interfaces and similar authentication workflows. Once identified, these weaknesses are rapidly operationalized across the ecosystem, generating coordinated exploitation waves rather than isolated incidents.

This activity is reinforced by a persistent defensive visibility gap. Network infrastructure is under-instrumented. Patching is often delayed, and standard remediation actions frequently fail to remove persistence embedded in configurations or system components.

Recommendations

To confront this threat, defenders must extend detection into the network control plane. Effective coverage includes monitoring for large-scale exploitation attempts against edge services:

- Configuration tampering
- Unauthorized account creation
- Credential and secret extraction
- Session hijacking
- Lateral movement initiated by network devices

Defenders also need to focus on identity integration abuse, such as unexpected Security Assertion Markup Language (SAML) or identity provider changes, and pre-breach exposure risks such as unpatched systems and exposed management interfaces. A defining signal of this threat is a network device behaving like an endpoint by initiating scanning, credential access, or lateral movement.

Information stealers: A continued threat

Information stealers are a type of malware designed to secretly gather sensitive data from a device. Over the past year, they have evolved from opportunistic credential grabbers into a broader, more adaptive access-and-monetization layer that often sits at the front of larger intrusion chains.

These campaigns increasingly target credentials, session tokens, and other high-value artifacts that enable ATO and follow-on compromise across endpoints, browsers, and cloud-connected services. There is also a growing emphasis on fast-turn monetization through cryptocurrency wallet and session theft. Many attempted ransomware attacks are disrupted during the infostealer stage, underscoring the value of early detection and intervention.

In parallel, Microsoft Defender Experts has seen increased use of cross-platform payloads, modular scripts, and delivery lures that abuse trusted platforms and common utilities to blend into typical user workflows. Together, these shifts signal a move beyond traditional Windows-focused campaigns toward scalable operations that also target macOS and developer environments with low-friction, stealthy collection at scale.

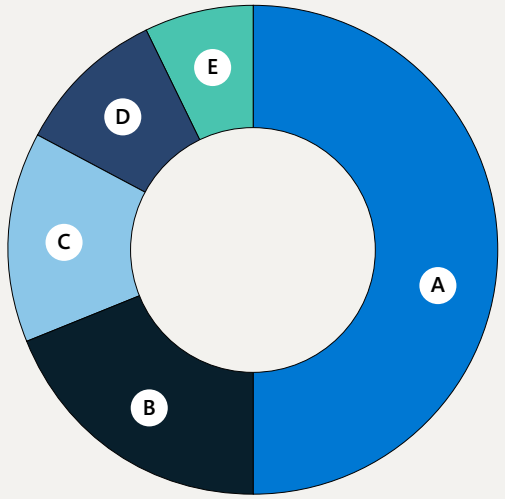
Infostealer activity is becoming more advanced

Since late 2025, Microsoft Defender Experts has observed an increase in macOS-targeted infostealer activity, driven largely by social engineering-based delivery. Campaigns associated with malware families such as Atomic macOS Stealer (AMOS), MacSync Stealer, and DigitStealer rely on user-initiated actions, often disguised as legitimate workflows. These operations are designed to enable ATO, financial fraud, and downstream compromise of cloud and developer environments.
A defining characteristic of these macOS campaigns is their use of native system tooling and fileless execution techniques. This allows attackers to blend into expected platform behavior and reduce visibility compared to conventional malware deployments. Overall, this trend highlights a strategic move by threat actors to align their tradecraft more closely with platform norms, signaling increased maturity and investment in macOS-specific intrusion paths.

Microsoft Defender Experts also continue to observe a broader shift toward cross-functional information-stealing operations, where the initial stealer stage is also used to validate access, profile the target environment, and enable follow-on payload delivery. This model increasingly aligns with nation-state tradecraft, where information theft is tightly coupled with long-term access objectives. In such campaigns, infostealers function as an entry and validation layer within larger intrusion chains, leveraging cross-platform scripts, modular tooling, and developer-focused lures to reach high-value identities and sensitive assets.

Top five stealer families observed in infostealer activity among Microsoft Defender Experts customers

Key		%
A	Atomic macOS Stealer	50
B	OtterCookie	19
C	PS1Bot	14
D	Lumma Stealer	10
E	MacSync Stealer	7



Source: Microsoft Defender Experts



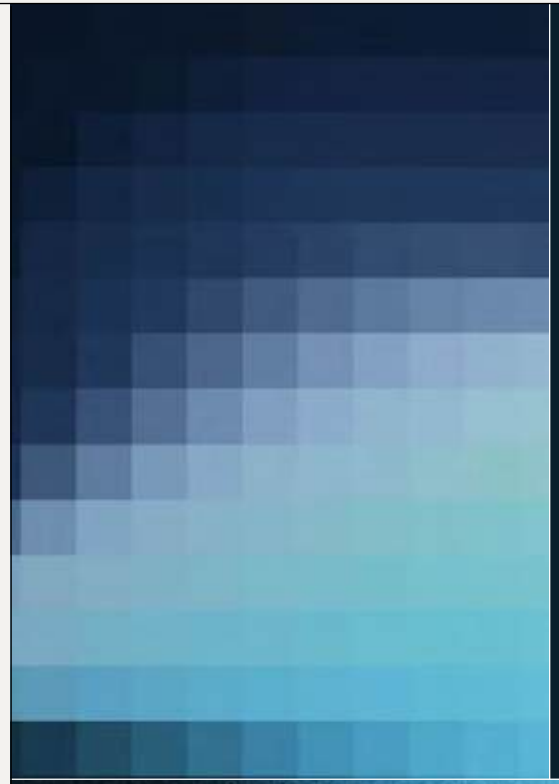
How modern intrusions begin

Information stealers: A continued threat (continued)

Trusted platforms and common utilities are abused for stealthy delivery

Microsoft Defender Experts have observed threat actors increasingly using trusted platforms and familiar consumer utilities as delivery cover, blending theft activity into routine workflows and delaying suspicion. In the CrystalPDF campaign, for example, attackers promoted a fake PDF converter via malvertising and search engine optimization (SEO) poisoning. It established persistence, targeted browser credential stores, and initiated C2 traffic. Similar abuse of messaging platforms and everyday tools shows how common install-and-use behavior is being repurposed to reach financial and cryptocurrency accounts.

These intrusions often start with low-signal artifacts that resemble nuisance software or ordinary installation noise (for example, simple file creation alerts), buying time for attackers to progress into persistence, credential and wallet theft, and C2. Delivery and early-stage behavior are intentionally designed to look routine, making "annoying vs. malicious" distinctions less reliable for defenders.



Recommendations

Prioritize controls that reduce initial execution success, limit credential/token exposure, and shorten time-to-detection for wallet and developer-secret theft.

Harden initial access paths.

Restrict script abuse (PowerShell/JavaScript), block common LOLBins where feasible, and enforce application control for user-writable locations.

Strengthen identity.

Employ phishing-resistant MFA, conditional access, and rapid invalidation/rotation for tokens and session credentials.

Protect developer environments.

Move secrets to managed vaults, scan repos and endpoints for secrets, and enforce short-lived credentials for CI/CD and cloud CLIs.

Reduce crypto wallet exposure.

Prefer hardware wallets where possible, control/monitor wallet extensions, and alert on wallet-related file/extension access from unusual processes.

Detect and contain early.

Watch for infostealer tradecraft (credential store access, archive/exfil patterns, Telegram/C2 beacons), and isolate hosts quickly to prevent follow-on payloads.

Be ready to recover.

Maintain playbooks for token/key compromise (rotate SSH/API/cloud keys, invalidate sessions, rebuild development endpoints), and validate downstream access after any infostealer event.



Detection, intelligence, *and disruption*



Operationalizing detection at scale

A growing challenge for enterprise security operations centers (SOCs) is the operational burden created by high volumes of low-context alerts. Many incidents are built from isolated detections rather than coherent attack narratives, forcing analysts to manually enrich and correlate data to determine intent and impact.

This workflow reduces efficiency and increases the risk that coordinated intrusions remain hidden within benign noise or repetitive false positives. As a result, experienced analysts could be spending more time reconstructing context than disrupting adversary activity.

In response, modern SOCs are shifting from reactive investigation toward proactive disruption. Rather than treating alerts as independent events, advanced detection systems focus on identifying correlated behaviors across hybrid attack chains. Correlation-first detection models prioritize building high-confidence incidents that reflect attacker intent and the full attack scope instead of surfacing disconnected signals.

Automated disruption is a critical component of this model. Once a correlated incident reaches a defined confidence threshold, systems can automatically isolate devices, disable compromised accounts, revoke sessions, or quarantine cloud resources without waiting for manual approval. AI and machine learning enable these decisions by evaluating cross-domain telemetry, identifying likely intrusion patterns, and weighing the operational impact of rapid containment. By compressing detection-to-response timelines from hours to minutes, automated disruption significantly reduces adversary dwell time and limits attack progression.

To support this transition, scalable detection architectures must improve signal quality through suppression, enrichment, and prioritization.

One of the most important and often underdeveloped capabilities is internal infrastructure profiling. Detection systems that lack contextual awareness of the environments they protect struggle to distinguish malicious behavior from legitimate operational activity.

Effective profiling establishes authoritative baselines for:

- Corporate VPN ranges
- Cloud workloads
- Service accounts
- Administrative jobs
- Geographic workforce patterns
- Routine authentication flows

With this context, detection engines can suppress predictable operational noise while elevating behaviors that meaningfully diverge from established norms.

Threat intelligence integration further strengthens detection quality by transforming isolated indicators into actionable context. An outbound connection gains significance when tied to known C2 infrastructure. A PowerShell execution becomes more urgent when associated with documented adversary tradecraft. A suspicious sign-in warrants escalation when linked to credential harvesting campaigns.

To deliver value, threat intelligence must operate as a continuous input throughout the detection lifecycle rather than as a post-incident reference source. Strategic use of threat intelligence should guide detection engineering priorities, while tactical indicators should inform alert generation, enrichment, scoring, and suppression decisions in real time. Organizations that deeply integrate intelligence into detection pipelines can significantly improve analyst efficiency because surfaced incidents already contain the context required for rapid action.

AI is reshaping day-to-day SOC operations beyond detection and disruption

LLMs and specialized agents can summarize incidents, generate investigative narratives, write log hunting queries, and automate executive reporting. AI-assisted triage can classify incidents, recommend investigative steps, and surface relevant evidence across multiple data sources without requiring analysts to manually pivot between tools.

For detection engineering teams, AI accelerates the development, testing, and validation of analytics, reducing the time required to deploy new detections. Operational leaders can use AI-driven reporting and queue analysis to reduce administrative overhead and improve visibility into response performance. This allows skilled defenders to focus on higher-value activities such as threat hunting, hypothesis development, and adversary disruption rather than repetitive operational tasks.

Ultimately, operationalizing detection at scale requires a shift toward quality-centric security architectures where fidelity, context, and correlation are prioritized alongside coverage. The number of rules deployed or alerts generated is no longer a meaningful measure of SOC maturity. What matters is the percentage of incidents that reflect genuine adversary activity, the speed of correlation and containment, and the ability to suppress low-value signals before they reach analysts.

Detection, intelligence, and disruption

Operationalizing detection at scale (continued)

Recommendations

Organizations should modernize SOC operations by integrating infrastructure context, threat intelligence, AI-driven correlation, and automated disruption into a unified detection strategy.

Build authoritative infrastructure context.

Maintain current baselines for identities, devices, cloud workloads, network ranges, administrative activity, and normal authentication patterns.

Prioritize correlation and signal quality.

Measure success by the proportion of incidents that reflect genuine adversary activity, not by the number of rules or alerts generated.

Integrate threat intelligence throughout the detection lifecycle.

Use intelligence to guide detection priorities and enrich, score, and suppress signals in real time.

Use AI to accelerate investigations.

Apply AI-assisted triage, incident summarization, hunting-query generation, and detection engineering so analysts can focus on higher-value work.

Automate high-confidence containment.

Where appropriate, enable rapid actions such as isolating devices, disabling compromised accounts, revoking sessions, and quarantining affected resources.

Track operational outcomes.

Evaluate SOC effectiveness by detection fidelity, time to correlation and containment, and the ability to disrupt malicious activity before it progresses.



How shared threat intelligence is reshaping the fight against fraud

Cyber-enabled fraud operates across a fragmented ecosystem of platforms, services, and infrastructure providers. As a result, no single organization has sufficient visibility to fully understand or disrupt fraudulent activity on its own.

For example, individual platform providers may observe isolated indicators of compromise (IOC), but they often lack visibility into the full scope of harm, which makes coordinated data sharing essential and underscores the need for public and private partners to act together to disrupt fraud. Without this collective view, enforcement actions risk being delayed, incomplete, or misaligned.

At Microsoft, the Digital Crimes Unit (DCU) plays a central role in advancing data sharing as a core pillar of its strategy to combat cyber-enabled fraud. At its foundation, the DCU leverages threat intelligence to identify criminal networks and disrupt their infrastructure through affirmative litigation, criminal enforcement, and global coordination. A critical component of this model is the systematic collection, enrichment, and dissemination of fraud-related data and indicators.

Fragmented visibility across the fraud attack chain reinforces the need to share data

Threat activity is increasingly distributed, with actors leveraging multiple services such as hosting providers, communication platforms, payment channels, and advertising networks to run campaigns at scale and evade detection.

> **No single organization has sufficient visibility to fully understand or disrupt fraudulent activity on its own.**

Detection, intelligence, and disruption

How shared threat intelligence is reshaping the fight against fraud (continued)

As a result, threat hunters across the cybersecurity ecosystem generally detect only distinct elements of a campaign across the attack chain, such as suspicious account activity or transaction anomalies. In isolation, these signals can appear inconclusive or low-priority. However, when combined and analyzed in context, they can reveal coordinated activity or repeated targeting patterns.

Data sharing is essential to overcoming this blind spot, but not all data is equally actionable. Shared data must be vetted, contextualized, and responsibly handled. Effective data-sharing programs require sustained investment in analysis, enrichment, and governance. The ability to transform raw data into actionable intelligence is what ultimately enables coordinated disruption.

Microsoft is increasingly investing in mechanisms to share data and intelligence across sectors, ranging from exchanges during active investigations to programs designed for continuous threat intelligence sharing and collaboration. Platforms such as the Global Signal Exchange (GSE) can facilitate the sharing of fraud and scam signals across organizations such as members of the Global Anti-Scam Alliance (GASA). Examples include private-to-private collaboration, private-to-public partnerships—particularly where law enforcement is involved—and consumer reporting, where data from sources such as the Better Business Bureau (BBB) or the American Association of Retired Persons (AARP) provides critical visibility into victim experiences and emerging trends.

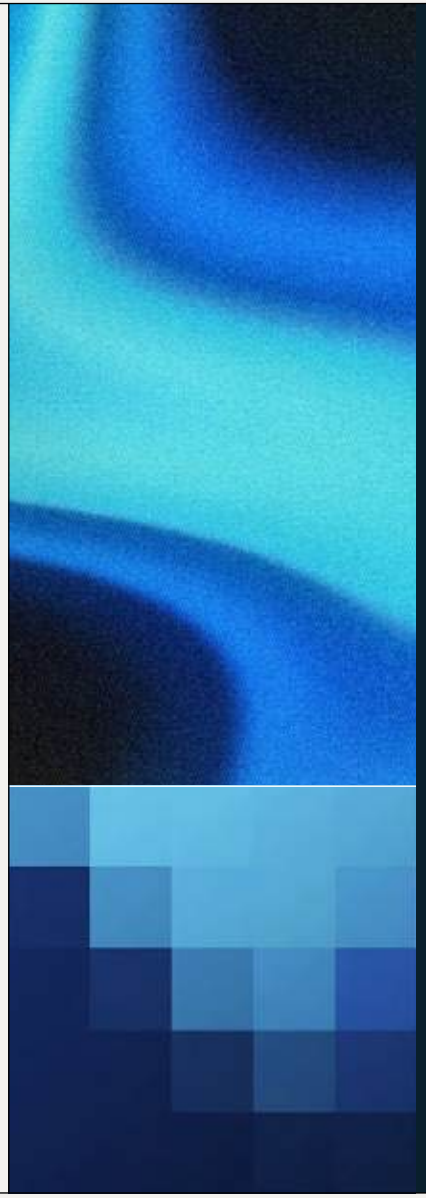
Each intel-sharing type contributes a distinct layer of insight. Together, they enable defenders to move from fragmented observations to a cohesive understanding of the threat landscape.

Coordinated enforcement supports detection and prevention

A defining feature of the DCU's approach is its proactive sharing of threat intelligence derived from investigations and disruption operations. The unit gathers insights into malicious infrastructure, fraud patterns, and threat actor behavior. Internally, these insights strengthen Microsoft's detection and prevention capabilities by feeding into security systems and product defenses. Externally, the DCU provides actionable intelligence to global law enforcement and key partners to support investigations and prevent the reuse of criminal infrastructure. In FY26, Microsoft worked with 16 partners to curate fraud data to support investigation, detection, and prevention efforts, including sharing victim reports, threat actor email addresses, and fraud infrastructure.

How the DCU is combatting tech support scams through data sharing

One way the DCU shares data with partners is through the Cyber Threat Intelligence Program (CTIP). Recently, it introduced a new feed, the Tech Scam Feed (CTIP-TSF), which incorporates victim reports gathered from our tech support scam reporting tool. This adds to existing feeds that focus on C2 and infected-device feeds, which are drawn from the DCU's civil disruption actions. The DCU shares information with CTIP subscribers to facilitate proactive fraud detection and inform disruption, mitigation, and enforcement strategies.



Recommendations

As fraud actors become more coordinated and adaptive, defenders must match that coordination with equally integrated approaches relying on a two-sided commitment: organizations must both share high-quality, actionable intelligence and operationalize incoming signals, integrating them into their own detection and enforcement workflows.

Expanding participation in responsible cross-sector data sharing is critical to achieving this goal. This includes:

- Strengthening industry-to-industry collaboration
- Expanding partnerships with law enforcement and international bodies
- Investing in systems that improve data quality, validation, and context
- Ensuring responsible governance to protect sensitive information while enabling action

Targeting fraud's backbone: Scaling disruption through law, technology, and global partnerships

The cyber-enabled fraud landscape is an adaptive industry shaped and accelerated by specialization, service models, and the commoditization of cyber capabilities.

The various stages of an attack, from phishing to ATO, have converged into a fraud supply chain. Threat actors can now manage and market access to malicious products that use infrastructure and tools that are bundled into cohesive service offerings.

Responses that focus only on serial incident response—or on the takedown of a single scam operation—can struggle to create lasting impact because the fraud ecosystem is resilient and can reconstitute itself. A hyperscale, cyber-enabled fraud leverages a shared services model that combines hosting, elastic compute, anonymized network layers, phishing-as-a-service (PhaaS), bot networks, and credential underground markets. These fraud enablers function as the backbone of the ecosystem. So long as those services remain intact, downstream fraud actors can cycle quickly through new accounts, domains, infrastructure, and ultimately new tactics. On the other hand, when enablers and cybercrime services are degraded, many seemingly unrelated fraud operations can falter simultaneously as costs rise, efficacy decreases, and operational friction increases.

Disrupting fraud: How coordinated action brought down RedVDS

RedVDS was a global cybercrime subscription service that rented out cheap, disposable virtual dedicated servers (VDS). In 2026, Microsoft took it offline as part of coordinated action across civil and criminal law, automation, and partner ecosystems. This shows how joint action can produce more durable outcomes than any single lever alone.

RedVDS exemplified the modern enabler model: it operated as a cheap, subscription-based virtual desktop hosting service. Thousands of cybercriminal customers used it worldwide to launch fraud campaigns at scale, including phishing, ATO, and payment diversion such as BCI. By providing anonymous, disposable virtual machines (VM) optimized for abuse, RedVDS functioned as shared criminal infrastructure that powered hundreds of cyber-enabled fraud operations simultaneously.

Between 2023 and 2026, RedVDS provided over 10,000 users access to up to 8,000 virtual desktops a month. These users were particularly heavily concentrated in Africa and the United States.

How Microsoft dismantled the threat

The DCU filed a civil action in the United States and seized the key domains on which RedVDS relied. Simultaneously, the DCU leveraged intellectual property rights laws in the United Kingdom in a novel and scalable way. This was designed to disrupt operational infrastructure and compel the disclosure of information about the service's operators and the broader criminal ecosystem. This approach shows that existing legal frameworks can be adapted creatively to generate intelligence, impose accountability, and complement criminal investigations. When civil action is structured strategically, it can do more than remove assets—it can expose operational dependencies, create pressure points, and enable a broader understanding of how an enabler sustains itself and how its customers exploit the service.

At the same time, legal action alone is rarely sufficient against modern fraud infrastructure because enablers can migrate, rebrand, or spin up replacement capacity quickly. That's why the RedVDS action was paired with technology-enabled disruption, including the use of the DCU's Statutory Automated Disruption (SAD) program. Through SAD, Microsoft systematically identified and disabled previously created malicious VMs associated with RedVDS to prevent reconstitution at scale.

This reduced the operational resilience of downstream fraud campaigns that depended on this infrastructure. The combination of legal authority and automated technical enforcement helped translate disruption from a one-time event into sustained pressure, limiting the ability of the ecosystem to rebound quickly.

> **When civil action is structured strategically, it can do more than remove assets—it can expose operational dependencies, create pressure points, and enable a broader understanding of how an enabler sustains itself and how its customers exploit the service.**

Detection, intelligence, and disruption

Targeting fraud's backbone: Scaling disruption through law, technology, and global partnerships (continued)

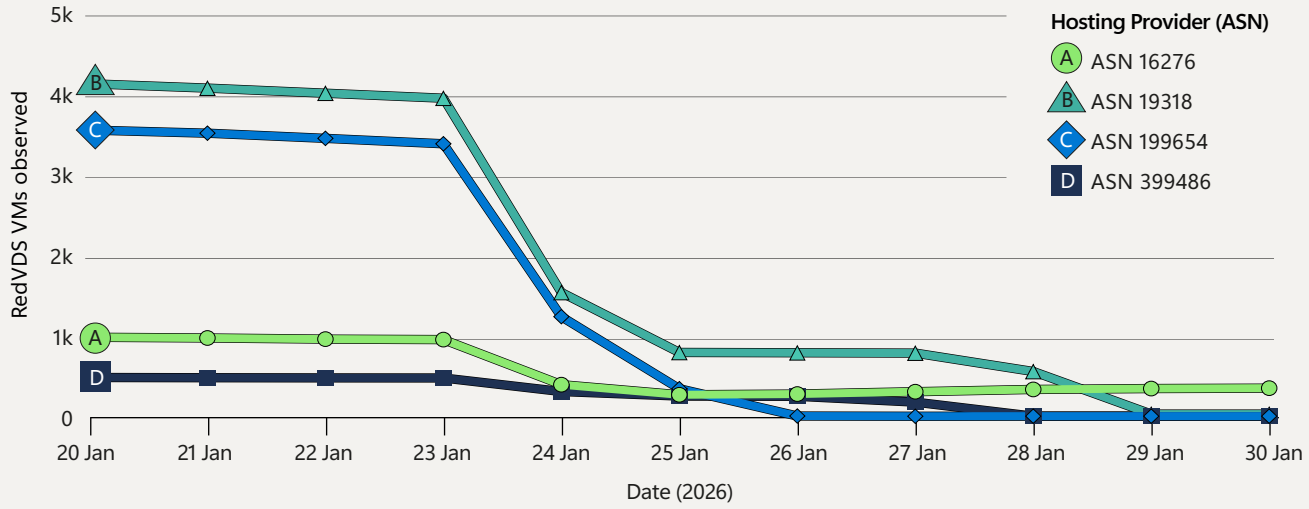
Partnerships are central to making disruption durable

Microsoft's actions were closely coordinated with law enforcement partners to ensure that civil, technical, and criminal authorities reinforced one another. Coordination supports deconfliction and sequencing so that civil action and technical disruption do not unintentionally undermine active criminal investigations. It also enables complementary outcomes:

- Civil processes can drive immediate disruption and produce information that supports accountability.
- Law enforcement can pursue criminal investigations, seizures, and deterrence.

In a cross-border environment where infrastructure, operators, victims, and financial flows frequently span multiple jurisdictions, aligned public-private action can reduce the safe havens on which enablers rely and increase the probability that disruption translates into lasting consequences.

RedVDS malicious VMs decline after takedown action



Beginning in late January 2026, Microsoft sent notifications to hosting providers demanding they take down VMs associated with RedVDS. Most hosting providers, identified by Autonomous System Numbers (ASN) in the chart, began complying with the notices, resulting in a dramatic decrease in available malicious VMs.

Private sector collaboration plays an equally important role

The fraud ecosystem depends on a wide range of commercial services and platforms. Hosting providers, registrars, security vendors, communications platforms, identity providers, and payment ecosystems all intersect with the infrastructure that fraud actors use. In the RedVDS operation, the DCU shared RedVDS-related IOCs with the National Cyber-Forensics and Training Alliance (NCFTA). This helped other providers identify and disrupt infrastructure associated with RedVDS and limit opportunities for rapid reconstitution.

In practice, this means building repeatable operational mechanisms for identifying abuse, validating signals, coordinating timing, and taking action in ways that reduce collateral impact while maximizing disruption to criminal operations. When partnership models mature, they also generate compounding benefits: each disruption yields insight into attacker tradecraft, reconstitution patterns, and infrastructure dependencies, which can be fed back into prevention, detection, and the next cycle of enforcement. Over time, this turns individual interventions into a durable system of collective defense—one that raises the cost of abuse, limits criminals' ability to rebuild, and strengthens resilience across the broader digital ecosystem.

Securing the *foundations*



The race to retire NTLM

NT LAN Manager (NTLM) has been part of Windows authentication for more than three decades. This legacy authentication protocol uses challenge-response verification for access to network resources. However, as security threats have evolved, so have our standards to meet modern security expectations. Microsoft is phasing out NTLM in favor of stronger, Kerberos-based alternatives, creating a more secure authentication model.

Despite years of investment in modern authentication, NTLM remains a persistent identity risk in enterprise environments. Attackers abuse NTLM to gain footholds in enterprise networks, steal credentials, move laterally between systems, and escalate privileges. Microsoft security data shows nearly a sixfold increase in NTLM-related security cases over the past decade. In the first six months of 2026 alone, reported cases exceeded the total number observed during 2025. Publicly disclosed vulnerabilities and attack techniques continue to emerge, underscoring the ongoing security risks posed by NTLM.

Authentication advances are supporting organizations to make a shift

Organizations continue to use NTLM because they are unaware of the risk it poses, are not clear about where it remains in use, and/or still depend on it for compatibility and operational continuity. Advances in the Windows authentication platform are helping to reduce these dependencies, making the transition to modern authentication more practical. As organizations make this shift, they can reduce opportunities for identity-based attacks and strengthen resilience against the evolving threat landscape. Moving away from NTLM represents a major evolution in Windows authentication, and a critical step toward a passwordless, phishing-resistant future.

The cloud threat landscape

The cloud threat landscape is becoming faster, broader, and harder to contain. Over the past year, attackers increasingly used trusted software supply chains to reach many organizations at once, exploited exposed cloud systems within hours, and benefited from faster discovery of serious Linux vulnerabilities.

Major supply chain attacks are growing

Open-source supply chain attacks are becoming more frequent and more damaging. Within the past year, major developer ecosystems saw attacks roughly every ten days, with a clear acceleration through Q2 2026. Affected ecosystems include npm, one of the world's largest online libraries of reusable software components, the Python Package Index (PyPI), Crates.io, and GitHub Actions.

Open-source supply chain compromises are also cloud security incidents because their primary targets, blast radius, and impacts extend into cloud environments. The largest incidents show how quickly one compromise can spread:

- Shai-Hulud 2.0 affected 25,000 repositories.
- Megalodon backdoored 5,561 GitHub repositories in under six hours.
- Axios reached more than 100 million weekly downloads while under attacker control.

Securing the foundations

The cloud threat landscape (continued)

Month	Incidents	Largest wave	Dominant vector
Nov 2025	1	25,000	npm worm
Mar 2026	5	5	GitHub Actions
Apr 2026	6	8	API hijack
May 2026	7	5,661	npm account hijack

Attacks are following three new trends

1. **Attackers are moving upstream into the systems organizations use to build and release software.** A single stolen credential or compromised workflow can give attackers a path to many downstream organizations at once. Incidents affecting developer platforms, build services, and widely used open-source projects show how compromises that start inside development environments can quickly become a broader ecosystem risk.
2. **Attackers are manipulating trust signals.** In some incidents, malicious software appeared legitimate because verification records were stolen, misused, or faked. These cases show that while verification is still important, it cannot be treated as a complete integrity guarantee on its own.
3. **Developers' everyday tools are becoming new entry points.** Code editors and AI coding assistants now sit close to source code, credentials, and build workflows. Recent cases involving malicious extensions and poisoned AI-assistant configurations show that the developer workstation is becoming part of the supply chain attack surface.

Time-to-exploit is narrowing

A container packages an application with everything it needs to run, making containerized applications lightweight, portable, and ideal for modern development. Technologies like microservices, DevOps, and Kubernetes have made containers central to building and operating cloud-native applications. However, containerizing an application also introduces new risks.

Exposed container workloads are being found and attacked quickly. The median time from container start to first exploit attempt is just 5.3 hours, and more than half of compromised containers are exploited within 24 hours of starting. Defenders need to respond in hours, not days. Default credentials, exposed APIs, and unpatched images drive most of the successful compromises.

Linux kernel exploitation is increasing

Linux kernel exploitation is also accelerating. Attackers can use programming flaws in an operating system's core component to gain root access, bypass security controls, or take full control of the machine. Because the kernel manages hardware, memory, and system processes with the highest privilege, compromising it gives an attacker total dominion over the system.

In a two-week period between April 29 and May 13, 2026, three publicly disclosed flaws gave local users a path to root access across major Linux distributions. Each was published with working proof-of-concept code and no embargo, leaving defenders very little time to assess exposure and patch. Attackers found the Copy Fail vulnerability with help from AI, and the other two fit the same broader pattern, suggesting that AI will increase the pace at which serious vulnerability classes are discovered.

Generative AI services are becoming cloud attack surfaces

In Microsoft data, nearly all attacked AI subscriptions in 2026 involved generative AI services: OpenAI and Azure OpenAI, custom AI agents and applications, other generative services, and AI Foundry. Non-generative cognitive services, including Speech, Vision, Document Intelligence, Translator, and AI Search, accounted for less than 2% combined.

Attacked customer subscriptions, by AI surface

AI surface	Share
OpenAI / Azure OpenAI	39%
Custom AI — agents, copilots, LLM apps	31%
Other generative	20%
AI Foundry	8%
Non-generative (combined)	< 2%

Rethinking defense models to tackle intrusions in hybrid cloud environments

Modern threat actors no longer need to break security controls if they can leverage trust and governance instead.

By operating inside legitimate identity and management frameworks, threat actors can turn defensive systems into enablers, extending dwell time and reducing operational risk. This model of intrusion challenges longstanding assumptions about where detection and prevention should be concentrated and highlights why identity governance, alert integrity, and hybrid visibility are now central to enterprise security resilience. Defenders need to be prepared for these intrusions.

Uncovering Storm-2885's hybrid intrusion

In February 2026, DART uncovered how the threat actor Storm-2885 conducted a prolonged and highly effective hybrid intrusion. The threat actor abused cloud identity and management systems as the primary control plane for operations. From there, they operated almost entirely through legitimate Microsoft services, using them to establish persistence, evade detection, and pivot from cloud environments into on-premises infrastructure.

Storm-2885 compromised a highly privileged cloud identity, then used that foothold to repeatedly authenticate, reset credentials, and register additional authentication methods, ensuring long-term access. The same privileged access also enabled the threat actor to interact with security tooling—closing alerts, downgrading incidents, and resolving detections as false positives before security teams could respond. This allowed malicious activity to continue without triggering investigative action.

From its cloud foothold, Storm-2885 leveraged device and identity trust relationships to expand its reach. It used Microsoft Intune device management functionality to deploy scripts and tools across managed systems, establishing secure tunnels that connected threat actor-controlled infrastructure with on-premises environments. These tunnels enabled quiet, resilient access paths that bypassed traditional perimeter defenses and reduced reliance on exposed network services.

Intrusions that work within expected cloud behavior are hard to detect

Storm-2885 leveraged Microsoft Graph API access through a legitimate Microsoft Entra ID application to conduct large-scale data access and reconnaissance. Hundreds of thousands of emails were accessed across hundreds of mailboxes, alongside targeted searches for credentials, infrastructure documentation, and operational details. This activity blended into expected cloud application behavior, making it difficult to distinguish from legitimate administrative or business use without careful correlation.

Within the on-premises environment, Storm-2885 used standard administrative techniques. It used SQL Server, Server remote execution, RDP, and SSH-based tunneling to move laterally, deploy additional access methods, and maintain redundancy. Multiple persistence mechanisms were established across Windows and Linux systems, including scheduled tasks, services, and cron jobs (automated tasks and scripts). This meant access continued even if individual artifacts were discovered and removed.



Recommendations

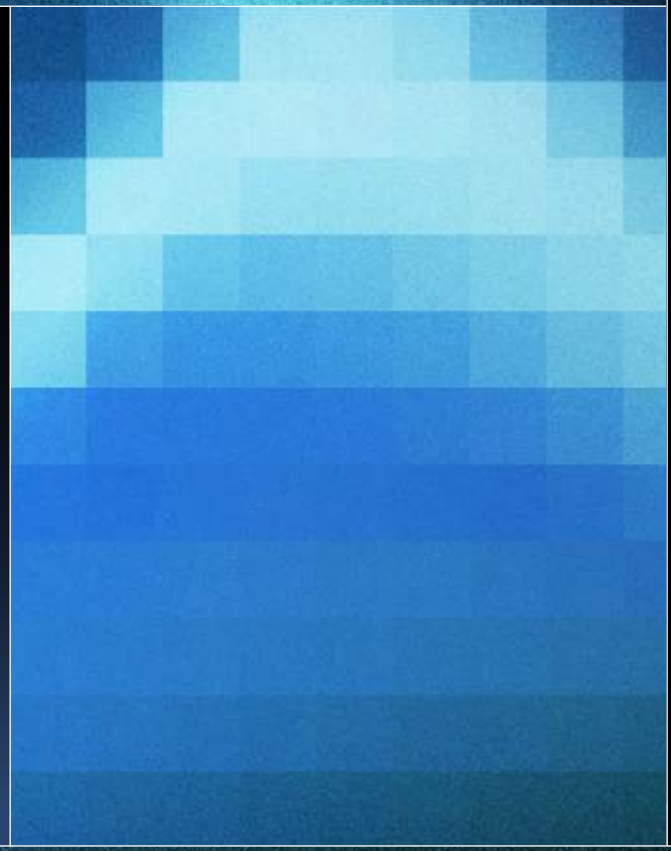
Cyber risk is increasingly concentrated in identity and cloud governance, not just networks or malware. This type of intrusion creates material risk even in well-defended environments. Sensitive information can be accessed quietly, response timelines can be extended, and data exposure can occur without visible operational disruption. Governance and response models need to be designed for this kind of threat, to make sure that security investment is not neutralized by identity compromise.

For CISOs, the takeaway is that identity governance, alert handling, and administrative workflows must be treated with the same prioritization as Tier 0 controls. Endpoint detection, threat intelligence, and response tooling are only effective if threat actors are stopped from operating inside trusted identity boundaries.

Cybercrime

> “Disrupting cybercrime requires more than understanding the threat, it requires action. Joint efforts between Europol and Microsoft have shown how law enforcement and industry can work together to target the infrastructure, tools, and services that enable cybercriminal activity. Recent actions against prolific facilitators such as Tycoon2FA and the malware ecosystem through Operation Endgame demonstrate what can be achieved when technical insights and investigative action are aligned against a common threat.

Edvardas Sileris, Head of the European Cybercrime Centre (EC3)
Europol EC3



Identity is at the center of modern cyber risk 63
Defending against identity compromise 63
Passkeys are the safer replacement for passwords 64
Phishing landscape 65

How attackers gain access 66
QR code phishing 66
CAPTCHA phishing 66
Disrupting phishing: Lessons learned from the Tycoon2FA disruption 67

Legitimate access becomes the threat 69
RMM tools as a post-compromise standard 69
The hidden operational risk 69
Hybrid attack — Identity and cloud compromise 71
Cloud identity abuse increases 73

Ransomware: From access to impact 75
Ransomware landscape 75

Cybercrime disruption and collective defense 78
Information sharing for the modern age 78
Policy, partnerships, and enforcement 78
Cyber mercenaries and the commercial cyber intrusion market 79

Identity is at the center *of modern cyber risk*

Defending against identity compromise

DART continues to observe identity compromise as the leading threat for organizations, despite the attack surface increasing at scale.

As organizations rely more on cloud platforms, AI systems, and automated processes, a single cyber incident can rapidly impact more systems, data, and business operations than in the past. Identity as a common attack vector pre-dates organizational dependencies on cloud, AI, and automation, but threat actors are now operating at a bigger scale and with quicker speed.

As threat actors target non-human identities, they are increasingly leveraging API and trusted workflow integrations to escalate privileges, access sensitive data, and operate at scale. This reflects threat actors' adoption of emerging technology and adaptation to organizational infrastructure models while still using the same time-tested attack methods against the identity plane.

As organizations expand machine-to-machine access and adopt AI-driven services, the attack surface continues to grow. This means the most effective ways to reduce risk and limit business impact stay the same, by:

- Strengthening identity-centric controls
- Reducing reliance on reusable credentials
- Improving visibility across both human and non-human identities

Scale, speed, and automation are changing the game

Unchanged	Changed
Identity compromise	Scale of access
Credential reuse	Speed of execution
Privilege escalation	Increase of use of non-human identities
Access misuse	Automation and orchestration

DART observed a consistent pattern

1. Initial compromise of a human identity
2. Discovery of non-human identity credentials (For example, service principals, API keys)
3. Escalation of privileges using existing access
4. Actions on objectives such as data exfiltration and/or command-and-control deployment

In the past year, DART observed an uptick in threat actor use of built-in workflow automation capabilities and trusted integrations to orchestrate data exfiltration at scale. Actors are able to operationalize access through trusted automation paths, collecting and exfiltrating data without requiring persistent interactive access. Because this activity leverages trusted identities and native services, it can blend into normal operations and evade traditional detection approaches.

Recommendations

Prioritize identity-centric controls.
Implement phishing-resistant authentication and reduce reliance on reusable credentials across human identities. Reduce reliance on long-lived credentials by adopting short-lived tokens and automating credential rotation.

Strengthen visibility into identity usage.
Monitor for anomalous access patterns, new permissions, and previously unseen application or API activity.

Reduce the attack surface of non-human identities.
Enforce least privilege, rotate credentials regularly, and restrict access to trusted locations.

Validate exposure through proactive testing.
Conduct assume-breach exercises (such as tests where you act as if a threat actor is already inside your network). Focus on cloud services, SaaS platforms, APIs, and workflow automation capabilities, prioritized based on where the highest-value data resides, is accessed, or can be aggregated. Ensure these exercises are aligned to known threat actor behavior, including searching for plaintext and otherwise exposed credentials.

Passkeys are the safer replacement for passwords

Passkeys represent a strong countermeasure to both password and phishing attacks.

Passwords continue to be the first vulnerability many attackers test. In the months immediately following the launch of ChatGPT, Microsoft tracked a significant but short-term spike in password attacks, suggesting attackers were experimenting with using generative AI to launch them. The effectiveness of password attacks is waning, however. Overall, password attacks are down 26% year-over-year. This provides strong evidence that security measures such as MFA and passkeys are forcing attackers to pivot to other approaches to achieve access to high-value targets.

Phishing has evolved. In the last few years, attackers have shifted to techniques like fatigue attacks and AiTM phishing to circumvent MFA. In the first six months of 2026, the share of AiTM phishing and token theft attacks more than doubled. This is based on their share as a percentage of the overall attacks Microsoft detected and, of course, is still comparatively small. Nonetheless, it shows the direction of travel. Passkeys are phishing-resistant credentials that aren't replayable. They only work on the website or app for which they were created, meaning if a user lands on a lookalike phishing site, their passkey won't work there.

Organizations need to choose the right passkey for their users

The FIDO Alliance, the National Institute of Standards and Technology (NIST), and the UK's National Cyber Security Centre (NCSC), along with many companies that steward accounts for consumers and businesses, strongly advocate passkey adoption. As organizations standardize using passkeys, choosing the right type—such as device-bound or synced—must be based on user roles and risk.

Passkeys enable phishing-resistant authentication and offer a viable alternative to passwords. However, as organizations secure authentication, threat actors try to take over accounts by attacking other vulnerable areas—for example, new-hire onboarding and account recovery. These rely on weaker credentials like SMS or email-based one-time passcode (OTP). The next step is therefore to remove all phishable credentials from an account to make not only authentication phishing-resistant but also make the account itself phishing-resistant.

Risk class	Synced passkeys	Device-bound passkeys
What it means	Passkey syncs through the passkey provider's cloud to the user's other devices	Private key remains on one device or authenticator and cannot be synced elsewhere
Pros	• Easy to use across phone, laptop, etc. • Built-in backup if a device gets lost • Less chance of getting locked out Synced passkeys can satisfy the requirements for a multifactor cryptographic software authenticator and can be used at AAL2 when implemented in accordance with NIST SP 800-63B-4 requirements.	• More control over where credentials live • Harder for attackers to copy or move Device-bound passkeys implemented as multifactor cryptographic devices can satisfy AAL2 or AAL3 requirements.
Cons	• Dependent on cloud/ account security • Harder for IT to tightly control	• Only works on that one device • No automatic backup or recovery • Can be frustrating for users with multiple devices
Best for	Most users with non-admin privileges in the organization	Admins, sensitive systems, high-risk access

Phishing landscape

Today's phishing landscape is no longer centered on infecting systems. Instead, it's focused on guiding users to compromise their credentials, undetected. Across QR code phishing, CAPTCHA-gated phishing, and malicious payload delivery, a consistent pattern emerges: attackers are prioritizing techniques that evade automated inspection, exploit trusted business file formats, and move credential theft beyond the reach of traditional security controls.

Phishing has evolved from simple malicious attachments, with attackers increasingly designing attacks to slip past automated security tools rather than deliver malware directly. Underpinning much of this shift is the rise of PDF attachments, now the dominant vehicle across multiple phishing techniques. PDFs are trusted in business workflows, rarely blocked by policy, and can direct users to browser-based phishing experiences without triggering many of the behavioral detections commonly associated with HTML or Scalable Vector Graphic (SVG) files.

The increased availability of generative AI is also changing the phishing landscape. Threat actors are using it to automate the creation of context-aware lures—such as emails or messages designed to look real by using specific details about a person or company—that can bypass behavioral triggers. Attackers can then deploy adaptive backends that morph their signatures to evade automated sandboxing. This evolution affects not only spear phishing, but it has turned AiTM into a scalable, automated identity theft pipeline where the goal is no longer just the password, but the persistent, authenticated session cookie itself.

Email bombing is a social engineering pretext

Phishing that targets Teams is a growing problem. In the majority of these phishing cases, attackers initiate the attack with email bombing, followed by a phishing message in Teams from a newly created account that poses as support or helpdesk to remediate the issue. In the last year, Defender Experts observed multiple threat actors using this technique. They then attempted to deploy backdoor and remote access tools using an RMM tool.

Device code phishing adoption is on the rise

Once niche, device code phishing is now a preferred credential theft method for both nation-state and financially motivated actors. Defender Experts have observed this tactic level up in scale and sophistication. Around February 2026, attackers abused railway.com, a legitimate platform-as-a-service provider, to host phishing pages and token-harvesting services linked to the Microsoft 365 OAuth device code flow. Targets received lures such as DocuSign requests, voicemail alerts, and construction bid documents that were often highly personalized and potentially AI-generated. These redirected users to Microsoft's legitimate device sign-in page. Upon authentication, attacker-controlled services silently collected OAuth access and refresh tokens, which were replayed to access Microsoft 365 resources, bypassing passwords and MFA without traditional credential theft.

> **Phishing has evolved from simple malicious attachments, with attackers increasingly designing attacks to slip past automated security tools rather than deliver malware directly.**

This approach is particularly insidious because it exploits legitimate authentication flows end-to-end. No malware is delivered, no credentials are harvested, and no obviously malicious infrastructure is presented. Instead, attackers obtain valid tokens directly through user actions, resulting in persistent, high-fidelity access that appears entirely legitimate.

In the next section, we will look at how phishing operates through:

- QR codes
- CAPTCHA
- Malicious payloads

As cybercriminals gain access through varied methods, defenders need to stay one step ahead.

QR code phishing

QR code phishing (or "quishing") continues to be an effective way for attackers to bypass traditional email defenses.

By embedding a malicious URL within an image, attackers can evade text-based security scanning and direct victims to phishing sites on devices that often fall outside enterprise security controls. Between July 2025 and June 2026, Microsoft Defender for Office 365 detected more than 145 million QR code phishing attacks.

QR code phishing attacks come in waves. The volume of these attacks declined throughout the second half of 2025 and reached its lowest point in January 2026. It then rebounded sharply, reaching an all-time high in March 2026. The trend reversed in May, with volumes declining by more than half by June. Despite this drop, activity remained substantially higher than levels observed during most of 2025, underscoring how widely adopted QR code phishing has become.

The way these attacks were delivered also shifted during the year. PDF attachments were the dominant vehicle overall, with their share growing to 79% of attacks by April 2026 after four consecutive months of gains. Meanwhile, DOC and DOCX attachments, which had gained traction in late 2025, steadily declined.

For defenders, QR code phishing is particularly challenging to confront because it exploits the gap between where organizations enforce security controls and the places where credential theft actually occurs. A user may receive a phishing email on a managed corporate device, open an attached PDF, scan the QR code with a personal phone, and enter credentials on a phishing site without ever triggering traditional network-based detection. As attackers increasingly leverage QR codes to move users outside monitored environments, organizations should ensure that email security solutions can detect and analyze QR codes across attachment types, particularly PDFs. Mobile security policies also need to account for the growing use of personal devices in phishing attacks.

CAPTCHA phishing

CAPTCHA-gated phishing has emerged as a popular method for concealing malicious content from automated security tools.

By placing a CAPTCHA page between the victim and a phishing site, attackers create what appears to be a routine verification step while simultaneously preventing automated scanners from accessing the malicious content behind it. Between July 2025 and June 2026, Microsoft detected more than 100 million phishing attacks that used CAPTCHA pages as an intermediary step before presenting credential harvesting or malware-related content.

During this period, attackers continuously rotated delivery methods. In August 2025, embedded URLs accounted for 58% of these attacks. By November, SVG files had become the leading delivery mechanism (47%). By April 2026, PDFs accounted for nearly two-thirds of all CAPTCHA-gated phishing attacks, the highest share achieved by any single delivery format during the reporting period.

The PhaaS platform Tycoon2FA, discussed later in the report, played a significant role in driving CAPTCHA-gated phishing activity throughout late 2025. It grew from 9% of attacks in July to 76% by December. However, that influence declined rapidly during 2026, falling to 12% by June after law enforcement activity against the group. CAPTCHA-gated phishing activity weakened as the platform's influence diminished, suggesting that other operators have not yet filled the gap as effectively as they did the QR code phishing gap left by the service.

For defenders, the rapid cycling of delivery formats highlights the limitations of file-type-based detection strategies. Techniques tuned to a specific attachment type can become obsolete within weeks. More resilient defenses focus on identifying the behavior itself: CAPTCHA interstitial pages, multi-stage redirect chains, and the separation between initial user interaction and the final credential harvesting page.

Malicious payloads

The role of malicious email attachments has fundamentally changed. A decade ago, malicious attachments were primarily associated with malware delivery. Today, they are overwhelmingly used to facilitate credential theft.

Between July 2025 and June 2026, malicious payloads detected by Microsoft Defender for Office 365 were almost exclusively associated with credential phishing rather than malware delivery.

Instead of directly infecting systems, each month, between 89% and 96% of attachments directed users to phishing pages or presented convincing sign-in experiences designed to steal credentials. The attachment is no longer the weapon. Instead, it's a vehicle that delivers a user to the point of compromise.

HTML and PDF files are common vehicles for malicious payload

Between July 2025 and June 2026, Microsoft Defender for Office 365 processed hundreds of millions of malicious attachments. HTML files were the most common malicious payload in 10 months, though volumes fluctuated significantly because activity was often driven by a relatively small number of large phishing campaigns. HTML remains attractive to attackers because it can render convincing phishing pages, redirect users through multiple stages, load external content, or present CAPTCHA challenges before revealing the final phishing destination.

At the same time, PDF usage surged. After six consecutive months of decline, malicious PDF volume increased by 38% in February 2026 and another 50% in March. Unlike HTML activity, which often spikes around specific campaigns, elevated PDF activity remained consistent throughout the period. SVG payloads declined from 23% of malicious attachments in July 2025 to 7% by June 2026, while DOC/DOCX files peaked at 17% in December 2025 before steadily declining throughout early 2026, mirroring the displacement of Word documents by PDFs seen across other phishing techniques.

The growing dominance of PDFs across multiple phishing techniques underscores their central role in the modern phishing ecosystem and highlights the need for defenders to focus on user interaction and credential theft behaviors, not just malware detection.

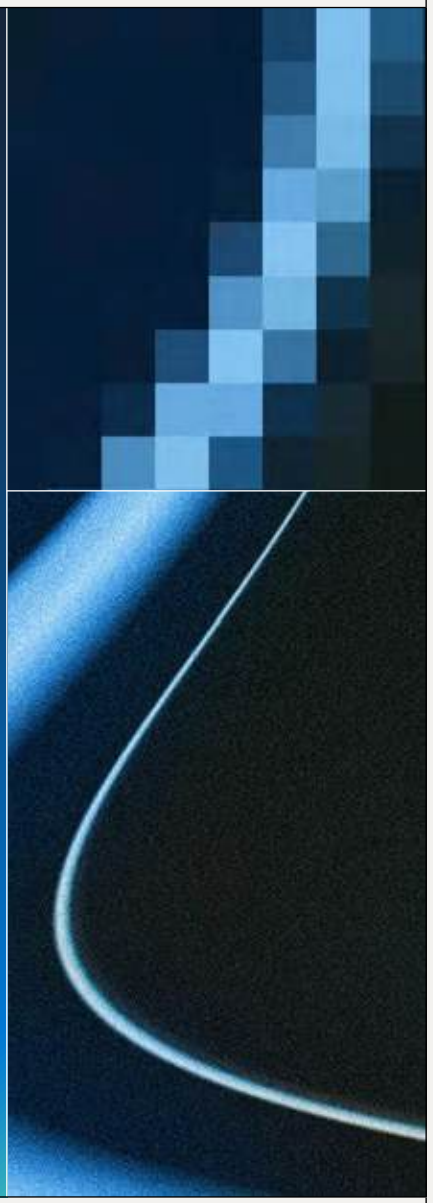
Initial access brokers and bulletproof hosting

Initial access brokers (IABs) continue to play a key role in the cybercrime ecosystem. From July 2025 to June 2026, Intel 471 observed 2,591 advertised access sales affecting organizations in at least 117 countries. This was down from 4,356 offers last year as a handful of prominent brokers reduced their bulk offerings. The United States accounted for the largest share of observed victims, followed by Canada, Italy, Brazil, and the United Kingdom.

Professional services, IT, manufacturing, telecommunications, and government organizations were among the most frequently targeted sectors. 45% of offers were obtained via the use of legitimate but compromised login credentials, while 37% of offers provided unauthorized access via compromised corporate remote access portals such as VPN-based devices.

Bulletproof hosting providers are a foundation of the cybercrime ecosystem

Bulletproof hosting providers enable ransomware, phishing, malware distribution, and other malicious activity by providing infrastructure designed to resist disruption. Since mid-2025, coordinated law enforcement actions targeting several prominent providers have disrupted the landscape and increased pressure on criminal infrastructure networks.



Disrupting phishing: Lessons learned from the Tycoon2FA disruption

Tycoon2FA was a significant PhaaS platform throughout 2025, used for fraudulent emails reaching over 500,000 organizations each month worldwide. In 2026, Microsoft, Europol, and industry partners conducted a coordinated action to disrupt the service.

Tycoon2FA provides AiTM infrastructure, sold on Telegram. During the second half of 2025, Tycoon2FA was linked to more than 15.1 million malicious emails per month and accounted for up to 9% of all malicious messages Microsoft detected. By December 2025, Tycoon2FA was responsible for 76% of all CAPTCHA-gated phishing activity and roughly a third of all QR code phishing attacks.

Disrupting Tycoon2FA infrastructure

In March 2026, the DCU, coordinating with international law enforcement and industry partners, targeted Tycoon2FA's hosting infrastructure. The team went after the servers where phishing pages were rendered and credentials intercepted, the core of what Tycoon2FA sells to its customers. As a result of this disruption operation, March phishing email volume fell 15%, with most activity concentrated in a brief early-month window before the operation took full effect. April dropped another 22%, and by June, Tycoon2FA activity had fallen 95% from its November 2025 peak.

The service's overall effectiveness collapsed. Since Tycoon2FA only provides a hosting infrastructure and email templates and not an actual email sending capability, it's likely many customers were not aware of the disruption and continued sending phishing emails without knowing that the pages behind those URLs were largely gone. As a share of all malicious email, Tycoon2FA went from roughly 11% at its mid-2025 peak to barely 1.5% by April 2026.

The platform's impact in the broader phishing ecosystem was eroded

Tycoon2FA's share of CAPTCHA-gated phishing dropped from 76% in December to 12% by June, while its share of QR code phishing fell from 33% to 14%. Tycoon2FA domains were hosted almost exclusively on Cloudflare, which provided a degree of protection against automated analysis and takedown using CAPTCHAs designed to distinguish human visitors from bots. As a result of the disruption, however, Tycoon2FA moved away from Cloudflare to a variety of alternative platforms, attempting (unsuccessfully) to identify replacement services that offered comparable anti-analysis protections.

The result of this transition has been a significant degradation in Tycoon2FA's ability to evade detection and successfully facilitate credential compromise. The operation also led to the arrest of two key figures from Tycoon2FA, who were detained by the Punjab National Cyber Crime Investigation Agency (NCCIA).

> "By June, Tycoon2FA activity had fallen 95% from its November 2025 peak.

Cybercrime ecosystems thrive when infrastructure, hosting providers, distributors, and operators can act with minimal interference. By coordinating across jurisdictions and sectors, defenders can disrupt those ecosystems at scale. The sustained decline in Tycoon2FA activity observed in the months following the operation demonstrates that these efforts can produce measurable improvements across the wider phishing threat landscape, not just against a single threat actor or platform.



Legitimate access *becomes the threat*



RMM tools as a post-compromise standard

Threat actors have long abused commercial RMM software for persistent, hands-on-keyboard access. The operational breadth and volume of abuse continues to grow.

RMM tools now appear across a wider range of intrusion sets and attack chain stages, compressing the path from initial access to impact and further eroding the value of signature-based defenses. In Microsoft endpoint investigations in the past year, the appearance of an unsanctioned RMM agent has become one of the most reliable mid-stage indicators of a human-operated intrusion.

RMM agents are typically introduced within minutes of a threat actor establishing a foothold and are configured to auto-launch on logon. Since RMM platforms ship with enterprise-grade file transfer, scripting, screen sharing, unattended logon, and session recording, threat actors have the operational reach of a senior administrator from a single agent. Across Microsoft post-incident reviews, the introduction of an unsanctioned RMM agent consistently preceded:

- Credential harvesting
- Lateral movement through the RMM's built-in remote command and file copy features
- Token theft and persistence in cloud identity with mailbox rule or OAuth consent abuse
- Encrypted exfiltration through the agent's transfer channel

Attackers often finish by deploying ransomware from the same operator session. RMM tools make it quicker and easier for attackers to bypass detection. Altogether, this means that a single RMM tool now supplants what once required a dozen discrete tools, collapsing most of the attack chain into one trusted channel that existing detections might not flag. The pattern is particularly pronounced in ClickFix-themed social engineering campaigns, where victims paste clipboard commands that invoke PowerShell to fetch and silently install an RMM client.

The most consequential RMM-enabled intrusions concentrate on high-value assets—domain controllers, identity-synchronization servers, and management infrastructure. Although RMM agents are deployed across all asset classes, the majority of follow-on activity lands on domain controllers, where a single trusted remote session offers the shortest path to tenant-scale impact. The endpoint evidence is often minimal: routine-looking RMM and remote tunnel activity, signed where possible, and indistinguishable from sanctioned administration.

The hidden operational risk of identity and access management in OT environments

OT environments exist wherever physical processes are automated and monitored by digital systems. When it comes to the hardware and software that directly monitor and control physical equipment and industrial processes, identity has become the most consequential and least governed control point.

Enterprise identity controls have been hardened, but OT floors still run on default credentials, shared service accounts, and maintenance sign-ins that haven't been rotated since commissioning. Every organization that runs physical operations on digital controls carries an identity exposure risk.

In May 2025, the US government warned that attackers were accessing industrial control systems (ICS) and supervisory control and data acquisition (SCADA) systems in the US oil and gas sector using basic techniques and succeeding because of poor cyber hygiene and internet-exposed assets. Seven months later, a second federal advisory highlighted pro-Russia groups had accessed OT control devices through minimally secured connections, with impacts including physical damage.

Legitimate access becomes the threat

The hidden operational risk of identity and access management in OT environments (continued)

The recommended fix came down to eliminating default credentials and requiring strong authentication. Threat actors don't have to be sophisticated if they're able to find unlocked doors. The barrier to entry is not technical; it is the absence of identity governance.

While this problem is not new, its consequence profile is. As the boundary between IT and OT continues to dissolve—through digital transformation, remote monitoring, vendor access, and connected supply chains—a compromised identity means exposure to both data and physical systems. When physical systems are manipulated, they don't generate breach notifications; they generate major incidents: production halts, equipment damage, worker safety incidents, and in the most severe cases, public safety events.

Many organizations have responded to this exposure by separating IT and OT onto different networks. Network segmentation governs where traffic can travel, but it becomes irrelevant when a valid credential authenticates on either side of that boundary. An adversary doesn't need to breach a segment if they can sign in.

Many devices were not designed for modern authentication processes

In OT, a significant portion of devices currently running were engineered decades before modern identity concepts existed. Programmable logic controllers (PLCs), human-machine interfaces (HMI), embedded sensors, and SCADA components running on firmware that predates MFA are not candidates for modern authentication enrollment. They don't support certificate-based access controls. In many cases, the concept of an identity-aware access policy is architecturally incompatible with the device itself.

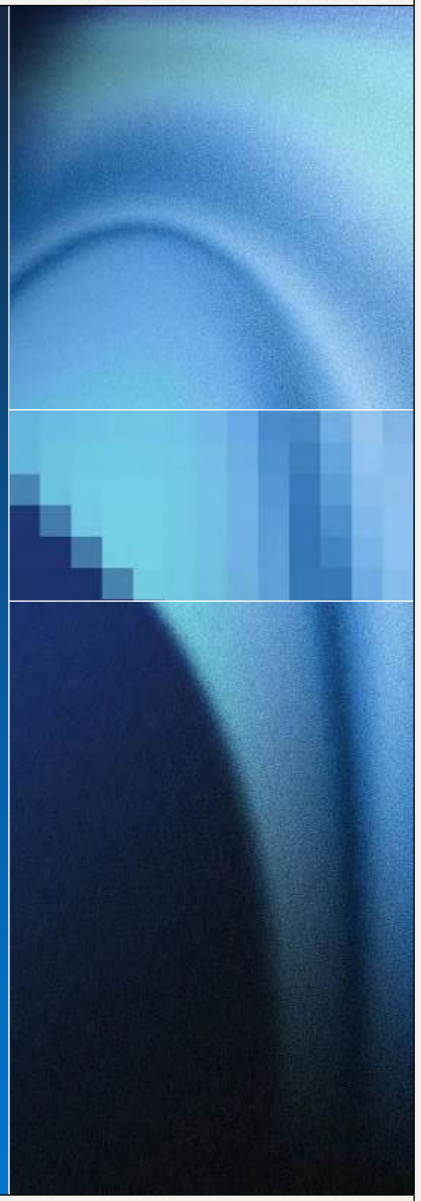
Default credentials on many OT devices are the result of physical infrastructure that was not designed to be secured the way we now require. Changing a default password might disrupt the process in which it's embedded. But this is the gap adversaries are now actively targeting. Industrial environments have become preferred, high-leverage targets precisely because the operational consequence of disruption is immediate and costly. Aging devices, default credentials, limited patching cycles, and identity blind spots are sector-agnostic.

The threat is outpacing the field's ability to respond. In 2023, the CISA ICS advisory program issued 380 advisories. That rose to 423 in 2024 and 508 in 2025.

The foundational architectural decision is to give OT its own purpose-built identity environment. When operational workloads often run on identity infrastructure meant for other purposes, the mismatch shows. Technical debt piles up, credential and access changes become risky, engineering teams fall back on workarounds, and the tools never get used to their full capability. A purpose-built identity environment, on the other hand, closes all of those gaps. Policies are configured for OT from the start, and modern identity capabilities can be applied as intended, not bent around a framework that belongs to a different environment.

The three pillars of OT identity governance: From awareness to accountability

1. **Treat OT identity as an operational risk, not an IT problem.** A compromised OT credential is a production event, a safety liability, and a regulatory exposure. Governance needs to reflect that distinction. Who owns OT identity risk today, and is it sitting in the right conversation with the right consequence framing?
2. **Get a full inventory of every OT identity today.** You cannot govern what you cannot see. Discover default credentials, dormant vendor accounts, shared service sign-ins, machine identities, and embedded device credentials.
3. **Connect OT Identity Access Management (IAM) maturity to business continuity and insurance posture now.** As OT connects to cloud infrastructure, exposure to ransomware reaches operational systems directly. Cyber insurers are beginning to require evidence of OT identity controls. Regulators are moving in the same direction. This is a business posture decision, not a security backlog item.



Hybrid attack: Identity and cloud compromise

Hybrid attacks have become common in modern intrusions. Rather than remaining confined to a single domain, many campaigns span infrastructure, identity systems, and cloud resources in a connected attack path. Hybrid attacks commonly begin with exposed infrastructure, trusted business applications, or workloads, then progress into identity and cloud environments. Persistence is distributed across endpoint, infrastructure, identity, and cloud layers.

Hybrid attack: On-premises to cloud compromise

Adversaries fully compromise on-premises, then steal cookies, tokens, and secrets to pivot into the cloud, where lateral movement runs through identities and secrets (not RDP/WMI), escalating cloud privileges before exfiltrating data to attacker-controlled storage.

On-premises	Pivot • On-Prem -> Cloud			
On-premises compromise	**Pivot to cloud**	**Cloud (cloud lateral movement)**	**Cloud (cloud escalation)**	**Cloud (data exfiltration)**
Initial access to persistence	**Credential access**	**Lateral movement**	**Privilege escalation**	**Exfiltration / impact**
Exploit server or VPN appliance Perform domain lateral movement (Kerberoasting, RDP, WMI) Achieve kernel rootkit persistence	Steal cookies, tokens, secrets Access exposed apps and containers to reach cloud	Move through identities and secrets—not RDP/WMI Replay stolen tokens cookies	Manipulate cloud roles Add credentials to identities and service principals	Platform-as-a-Service (PaaS) or storage data theft Transfer to attacker-controlled cloud account
T1190 - Exploit Public-Facing App **T1558.003 - Kerberoasting** **T1021.001 - RDP** **T1014 - Rootkit**	**T1539 - Steal Session Cookie** **T1528 - Steal Application Access Token** **T1552 - Unsecured Credentials**	**T1078.004 - Cloud Accounts** **T1550 - Alternate Authentication Material**	**T1098.003 - Add Cloud Roles** **T1098.001 - Add Cloud Credentials**	**T1537 - Transfer Data to Cloud**

Legitimate access becomes the threat

Hybrid attack: identity and cloud compromise (continued)

After establishing persistence, threat actors use credential theft, token access, secret retrieval, and identity enumeration to expand control.

The defender operating model

Prevention focuses on reducing exposed infrastructure, enforcing least privilege, and securing identity boundaries.

Detection must correlate signals from infrastructure, identity, and cloud to reconstruct the full attack path.

Containment requires a coordinated response across identity, endpoint, and cloud systems, including revoking tokens and removing malicious applications.

Recovery involves rebuilding trust in identity systems and analyzing cross-domain impact.

The challenge comes from a lack of correlation across layers

Security teams should correlate infrastructure, identity, endpoint, application, and cloud signals into a single investigation path. They can then prioritize controls that reduce exposed entry points, harden identity boundaries, revoke compromised tokens, and remove malicious or over-permissioned applications. The operating priority is to make identity a core control plane for detecting, containing, and disrupting hybrid attacks before access expands across on-premises and cloud environments.

Persistence spread across layers

Attack is hybrid—persistence is distributed across endpoints, infrastructure, identity, and cloud.

1. Endpoint layer	2. Infrastructure layer	3. Identity layer	4. Cloud layer
GOAL Sustained foothold on compromised devices while evading detection and endpoint protections	**GOAL** Resilient control over critical infrastructure components beyond traditional endpoint visibility	**GOAL** Persistent access through identity mechanisms that often bypass traditional security controls	**GOAL** Ongoing access to cloud workloads, data, and services
• Stealthy implants with kernel-level components • Process injection and hollowing • Registry-based persistence • Scheduled task execution • Credential harvesting and reuse	• Compromise of unmanaged devices (ESXi, load balancers, network appliances) • Firmware or configuration manipulation • Creation and abuse of backdoor accounts • SSH/web shell access persistence	• Addition of MFA methods • Creation of app secrets and OAuth consent grants • Service principal provisioning • Token theft and refresh token abuse	• Dormant or hidden cloud applications • Backdoor service principal name (SPN) and long-lived credentials • Exposure of storage access keys • Misconfigured roles and excessive permissions

Hybrid persistence: Multi-layer, multi-platform, hard to fully eradicate

Avoid single-point remediation. Survive partial incident response actions. Enable rapid re-entry after eviction attempts.

Cloud identity abuse increases

Attackers continue to pursue identity as a primary vector into and within the cloud. In cloud identity abuse, threat actors leverage cloud-based identities (users, applications, and workload identities) to gain unauthorized access, escalate privileges, and maintain persistence across SaaS, cloud, and hybrid deployments.

Cloud identity abuse was once a technique limited to highly advanced nation-state actors. Now, it's used by all threat actor types against all industries. Cloud identity abuse typically starts with credential access techniques such as password spray or AiTM phishing. Post-access, threat actors operate within legitimate identity systems and APIs to establish persistence, move laterally, and access sensitive data across SaaS and infrastructure-as-a-service (IaaS) services. Threat actors use automation, including AI-assisted content generation, to scale attacks. They also increasingly abuse legitimate identity features and cloud APIs, helping malicious activity blend into normal operations.

Data from Microsoft Defender XDR incidents between December 2025 and June 2026 reinforces that identity is the way in for attackers.

Initial access comes from password spray attacks

Password spray attacks account for over 99% of observed cloud initial access efforts. Password spray can also be an effective post-compromise technique. Successful password spray attacks can enable ATO, data access, and further lateral movement across cloud services. AiTM credential phishing is the second most common enabler of cloud identity abuse (0.15%) and is a more efficient initial access method because it intercepts authentication flows in real time to capture credentials, MFA responses, and session tokens that provide immediate authenticated access to cloud services.

Threat actors persist post-compromise

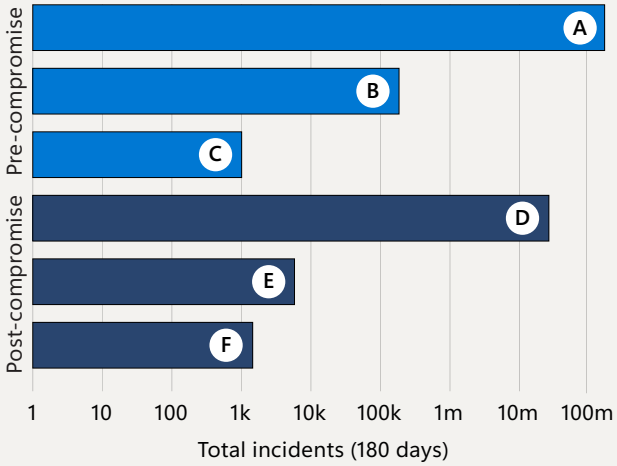
In post-compromise cloud identity abuse, threat actors misuse user and workload identities to operate within trusted authentication boundaries. Threat actors maintain persistence by adding credentials to service principals, abusing cloud applications, and replaying tokens from MFA-authenticated sessions, enabling long-lived access to Microsoft Graph and cloud control planes.

Cloud-related threat tactics by total alerts

December 2025 – June 2026 (180 days)

Key

A	Password spray attacks	124,146,529
B	AiTM credential phishing	151,168
C	Device code phishing	896
D	Threats targeting or leveraging Azure Blob Storage	19,482,467
E	Using token theft and token replay for unauthorized access	5,089
F	Malicious use of Microsoft Graph API	1,262



Top cloud-related threat tactics by total alerts from Microsoft Entra ID Protection, Defender for Cloud, Defender for Office 365, and Defender for Cloud Apps between December 2025 and June 2026.

Source: Microsoft Threat Intelligence

Legitimate access becomes the threat



Cloud identity abuse increases (continued)

Threats targeting or leveraging Azure Blob Storage

Threat actors target Azure Blob Storage (Microsoft's cloud-based object storage) and related cloud resources because the service plays a central role in storing and processing large volumes of unstructured data. This data underpins critical workloads such as analytics, backups, and AI pipelines. Threat actors exploit misconfigurations, exposed credentials, and overly permissive access mechanisms (such as shared keys or shared access signature (SAS) tokens) to gain access. They then use the platform for large-scale data exfiltration, malware hosting, and persistence within cloud estates.

By leveraging Azure-native tooling and workflows to blend into legitimate activity, threat actors can conduct stealthy lateral movement, automation abuse, and downstream compromise of dependent services. This can ultimately affect data integrity and business continuity. Azure Blob Storage may also serve as a C2 channel or staging ground, amplifying its impact beyond a single resource to the broader cloud environment.

Cloud application and SharePoint abuse

Threat actors are increasingly conducting cloud application and SharePoint abuse by leveraging compromised identities and application permissions in Microsoft Entra ID and Microsoft 365 to access data, abuse APIs, and maintain persistence while blending into normal activity. They then move laterally by abusing over-privileged accounts and application permissions, pivoting across cloud resources through legitimate API activity. Meanwhile, workload identities (often lacking MFA and holding broad permissions) allow for low-visibility control of the environment.

Exchange mailbox exfiltration (including Microsoft Graph API)

In Exchange mailbox exfiltration, threat actors leverage access to Microsoft Exchange Online accounts to systematically collect and extract sensitive communications and attachments. They do this using legitimate platform features such as mailbox forwarding rules, API access (including Microsoft Graph), and automated workflows (for example, Power Automate) without triggering traditional security controls.

This activity is particularly difficult to detect because it relies on sanctioned user behaviors and trusted services, enabling sustained, low-visibility data collection over time rather than immediate large-scale exfiltration. Access to high-value business communications can then be used for BCI, espionage, or fraud.

The prevalence of Azure Blob Storage targeting and Exchange mailbox exfiltration highlights how quickly compromised identities translate into high-impact data theft and downstream compromise of dependent services. To disrupt identity-driven attacks before they become enterprise-wide incidents, defenders should:

- Prioritize identity hardening and continuous monitoring
- Strengthen MFA and conditional access
- Reduce standing privileges
- Tighten app and workload identity permissions

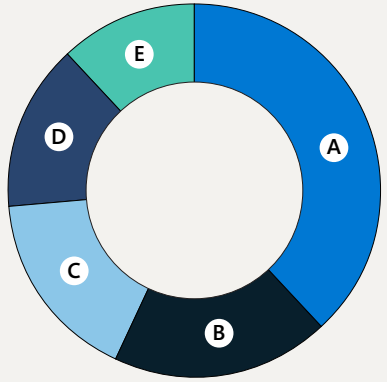
Ransomware: *From access to impact*

Ransomware landscape

Ransomware remains a persistent and evolving threat, shaped by a fluid ransomware-as-a-service (RaaS) ecosystem and expanding extortion tactics. This section examines the most active ransomware families, the countries and industries most affected, and the differences between public leak site claims and attacks detected by Microsoft.

Five most prolific ransomware families (from public leak site data)

Key		%
A	Qilin	16%
B	Akira	8%
C	Play (PlayCrypt)	7%
D	Gentlemen	6%
E	INC (INC Ransom)	5%



Source: Microsoft Threat Intelligence. July 1, 2025 – June 29, 2026

The rise of Qilin ransomware

Qilin was the leading ransomware this year, according to leak site data, up from fourth place last year. This shift was driven by several factors, including:

1. The disruption or weakening of competing RaaS offerings such as RansomHub in 2025.
2. Its competitive profit-sharing model; affiliates keep 80-85% of ransom payments, higher than most other RaaS splits.

Unlike most ransomware variants, Qilin has encryptors for both Windows and Linux/ESXi environments, allowing operators to target enterprise virtualization infrastructure in addition to endpoints. The Qilin service offers affiliates a robust support ecosystem that includes negotiation assistance. The packaging of tooling, infrastructure, and support into a unified platform significantly reduces the technical and operational burden on affiliates and streamlines their ransomware campaigns.

Ransomware: From access to impact

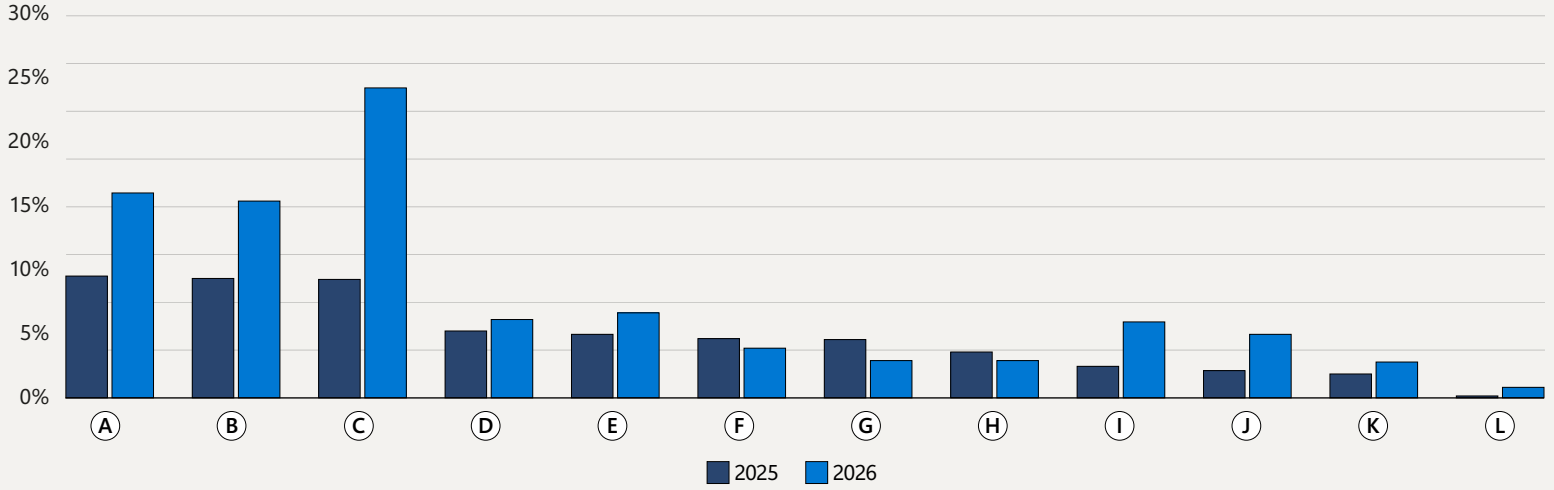
Ransomware landscape (continued)

Defender detections for 2026

According to public leak site data, Qilin listed the most victims in 2026 (16.4%, a 386% increase over last year). Akira had the second-most number of victims (8.3%), while INC, DragonForce, Gentlemen, SafePay, and Sinobi joined the top 10 for the first time. Within the top 10, Cl0p, SafePay, and Play were the only families that did not offer official RaaS affiliation, highlighting how dominant the RaaS model has become.

While leak site data identifies the most active ransomware offerings, RaaS operators may selectively post victims to advertise, coerce, or lie about victims, and so it may present an incomplete or distorted view. Analysis of ransomware events detected by Microsoft Defender impacting enterprises, for example, found Akira was the most active ransomware (22%, up 150% compared to 2025), followed by Qilin (14%, up 111%). Since new families like Gentlemen and Sinobi accounted for a significantly smaller proportion of events in our telemetry, this may suggest that leak site activity was heavily pushed to attract attention to these new offerings.

Impacted industries by percent, year over year



Industry	2025	2026
(A) Critical manufacturing	255	429
(B) Other business entities	250	412
(C) Information technology	248	649
(D) Consumer retail	140	164
(E) Transportation systems	133	178
(F) Energy	124	104
(G) Healthcare and public health	122	78

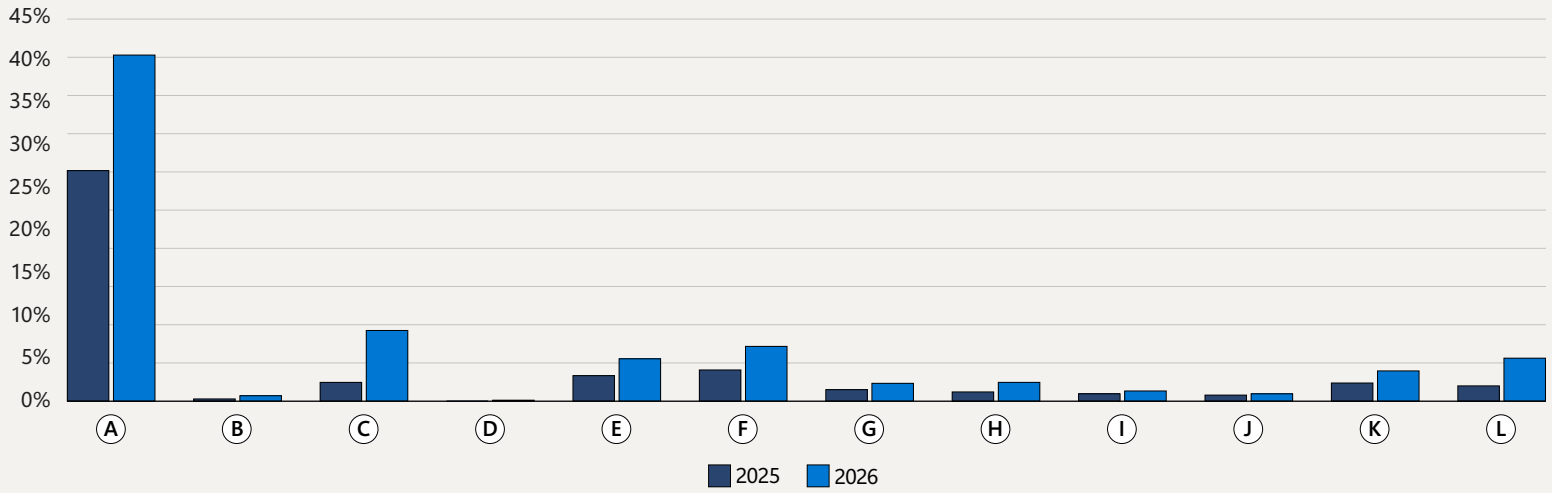
Industry	2025	2026
(H) Research and academia	96	78
(I) Communications infrastructure	66	159
(J) Government agencies and services	57	133
(K) Financial services	50	75
(L) Defense industrial base	4	22
Total	8,749	8,521
Unknown	7,100	5,819

2025 reporting period: July 2024–June 2025; 2026 reporting period: July 2025–June 2026.
Source: Microsoft Threat Intelligence

Ransomware: From access to impact

Ransomware landscape (continued)

Impacted countries by percent, year over year



Country	2025	2026	Percent change from 2025–2026
A United States	724	1,087	50%
B South Africa	7	17	143%
C Germany	59	222	276%
D Türkiye	0	3	N/A
E United Kingdom	80	133	66%
F Canada	98	172	76%
G Netherlands	36	56	56%

Country	2025	2026	Percent change from 2025–2026
H Switzerland	29	59	103%
I Costa Rica	23	32	39%
J Israel	19	23	21%
K Japan	57	95	67%
L Italy	48	135	181%
Total	8,749	8,521	-3%
Unknown	7,089	5,798	

2025 reporting period: July 2024–June 2025; 2026 reporting period: July 2025–June 2026.
Source: Microsoft Threat Intelligence

Microsoft Defender observed a 15.8% increase year-on-year in ransom detonations against enterprises. The United States remains the country most impacted by ransomware. While ransomware is often used against opportunistic targets, continued attacks against Israel (which increased by 21% this year) may reflect Iran's use of ransomware as a form of hybrid warfare.

Ransomware leak site content is no longer confined to the dark web

Threat actors and third parties are increasingly distributing and amplifying stolen data via social media platforms such as X and open-source websites as part of double extortion. Additionally, ecosystems like the "Com" (short for The Community) have used tactics like sextortion and swatting to get victims' attention and for financial gain. In the last year, some actors leveraged social engineering tactics such as aggressive vishing for initial access, consistent with general cybercrime trends.

Cybercrime disruption *and collective defense*



Information sharing for the modern age

Cybersecurity in 2026 is shaped by persistent geopolitical competition and a threat landscape that evolves faster than traditional institutional responses.

State-sponsored actors, cybercriminal groups, and hybrid campaigns increasingly operate across borders and sectors, targeting critical infrastructure, exploiting emerging technologies, and leveraging complexity to evade accountability. In this environment, no single entity—government or private sector—has a complete picture of the threat, nor the capacity to respond alone.

As we've seen already in this report, this reality has elevated information sharing to a central pillar of cybersecurity and digital resilience. Effective defense depends on the timely exchange of technical signals, strategic assessments, and operational insights across a diverse network of stakeholders. Public-private partnerships are foundational to national and collective security. At the same time, expectations of these partnerships are evolving. Governments are seeking more actionable, high-speed intelligence to support decision-making, while the industry is being asked to play a more proactive role in detecting, disrupting, and responding to malicious activity.

Structured, sustained collaboration supports real impact

This shift requires moving beyond episodic information exchange toward structured, sustained collaboration that integrates policy, operations, and technical expertise. Microsoft's initiative with NATO illustrates how such partnerships can be operationalized in practice and scaled to deliver meaningful impact.

In February 2026, NATO's Cyber and Digital Transformation Division (CDT) and Microsoft signed a Letter of Intent to strengthen cooperation in support of a rules-based approach to cyberspace, anchored in international law and agreed norms. Our experience highlights several lessons for operationalizing such partnerships:

1. **Political signaling must be matched by sustained execution.** Framework agreements only deliver value when accompanied by regular engagement, defined workflows, and shared operational objectives.
2. **Practical cooperation must be built through joint activities.** Exercises, incident simulations, and shared playbooks are critical for developing trust, testing coordination mechanisms, and ensuring that partners can operate effectively under real-world conditions. These tools translate high-level commitments into operational readiness.
3. **Information sharing must be bidirectional and timely.** Establishing trusted channels that enable the exchange of intelligence and guidance—while respecting legal and privacy constraints—is key to maintaining shared situational awareness.
4. **Coordinated attribution of malicious activity, supported by both technical evidence and political consensus, is essential for deterrence.** Closer collaboration between governments and industry can help ensure that attribution is credible, timely, and strategically effective.

Policy, partnerships, and enforcement: How to disrupt the ransomware economy

Because ransomware is an ecosystem problem, it is a policy challenge as much as a technical one. The same features that enable scale—shared infrastructure, specialized services, and jurisdictional arbitrage—also create leverage points for disruption when governments and industry act together. Sustained public-private collaboration, targeted operational disruption, and coordinated political pressure—particularly on safe havens—are essential to reducing its impact.

Government choices on legislation, cooperation, and enforcement directly shape the threat environment. Four policy areas have the most impact:

1. **Ransom payments.** Microsoft strongly discourages ransom payments because they reinforce criminal business models without guaranteeing recovery. However, blanket bans risk revictimizing organizations, suppressing reporting, and weakening cooperation. A more effective approach is to discourage payment while supporting victims through no-fault reporting, incident response assistance, and resilience investment.

Cybercrime disruption and collective defense

Policy, partnerships, and enforcement: How to disrupt the ransomware economy (continued)

2. **Safe havens and state accountability.** A small number of jurisdictions continue to function as safe havens, tolerating or even enabling ransomware activity. Microsoft continues to support the designation of persistent safe havens as "state sponsors of ransomware" to elevate ransomware to a matter of international accountability, unlock broader consequences, and reinforce deterrence through both stigma and action.
3. **Cross-border enforcement.** Ransomware investigations are inherently international, yet cooperation remains uneven. Existing legal frameworks already provide the necessary tools: mutual legal assistance, data preservation, asset seizure, and extradition. The constraint is not legal coverage, but inconsistent political will and operational capacity to use these tools effectively.
4. **Consequences and deterrence.** Visible consequences matter. Arrests, indictments, infrastructure takedowns, and asset seizures demonstrate that ransomware is neither anonymous nor risk-free. Recent coordinated actions in cooperative jurisdictions highlight what sustained enforcement can achieve—and contrast sharply with the impunity available in safe havens.



Cyber mercenaries and the commercial cyber intrusion market

Cyber mercenaries have matured into a structured global ecosystem built on a clear division of labor.

In this established market, commercial actors develop and deliver intrusion capabilities, with capital-backed companies carrying out activities ranging from exploit development to hacking-as-a-service and access-as-a-service. Crucially, this growth is accelerating. As the market grows, so do the harms it creates. That includes giving irresponsible actors access to high-end cyber weapons, commercially incentivizing efforts to break critical digital infrastructure, and providing cover for those looking to break international laws and abuse human rights. Given that, Microsoft is redoubling our efforts to call for effective international action to tackle this threat.

The persistence of this market reflects both an increase in demand by states for powerful cyber intrusion tools and weaknesses in mechanisms intended to constrain their abuse. Governments remain among the most influential customers and users of cyber mercenaries, with legal cases in Italy, Spain, Greece, and Poland highlighting not only the extent of public sector purchasing but persistent challenges in accountability and oversight.

Analysis of EU dual-use export controls has found that existing frameworks lack sufficient enforcement and oversight to match the scale and urgency of the challenge. Intermediaries further entrench the problem—brokers and resellers obscure supply chains and facilitate access, yet remain largely outside the scope of current policy responses.

Governments need to meet their commitments to enable change. The cyber mercenary ecosystem is sustained by systemic enablers: they issue licenses without scrutiny, make procurement decisions shielded from oversight, and invest without consequence. If these conditions persist, technical mitigations and isolated enforcement actions will fall short. Governments therefore can and should reform and align their policy controls to close loopholes and target misuse. Because the cyber mercenary market is structurally predictable, it is vulnerable to targeted intervention. Geographic concentration, repeat entrepreneurship, partnerships between hardware and spyware firms, corporate rebranding, jurisdictional arbitrage, and globally mobile capital all provide actionable points of leverage for policymakers.

Governments have acknowledged the problem and committed to action through the Pall Mall Process. This initiative, launched by the UK and France in 2024, is dedicated to tackling the proliferation and irresponsible use of commercial cyber intrusion capabilities. The process created a voluntary and non-binding 'code of practice' for states, but unfortunately, governments seem to have stalled in the implementation phase of their commitments.

Cybercrime disruption and collective defense

Cyber mercenaries and the commercial cyber intrusion market (continued)

In Microsoft's experience, disruption is most effective when it targets the economic and operational foundations of the market and when disruption is carried out in collaboration with other partners. By combining technical disruption, legal action, threat intelligence, product security, and policy advocacy, stakeholders can intervene across the full life cycle of intrusion capabilities.

We call on our partners to work with us to go further:

1. To ensure that commitments made through the Pall Mall Process and other fora are translated into enforceable implementation.
2. To collaborate on targeted action against vendors, brokers, investors, customers, and other enablers.

That means closing procurement, oversight, and accountability gaps; increasing transparency around government use of commercial intrusion capabilities; aligning sanctions, licensing, and enforcement decisions across jurisdictions; and using shared intelligence to disrupt repeat offenders before they can rebrand, relocate, or scale. The goal is to reduce demand, raise the cost of abuse, and make the cyber mercenary market less profitable, less permissive, and less able to operate in the shadows.

1. Setting spyware standards after Europe's spying scandal
2. CDT Europe Report: EU Export Controls Are Failing to Address the Spread of Spyware - Center for Democracy and Technology
3. Mythical Beasts: Investigating the role of intermediaries in the proliferation of offensive cyber capabilities

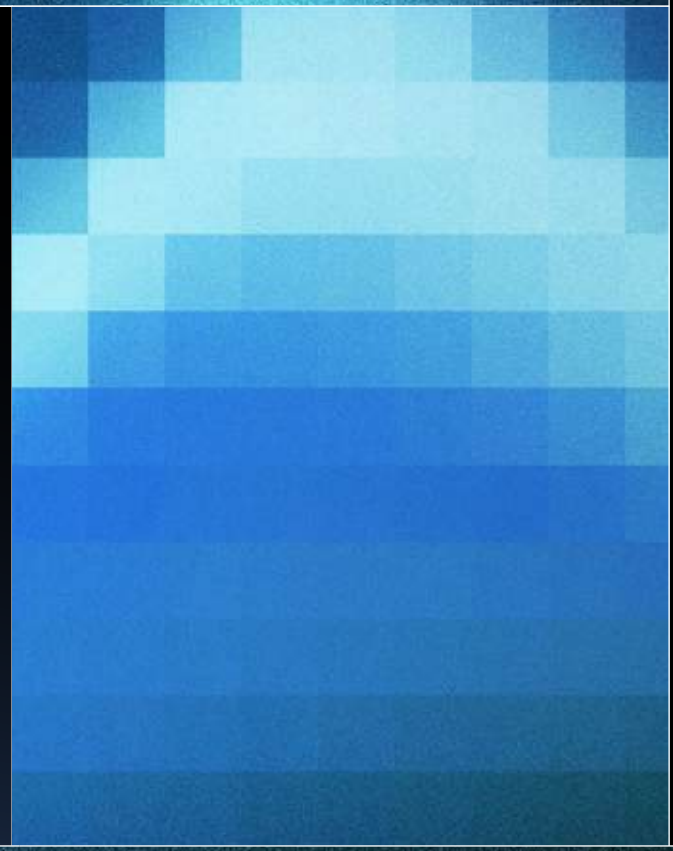


4

Resilience

> “Any comprehensive cybersecurity strategy requires proactive investment in systemic prevention and decisive action against threat actors. Through our collaboration with Microsoft, we’ve seen firsthand how industry leadership on disruption efforts can both reduce threats across the entire ecosystem and produce targeted threat intelligence to help organizations better understand and defend against future criminal and nation-state intrusions.

Nicholas Leiserson, Senior Vice President for Policy
Institute for Security and Technology



Trust, software, and supply chain security 82
Digital media integrity in the age of AI 82
Weaponizing trust 83
Supply chain compromise 84

Critical infrastructure and operational technology 85
Defending a nation starts with defending its critical infrastructure 85
Threat and vulnerability management for critical infrastructure asset visibility 87
Cyber-physical threats to critical infrastructure 89

National resilience and geopolitical security 90
From fragmentation to resilience: What digital defense looks like in an era of polycrisis 90
From reactive defense to proactive cyber defense 90
Evolving geopolitics of cybersecurity 91
Open-source software security 93
Operationalizing cyber resilience: Lessons from the ARC initiative 94
Supporting the future of the CVE Program 95
Subsea cables and global connectivity: Building resilience through a comprehensive approach to security 96

Preparing for the next security transition 97
Quantum-safe readiness for organizations: Start preparing today 97
Securing the quantum future through policy 98

Trust, software, and *supply chain security*

Digital media integrity in the age of AI

For decades, photographs, videos, audio recordings, and official digital communications were implicitly trusted. Today, that assumption is increasingly challenged.

Generative AI enables the creation of highly persuasive synthetic media, while the liar's dividend, the advantage gained by falsely claiming authentic content is fabricated, makes it easier to dismiss genuine evidence.

A critical challenge of the AI era is the growing difficulty of determining which content can be trusted. The implications extend beyond trust in everyday online content, affecting business continuity, brand reputation, crisis response, exposure to fraud, geopolitical stability, and an informed citizenry—a foundation of democracy. Several dynamics are contributing to this shift:

- **Synthetic content creation is growing rapidly**, increasing both the volume of content that must be evaluated and the baseline level of uncertainty.
- **Mixed media will become harder to interpret**, as authentic, altered, and fully synthetic elements are combined in ways that complicate verification.
- **Offline media creation introduces new security risks**, as content can be generated or modified outside observable systems and then introduced into digital environments with incomplete, compromised, or misleading provenance information.
- **Authentication signals are proliferating but remain inconsistent**, creating fragmentation across tools, platforms, and standards and making trust more difficult to establish.

These dynamics warrant sustained executive attention, as digital media integrity is a sociotechnical challenge, with vulnerabilities emerging at the complex intersection of technology and human behavior.

This is the context for Microsoft's report, Media Integrity and Authentication: Status, Directions, and Futures (MIA), which examines the current authentication ecosystem, including the opportunities and limitations of provenance, watermarking, fingerprinting, and detection. Building on Microsoft's contributions to the technical foundations of media provenance and its role as a co-founder of Coalition for Content Provenance and Authenticity (C2PA), the report helps leaders navigate advances in generative AI, evolving legislation, and the growing adoption of authentication technologies.

In practice, authentication methods serve different purposes and have important limitations. Content and associated signals can be altered, degraded, or lost as media moves across platforms. To assess the implications of these real-world disruptions, researchers evaluated hundreds of authentication scenarios through sociotechnical red-teaming exercises. Only a limited subset yielded high-confidence validation, with the strongest results observed when provenance and watermarking were linked.

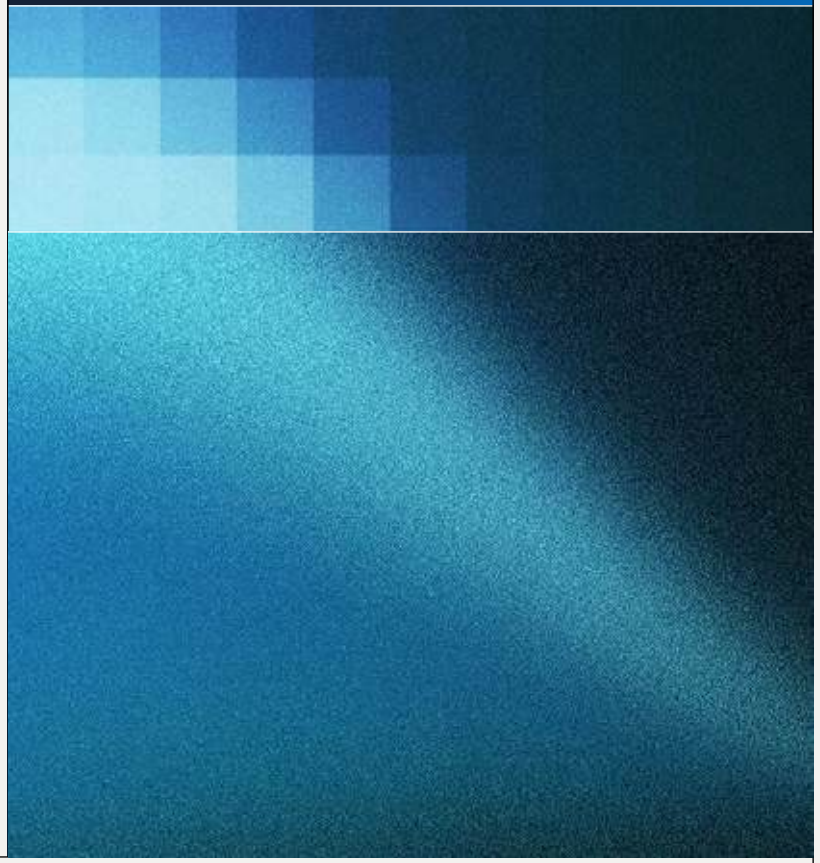
The MIA report concludes that trust in digital media will depend not on any single technology, but on layered defenses, transparent systems, continual red teaming to inform evolving standards, and sustained collaboration. Organizations that recognize media integrity as foundational infrastructure will be better positioned to strengthen resilience, protect their reputations, maintain operational continuity, and build confidence in AI-enabled innovation.

Resources

Media Authenticity Methods in Practice: Capabilities, Limitations, and Directions

A new study explores how AI shapes what you can trust online

Coalition for Content Provenance and Authenticity



Weaponizing trust: Malicious abuse of code signing infrastructure

In May 2026, Microsoft led a global effort to expose and disrupt a growing cybercriminal ecosystem. Fox Tempest, a financially motivated threat actor, operated a malware-signing-as-a-service (MSaaS). This model meant that cybercriminals, including ransomware groups, could access trusted, signed malware for use in attacks against schools, hospitals, and critical infrastructure.

Through technical investigation, legal action, and collaboration with industry and law enforcement partners, Microsoft identified and dismantled key elements of the threat actor's online infrastructure. In May 2026, the DCU filed a civil case enabling the seizure of a core domain and disabling hundreds of actor-controlled virtual machines and accounts. This legal action disrupted the actor's operations, cutting off access for other cybercriminals. Simultaneously, Microsoft revoked over a thousand fraudulent code-signing certificates and dismantled supporting infrastructure, preventing further abuse. These efforts culminated in the arrest of a primary suspect in Europe, demonstrating how coordinated public-private action can translate intelligence into real-world disruption.

The operation reveals an emerging threat to the software supply chain

Rather than attempting to bypass security controls, threat actors increasingly seek to exploit the trust mechanisms designed to protect users. By fraudulently obtaining legitimate code-signing capabilities, cybercriminals can make malware appear authentic, increasing the likelihood that malicious software will evade detection and reach victims.

Fox Tempest leveraged Microsoft's artifact signing processes by creating fraudulent developer identities and submitting malicious binaries for signing. By passing through legitimate verification steps, the threat actor was able to obtain authentic Microsoft-signed certificates for malware payloads, including ransomware loaders and infostealers. Fox Tempest industrialized this abuse process, using automation and social engineering to onboard new developer accounts and submit malicious artifacts.

Code signing depends on vigilance across the entire trust lifecycle

Attackers do not just bypass controls, they actively seek to inherit trust by manipulating legitimate processes. Trust mechanisms must evolve alongside attacker tradecraft, and defenders must be prepared to validate trust continuously rather than assume it indefinitely. This calls for continuous integrity across issuance, use, and monitoring.



Supply chain compromise: From isolated incidents to cascading ecosystem attacks

Software supply chain compromises have evolved. In the past, attackers focused on small, single-target hacks, targeting intrusion to a specific vendor.

Today they use massive, automated attack chains that spread quickly across multiple platforms, using one weak spot to automatically break into many connected systems. Attacks span multiple package registries (online digital storage hubs where developers download ready-made code pieces) and CI/CD platforms (automated systems that test and build new code changes for a project). What begins with a single stolen token or misconfigured workflow can cascade outward: compromising packages, tools, and organizations at scale.

Two common supply chain attack methods are:

1. **Poisoning trusted downloads.** Attackers can secretly replace legitimate software with tampered versions containing hidden malware. For example, in April 2026, the CPUID website, a popular source for computer hardware tools, was compromised. Attackers swapped genuine installers with versions that installed malware, giving them remote control over affected machines. For users, everything appeared genuine—the download came from an official site and the software seemed authentic. They had no obvious reason to mistrust it.
2. **Exploiting code libraries.** Large code libraries, like npm, are central hubs where developers share and download reusable code. Their massive scale makes them prime targets. Attackers can inject malicious code into these libraries which then automatically spreads to thousands of other projects. This is a particularly effective initial access vector because during installation, a package can automatically run its code, often before anyone reviews it. Access to these libraries relies heavily on individual developer accounts, the credentials for which can be stolen through phishing. This reinforces how attackers are increasingly exploiting trusted relationships within software ecosystems rather than targeting individual endpoints directly. As those ecosystems become more interconnected, compromises involving packages, developer accounts, AI-assisted development tools, and build infrastructure can rapidly extend beyond a single organization.

Recommendations

Prioritize controls that reduce the blast radius of package compromise, limit credential exposure in build environments, and shorten time-to-detection for supply chain intrusions.

Avoid exposing sensitive access tokens in development workflows, use short-lived credentials, and design build systems as temporary.

Strengthen publisher accounts. Enforce strong security measures like phishing-resistant MFA for all accounts with publishing access to code libraries.

Rotate credentials after exposure. If any software is compromised, immediately change all accessible credentials and rebuild affected systems.

Detect new attack methods. Monitor for unusual software activity, unexpected connections, and new evasion techniques.

Maintain a software inventory. Keep an up-to-date list of all open-source components used in production to quickly respond to new compromises.



Critical infrastructure *and operational technology*

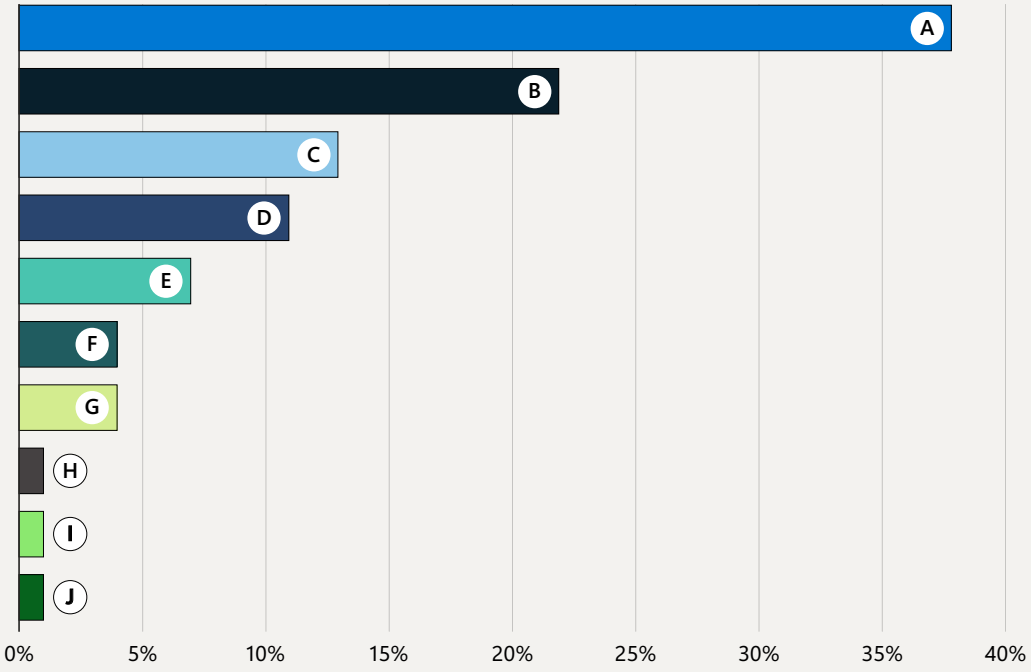
Defending a nation starts with defending its critical infrastructure

Nation-state actors, ransomware syndicates, and opportunistic criminals are pivoting from data theft toward disruption of the physical services that hold communities together.

Securing state and local government, education (SLED), and critical infrastructure is not just an IT problem. It's now a national security imperative that requires breaking the vertical model in which regional IT defends only regional systems, while municipalities, cities, schools, and utilities fend for themselves.

Research and academia is the most-targeted SLED sector (38%)—more than transportation (22%) and government (13%) combined. Educational institutions and systems sit at the intersection of open collaboration, federal research funding, and sprawling supplier ecosystems. This makes them a key entry point into the public-sector supply chain. According to Microsoft's Defender notifications, cloud identity abuse is the most observed technique against both SLED (62%) and critical infrastructure (78%), highlighting the importance of defending identity in these sectors (see Figures 2 and 3). Microsoft identified threat actor activity leveraging OAuth consent abuse, a technique that re-emerged prominently in the research and academia sector over the last year and is now considered a cross-industry method for exploiting trust relationships in SaaS ecosystems. The observed activity included overlapping tradecraft commonly associated with ShinyHunters.

Figure 1 – **Attacks against SLED and the critical infrastructure sector**



Key		%
A	Research and academia	38%
B	Transportation systems	22%
C	Government agencies and services	13%
D	Communications infrastructure	11%
E	Critical manufacturing	7%

Key		%
F	Energy	4%
G	Government facilities	4%
H	Chemicals	1%
I	Water and wastewater systems	1%
J	Emergency services	1%

Data is from November 1, 2025 to April 30, 2026. Source: Microsoft Threat Intelligence.

Critical infrastructure and operational technology

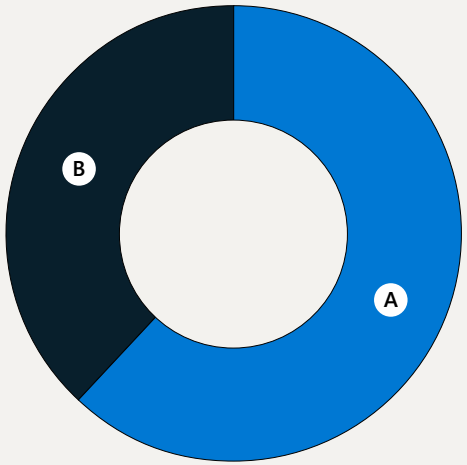
Defending a nation starts with defending its critical infrastructure (continued)

Once inside, adversaries move with commodity and bespoke tooling. Against critical infrastructure, Truebot malware (39%) and information stealers (38%) dominate the observed tool mix, with the AzureHound framework (12%) used to map cloud-attack paths. Information stealers accelerate the supply chain. They harvest credentials, which become the next intrusion's entry point, cascading across the vendors, integrators, and shared services that connect a municipality's smallest water utility to its largest hospital system.

Legacy water systems, residual community risk

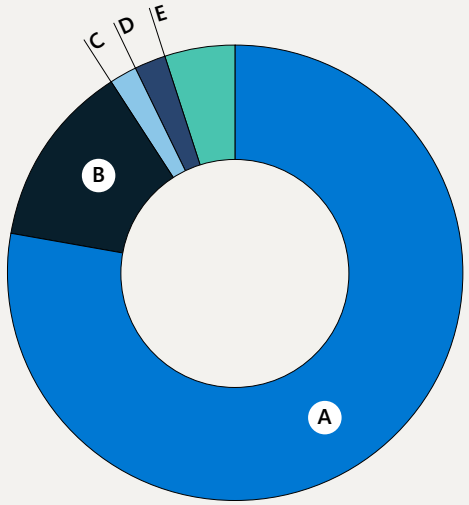
US drinking water and wastewater utilities run aging operational technology, often maintained by private operators and lower-capacity districts. In 2025, the Environmental Protection Agency reported that more than 70% of inspected water systems failed to meet basic cybersecurity requirements. Incidents such as the late 2024 American Water customer-portal outage, show how fast those gaps translate into service disruption and community safety risk.

Figure 2 – **Technique summary: SLED**



Key		%
A	Cloud identity abuse	62%
B	Password spray attacks	38%

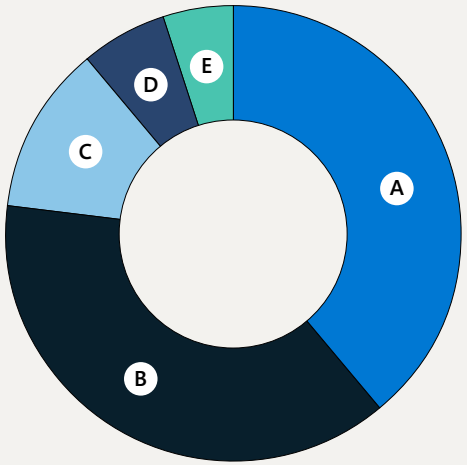
Figure 3 – **Technique summary: Critical infrastructure**



Key		%
A	Cloud identity abuse	78%
B	Password spray attacks	13%
C	Brute-force attacks	2%
D	Evolving phishing threats	2%
E	Other	5%

Figures 2 and 3 — Cloud identity abuse dominates the technique mix in both SLED (62%) and critical infrastructure (78%). Source: Microsoft Threat Intelligence. March 1 to June 30, 2026.

Figure 4 – **Tool summary: Critical infrastructure**



Key		%
A	Truebot malware	39%
B	Information stealers	38%
C	AzureHound framework	12%
D	Impacket	6%
E	LemonDuck	5%

Figure 4 — Truebot malware and information stealers lead the tooling used against critical infrastructure. Source: Microsoft Threat Intelligence. November 1, 2025 to April 30, 2026

Defending a nation starts with defending its critical infrastructure (continued)

Recommendations

Prime targets—like schools, water districts, local agencies, and the vendors that serve them—typically lack the resources to defend themselves. Collective defense is needed to meet the scale of the threat.

Protect lifelines. Treat water, power, communications, and emergency services as a single attack surface. Fund areas that are underserved for cyber investments, such as rural areas and small towns. Create whole-of-state shared service models that extend Zero Trust, identity, observability, and response to the smallest jurisdiction.

Defend identity first. Mandate phishing-resistant MFA, disable legacy and high-risk authentication flows (including unmanaged device code flow), and deploy continuous identity threat detection.

Stand up civil cyber defense and coordinate response across boundaries. Pool capability so no single jurisdiction defends alone across a state, through:

- SOCs
- Local defense
- National cybersecurity agencies
- Sector-based information sharing and analysis centers
- Universities
- Volunteer cyber corps

Civil cyber defense gives target-rich, resource-poor communities shared monitoring, surge response, and mutual aid. It moves away from isolated approaches, where every organization tries to build their own.

Build an AI-defended workforce. Pair AI-driven vulnerability management with student SOCs, community college pipelines, and apprenticeships to close the gap between adversary speed and defender capacity.

Prepare for the quantum horizon. Begin crypto-agility and post-quantum migration planning now so trust in public-sector computing survives the transition.

Threat and vulnerability management for critical infrastructure: From asset visibility to active defense

Critical infrastructure defenders face a fundamental mismatch: adversaries operate continuously, while traditional security programs remain periodic.

This gap is at the highest risk in datacenter environments supporting critical infrastructure. Here, undetected exposure does not just risk data loss but can cascade into physical disruption, service instability, and safety impacts. Nation-state actors are pre-positioning within infrastructure networks to enable future disruption or sabotage of critical operations. Meanwhile, financially motivated actors increasingly exploit unpatched systems, targeting availability rather than data.

As a result, threat and vulnerability management (TVM) must evolve into a continuous, integrated operational discipline, not a set of tools or periodic activities. Frontier AI is accelerating vulnerability discovery and exploitation, widening the gap between discovery and remediation. For critical infrastructure operators, closing this gap is now a strategic imperative. As a result, TVM must evolve from a periodic security function into a continuous operational capability. Executed correctly, TVM becomes a closed-loop defense system binding asset visibility, vulnerability identification, threat detection, and AI-accelerated insight. Each capability reinforces the others to create a compounding defensive advantage.

Reframing TVM as a continuous operational discipline

The key architectural shift is from tool-centric vulnerability management to continuous threat exposure management embedded into operations. Reporting metrics should shift from the number of patches deployed to exposure reduced, detection coverage increased, and time-to-mitigate compressed. A continuous TVM model:

- Integrates asset data, vulnerabilities, threats, and telemetry into a live system
- Operates as a feedback loop, not a sequence of steps
- Prioritizes risk using context, not static severity
- Adapts dynamically even when remediation is constrained

This shift aligns defense with how adversaries actually operate—continuously probing, adapting, and exploiting.

At the core of this model is a reinforcing cycle of four capabilities:

1. **Asset visibility.** Effective defense begins with complete and continuously updated awareness of all assets across IT and OT environments. That includes device relationships, network topology, segmentation boundaries, configuration state, and operational dependencies.
2. **Continuous vulnerability identification.** Vulnerability management here requires a shift from static scanning to continuous exposure awareness:
   - Integrating vendor advisories, ICS-specific disclosures, and CVEs
   - Prioritizing by asset criticality and exposure
   - Accepting delayed or constrained patching realities

Critical infrastructure and operational technology

Threat and vulnerability management for critical infrastructure: From asset visibility to active defense (continued)

3. **Threat detection and active defense.** Detection must operate as a continuous, intelligence-driven function rather than reactive alerting. Effective datacenter detection includes behavioral endpoint visibility across operational systems, identity and network telemetry to detect lateral movement, OT-aware detection logic, and threat intelligence mapped to adversary behavior—not indicators that expire.
4. **AI-accelerated insight.** At datacenter scale, manual analysis is insufficient. Frontier AI is creating both new risks and new defensive capabilities. As vulnerabilities outpace remediation, AI helps prioritize exposures, identify attack paths, correlate signals, and accelerate response, enabling teams to focus on the risks most likely to be exploited. AI prioritizes high-risk alerts (triage compression), accelerates detection engineering cycles, correlates identity, network, and OT signals in near real time, and maintains a continuously updated risk model.

The defining strength of this model is interdependence:

- Asset visibility improves vulnerability accuracy
- Vulnerability context strengthens detection prioritization
- Detection insights refine asset and exposure awareness
- AI continuously correlates and accelerates all layers

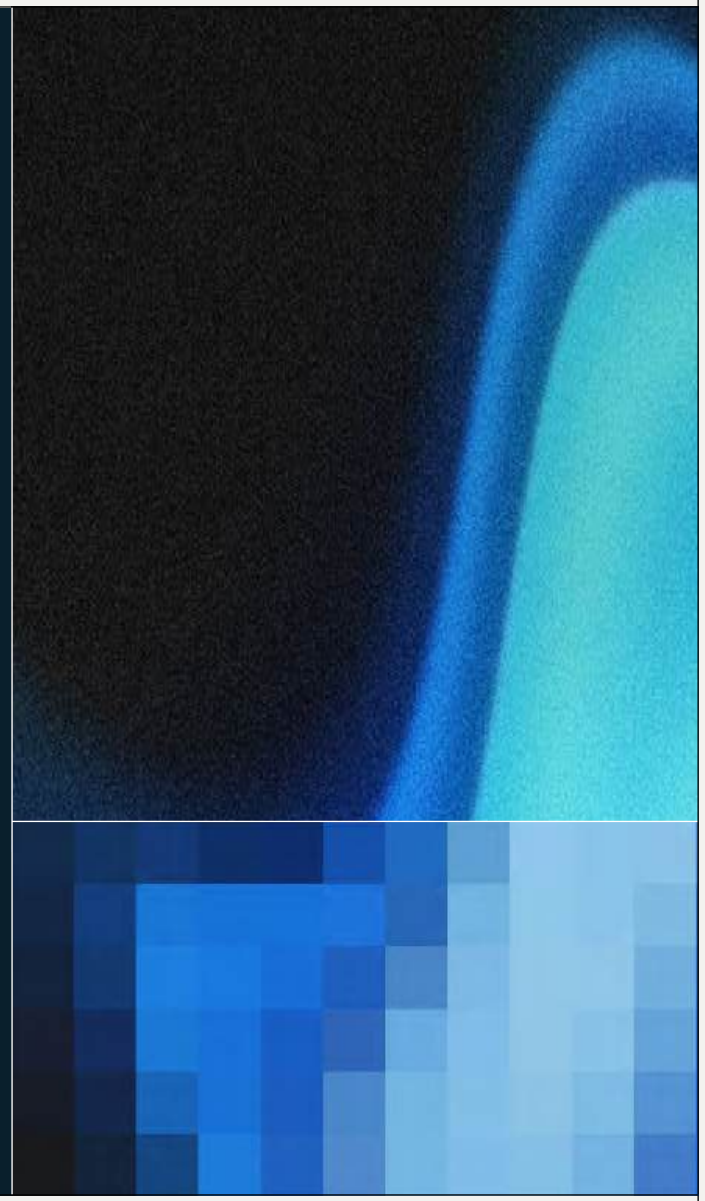
This creates compound resilience against increasingly AI-enabled adversaries. Organizations that integrate AI into TVM can reduce exposure windows, prioritize remediation based on real-world risk, and respond more effectively to threats operating at machine speed. Closing the loop between proactive exposure reduction and reactive threat detection covers both sides of the breach equation—shrinking attack surface while catching what slips through.

> **The key architectural shift is from tool-centric vulnerability management to continuous threat exposure management embedded into operations.**

Recommendations for operators

Critical infrastructure and datacenter operators should make it a priority to:

- Establish verified, continuously updated asset visibility across IT and OT.
- Integrate vulnerability data with asset context and threat intelligence.
- Build detection capabilities aligned to adversary techniques, not signatures.
- Measure effectiveness by coverage and detection gaps, not alert volume.
- Introduce AI-driven correlation early to avoid scale bottlenecks.
- Treat TVM as a core operational discipline, not a siloed security function.



Cyber-physical threats to critical infrastructure

Cyber-physical threats are a strategic risk to critical infrastructure. Datacenters connect cloud services, facility operations, physical security, power, cooling, and life-safety systems. Compromise of IT, OT, or digitally controlled building systems can move quickly from network access to physical consequence: power loss, cooling disruption, lockdown, lost visibility, equipment damage, or life safety risk to personnel.

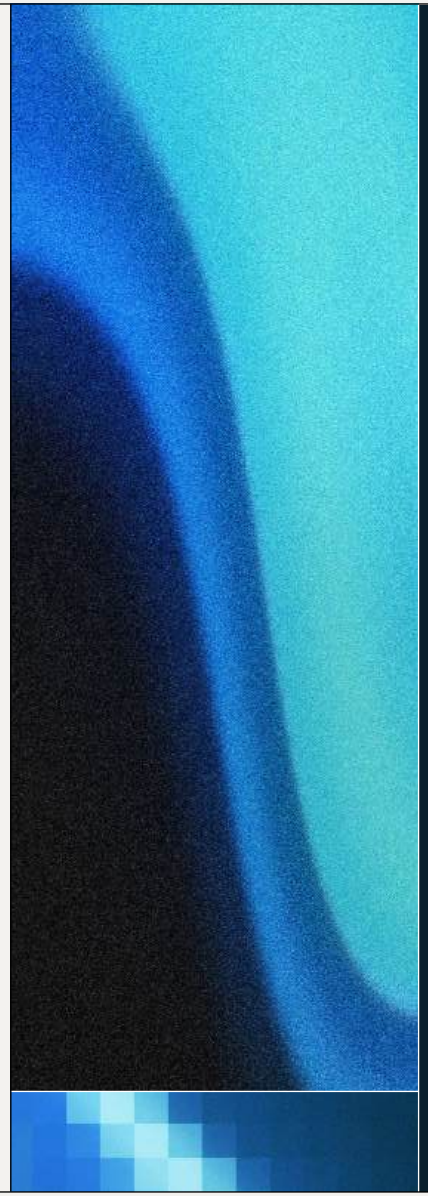
For datacenters, cyber-physical risk concentrates where digital control planes govern physical outcomes. Power, heating, ventilation, and air conditioning (HVAC), building management, access control, surveillance, fire suppression, and emergency systems are interdependent. Compromise of one system can cascade into downtime, unsafe conditions, lost visibility, or impaired emergency response.

Datacenter physical systems should be defended as interconnected critical services

The six physical security domains are now cyber-physical:

- Perimeter security
- Access control
- Surveillance and monitoring
- Asset protection
- Operational controls
- Environmental/facility safeguards

Each depends on controllers, sensors, connected panels, management platforms, or identity systems that adversaries can target using familiar IT techniques.



Recommendations

Organizations should adopt a resilience-focused defense strategy that goes beyond preventing intrusion. Key priorities include:

Detect and disrupt consequence-level manipulation by monitoring for unauthorized state changes and operational impacts, not just initial access.

Maintain independent verification of physical processes through trusted monitoring, manual validation, or redundant sensing when digital control systems are degraded or untrusted.

Strengthen coordination across cybersecurity, OT engineering, physical security, and facility teams to respond effectively to cyber-physical threats.

Reduce both external and insider attack paths through strong identity protections, least-privilege access, and continuous monitoring.

Minimize adversary dwell time with rapid detection, containment, and response capabilities.

Regularly exercise cyber, physical security, and facility recovery plans to ensure safe and timely restoration of operations following an incident.

National resilience *and geopolitical security*

From fragmentation to resilience: What digital defense looks like in an era of polycrisis

For much of the past decade, cybersecurity strategies have been shaped by focusing on distinct risks to cybersecurity—for example, ransomware, distributed denial-of-service (DDoS) attacks, and supply chain attacks.

While these threats remain very real, they increasingly fail to describe the environment in which digital risk now unfolds. We have entered an era of polycrisis. This context is not defined by isolated events but by overlapping disruption, leaving little to no breathing room to recover. Geopolitical instability, economic pressure, rapid technological change, and persistent cyber activity interact in ways that increase impact and reduce response time. Digital systems are no longer stressed by a single event; they are tested by overlap, dependency, and cascading failure.

This shift exposes a widening gap between how many organizations defend digitally and how risks emerge. Fragmented defenses—optimized for individual threats or compliance requirements—become less successful when crises start to unfold simultaneously. Today's environment demands a move from fragmentation toward persistent digital resilience.

Fragmentation is not just a technical issue—it's built into how many organizations operate. Over time, rules, silos, and quick fixes add complexity. Identity, infrastructure, and OT often run separately. In a polycrisis, these gaps become weak spots, slowing decision making when speed matters. Improvements in one area may address specific issues, but they don't always add up to resilience across the whole system.

From defense-in-depth to resilience-by-design

Defending against threats to interconnected systems requires visibility across domains, disciplines, and geographies, as well as an understanding that digital systems are inseparable from societal resilience. This reality challenges traditional approaches to cyber defense such as defense-in-depth, where controls are added in layers, and gives greater focus to resiliency. Resilience-by-design assumes that failures will occur, that dependencies matter as much as adversaries, and that recovery and continuity are as critical as prevention.

Cloud platforms can function as resilience enablers. They offer consistent security baselines, global situational awareness, the ability to adapt at speed, and the collective management of complexity.

In a polycrisis environment, digital defense must evolve from guarding individual assets to developing systems that will function under pressure so that services to the public will continue. This requires organizations to design security for disruption, not just compliance, to align cyber resilience with broader continuity planning, and to treat identity, data, and infrastructure as inseparable. It also demands deeper collaboration—between public and private sectors, across borders, and among organizations that share infrastructure and risk—to ensure that dependencies are known and managed.

> **Digital systems are no longer stressed by a single event; they are tested by overlap, dependency, and cascading failure.**

From reactive defense to proactive cyberdefense

Government SOCs are no longer just fighting human adversaries; they're now fighting automated, AI-accelerated attack infrastructure. Today, defenders need to be right every time across thousands of systems, while an attacker needs only one path—and generative AI is helping find that path.

Nations that have invested in structured public-private collaboration demonstrate faster detection and disruption of multi-sector campaigns. Collaboration comes through joint operations centers, automated indicator-sharing platforms, and sector-specific information sharing and analysis centers (ISAC). The most effective national threat intelligence architectures are federated, rather than centralized. They operate under legal safe harbors that encourage good-faith sharing, support machine-readable formats for automated ingestion, and function in near real time.

Collaboration must actively work in practice, and at pace. A detection authored inside one agency should be deployed to peer agencies in hours, not weeks. And where agencies cannot reveal raw data, privacy-preserving matching lets them compare threat intelligence without exposing sensitive telemetry, resolving the tension between sovereignty and collective defense.

From reactive defense to proactive cyberdefense (continued)

National resilience can come from shared agentic defense

Every department running its own Security Operation Center (SOC) duplicates cost and tooling, while smaller agencies—municipalities, health authorities, school boards, etc.—are left exposed by a chronic skills gap. From many fragile, isolated SOCs, we must move toward one shared, agentic, multi-tenant cyber defense capability that every agency consumes as a service.

In this ideal model, a central shared-services SOC operates the platform, the analysts, and an agent fleet on behalf of dozens of agencies. This can reduce duplicated tools and create economies of scale. The model preserves autonomy and sovereignty by design. Each agency keeps its own tenant, data residency, and Zero Trust boundary, while central analysts operate across tenants through delegated, just-in-time, least-privilege access. Logs remain within agency-controlled environments and nothing is centralized that does not need to be.

The defensive fleet follows a simple pattern: deny, disrupt, delegate.

- **Deny:** Agents deny adversaries by exposing and closing attack paths before they're reached.
- **Disrupt:** Agents disrupt threats through autonomous triage and containment.
- **Delegate:** Agents delegate the work of codifying every lesson into new detections.

In this agentic model, scarce analysts are freed from conducting manual triage and turned into strategic supervisors.

Evidence of the benefits of this type of organization already exists. Organizations leveraging Microsoft Security Copilot report summarizing threats 60 to 70% faster, sharply cutting the mean time to resolve critical alerts. Autonomous tools such as the Phishing Triage Agent filter out background noise, saving security teams nearly 200 hours each month by closing thousands of false-positive alerts. Microsoft's own security team now uses custom AI agents to autonomously investigate 75% of incoming incidents (Microsoft, State of the SOC, 2026), freeing experts to hunt novel threats and architect stronger defenses.

Of course, autonomy without accountability is a non-starter in government. This agentic capability must be secured and governed. Low-risk actions such as alert enrichment and threat intelligence summarization can run autonomously; medium-risk actions like isolating an account require human approval, while high-risk decisions such as declaring a breach remain human-led. Replayable reasoning trails record what each agent decided, on what data, with what confidence, and whether a human reviewed it—aligning with the NIST AI Risk Management Framework and the EU AI Act.

> **Microsoft's own security team now uses custom AI agents to autonomously investigate 75% of incoming incidents, freeing experts to hunt novel threats and architect stronger defenses.**

Recommendations for public sector leaders

Governments should strongly consider investing in shared services and federated SOC models rather than agency-by-agency builds and adopt bounded-autonomy governance now, before agents proliferate. They should establish collective-defense sharing under legal safe harbors and fund it sustainably through transparent metering and chargeback.

Evolving geopolitics of cybersecurity

Nation states use cyber operations to coerce, shape conflict, and signal state capability. This digitizes warfare in a way that draws in civilians and private technology providers and can contribute to restrictions on internet freedom as governments expand surveillance and control over digital spaces. In this environment, cloud platforms, communications networks, and AI-enabled services become key points of contestation, not just commercial assets.

The expansion of geopolitical competition in the cyber sphere underpins the launch of the UN General Assembly's Global Mechanism (GMECH) on information communication technologies (ICT) in international security in 2026. The process highlights the complementary roles of technology and diplomacy:

Technical capabilities enable detection, disruption, and recovery from cyberattacks. Diplomacy shapes the broader environment—legal frameworks, policy choices, and norms—that define responsible state behavior in cyberspace.

As a permanent forum for cyber diplomacy, the GMECH on ICTs' value will depend on its ability to deliver tangible progress, particularly in strengthening protections for critical digital infrastructure. That includes the cloud services that underpin essential public and economic functions, such as healthcare and energy.

National resilience and geopolitical security

Evolving geopolitics of cybersecurity (continued)

The future of cyberspace is shaped not only by operations but by governance. Much of the infrastructure and expertise underpinning resilience sits outside government. Meaningful stakeholder participation in the GMECH is essential for ensuring that the rules that govern cyberspace are practical, implementable, and grounded in operational reality. Excluding industry and civil society, for example, risks fragmented responses, weaker policy outcomes, and increased legal uncertainty.

The move to action and recognition of responsibilities in armed conflict

Advancing cyber stability requires moving toward clarity on how international law applies in cyberspace, who interprets it, and how it is enforced. We are entering a more mature phase of cyber governance centered on attribution, evidence, due process, and meaningful accountability.

A similar transition is underway in the governance of AI in security and defense. The United Nations is shifting toward practical implementation and sustained engagement. The emerging Framework of Responsible Industry Behaviour for AI in the Military Domain builds on this by offering voluntary, actionable guidance. It was developed through a multistakeholder process involving governments, industry, and civil society. In parallel, initiatives such as the International Committee of the Red Cross's (ICRC) Digital Emblem aim to enhance the protection of medical and humanitarian assets in cyberspace.

In an increasingly contested landscape, delivering cybersecurity demands clearer rules, credible accountability, and sustained cooperation. Strengthening protections for critical infrastructure, advancing operational accountability, and preserving an open, secure, and resilient digital environment must be shared priorities. The path forward will depend on translating commitments into practice—ensuring that norms are implemented, responsibilities are upheld, and resilience is built where it matters most.

Framework of Responsible Industry Behaviour for AI in the Military Domain

As AI becomes more embedded in security and defense, industry decisions shape real-world risk. Choices about system design, data practices, testing, deployment, maintenance, and support affect reliability, predictability, accountability, and the likelihood of harm. The UN's Framework of Responsible Industry Behaviour for AI in the Military Domain seeks to create a practical and voluntary baseline for responsible industry conduct. Its value lies in translating existing high-level principles into clearer expectations for companies and governments. The framework is being developed through an inclusive process involving governments, international institutions, civil society, academia, and industry. In a field where technical choices can have strategic consequences, this framework's operational guidance is essential to building trust, accountability, and predictability.

Digital Emblem

The Digital Emblem Project addresses a practical challenge in modern conflict: medical and humanitarian assets that are protected under international humanitarian law are increasingly dependent on digital systems, but those protections are not always visible in cyberspace. The Digital Emblem aims to create an authenticated, interoperable marker that can help identify protected medical and humanitarian digital assets online, making existing protections easier to recognize and respect during cyber operations.



Open-source software security

Open-source software underpins critical infrastructure across governments, economies, and essential services.

However, its security model has not kept pace with frontier AI. Advanced AI systems are increasingly capable of accelerating vulnerability discovery, analysis, and exploit development, potentially increasing the speed and scale at which threat actors can target widely used software. Sustaining trust and resilience will require a more active government role—not to control open source, but to strengthen the institutions, incentives, and guardrails that support it.

Enhance vulnerability discovery and response

Governments should complement community efforts with dedicated defense centers leveraging AI for continuous scanning, triage, and response. National cybersecurity bodies should partner with Akrites, a Linux Foundation coalition that coordinates confidential vulnerability remediation for open-source software upon which much critical infrastructure depends. Vulnerabilities must follow coordinated disclosure processes, supported by funding and technical assistance, to enable rapid remediation.

Improve adoption of updates and global alignment

Governments should provide supply chain transparency and standardized disclosures, and offer reliable update mechanisms while addressing technical debt and unsupported software. International alignment through bodies such as the G7 should focus on common principles, voluntary baselines, and interoperable tools. Meanwhile, incentive-driven public-private partnerships, rather than mandates, will best preserve innovation and global collaboration.

Recommendations

Strengthen coordination and prioritization. Governments should establish a National Open Source Security Office (OSSO) to coordinate policy, funding, and engagement. Today's fragmented approach limits visibility and accountability. An OSSO would enable national risk assessments, prioritize critical dependencies, coordinate internationally, and guide secure adoption.

Invest in security capacity and secure by design practices. Governments should accelerate adoption of AI-enabled secure-by-design tools across the ecosystem, including automated vulnerability detection and stronger supply-chain governance. Investments should support:

1. Initiatives: For example, Open Source Security Foundation (OpenSSF), a global project hosted by the Linux Foundation that works to improve the security of open-source software, and Alpha-Omega, a specialized initiative under it that finds and fixes security flaws in critical code.
2. Standards and guidance: For example, the NIST Cybersecurity Framework, Secure Software Development Framework, and related software supply-chain guidance could help organizations to better understand and improve how they manage cybersecurity risks.
3. Organizations advancing secure development:
   - Software Assurance Forum for Excellence in Code (SAFECode), a global nonprofit organization dedicated to improving software and hardware security through industry collaboration.
   - Open Worldwide Application Security Project (OWASP), a global open-source community focused on improving software security through community-led documentation, tools, and standards.

Expanding existing frameworks will further strengthen resilience. That includes software bills of material (SBOMs—a tool that lists the component parts and software dependencies of a software package) and coordinated vulnerability disclosure (CVD–a collaborative process used to safely report, fix, and reveal security flaws to the public).

A National Technology Fund could also provide sustained support to maintainers of critical projects, reducing systemic risk while preserving independence.

Operationalizing cyber resilience: Lessons from the ARC initiative

Microsoft's Advancing Regional Cybersecurity (ARC) initiative was launched to strengthen preparedness, resilience, and coordination through public-private collaboration tailored to governments' operational needs.

It responds to a growing reality: rapid digital transformation—particularly across the Global South—has expanded both opportunity and exposure to cyber risk. ARC focuses on building the institutional capabilities required both to prevent incidents and to respond and recover effectively when they occur. The initiative was first implemented through a partnership with Kenya's National Computer and Cybercrime Coordination Committee (NC4). It has expanded into Latin America through work with Mexico's Digital Transformation and Telecommunications Agency (ATDT).

Cyber outcomes are shaped by effective governance

Across digitally transforming economies, cyber outcomes are increasingly shaped not by technical controls alone, but by how effectively governments coordinate during an incident. Response effectiveness depends on institutional readiness—clear roles, decision-making authorities, and coordination mechanisms—in addition to detection and prevention capabilities. As public sector systems become more interconnected, incidents can trigger cascading effects across ministries and sectors, exposing gaps in real-world response that are not always visible in policy frameworks.

ARC engagements highlight recurring structural challenges: unclear roles across agencies, fragmented escalation pathways, and limited mechanisms for whole-of-government crisis management. These gaps are particularly visible in scenario-based exercises. In these, coordination often breaks down at the intersection of technical response and policy decision making—for example, around public communication, system prioritization, or cross-sector resource allocation. By addressing these governance challenges alongside technical issues, governments can build effective resilience.

ARC Kenya: Uncovering the value of scenario-based planning

The ARC partnership with NC4 illustrates how scenario-based planning can surface operational dependencies that are difficult to identify through policy alone. Kenya has strong foundations, including a maturing regulatory framework for critical information infrastructure. However, these elements are not yet fully integrated into a unified national operating model.

As a result, friction points emerge where technical response intersects with national decision making. ARC findings point to several practical readiness measures:

- Establish a shared incident response lifecycle and playbooks for high-impact scenarios.
- Create a streamlined "report once" mechanism for incident reporting.
- Link organizational SOCs to sector-level coordination structures.
- Define clear information-sharing protocols that enable speed while protecting sensitive data.

Regular tabletop exercises and after-action reviews are essential to operationalize and continuously improve these arrangements.

ARC Mexico: Strengthening national incident response frameworks

Similar coordination challenges are shaping ARC engagements in Latin America, where governments are seeking to strengthen national incident response frameworks amid rapidly expanding digital services. In Mexico, ARC supported work with the ATDT, focused on improving whole-of-government coordination in preparation for the FIFA World Cup. This brought together technical, civilian, and regulatory stakeholders to test crisis response workflows and identify institutional dependencies that could affect mitigation and recovery during major cyber incidents. This approach is now being scaled regionally.

> "At the Council of Europe, data from the Microsoft Digital Defense Report is used to inform the design and implementation of cybercrime capacity-building programs for law enforcement worldwide.
>
> **Aisling Kelly,**
> **Head of Cybercrime Division**
> Council of Europe

Supporting the future of the CVE Program

The Common Vulnerabilities and Exposures (CVE) Program is a critical piece of shared infrastructure. It provides unique identification numbers for specific software or hardware security flaws.

This powers everything from incident response and patch management to risk scoring, compliance reporting, and security research, and is a critical piece of shared infrastructure. It provides unique identification numbers for specific software or hardware security flaws. The common reference numbers enable stakeholders to consistently identify and manage vulnerabilities once they are discovered. As the volume of vulnerabilities expands, the CVE ecosystem must scale to be sustainable and trusted.

As software systems, supply chains, and discovery techniques continue to evolve, maintaining the effectiveness of that foundation will require careful attention to scalability, data quality, governance, and long-term sustainability. Strengthening the CVE Program means ensuring that its operating model continues to meet the needs of a larger, more complex, and more interconnected environment. This can support faster coordination, better prioritization, and more effective risk reduction worldwide.

Evolving the CVE ecosystem

Since the CVE Program was created, the program's governance and capabilities haven't significantly evolved. As discovery capacity accelerates via frontier AI model capabilities, the program needs to evolve its throughput, data quality, and governance.

Mounting structural, technical, and governance challenges threaten the CVE Program's long-term effectiveness. Governance and funding models have struggled to keep pace with the program's evolution, raising concerns about neutrality, continuity, and international participation when oversight or resourcing is uncertain. At the same time, the volume and velocity of vulnerability discovery have increased dramatically, exposing scalability limits in CVE Numbering Authority (CNA) workflows and manual processes, and contributing to delays and inconsistencies in published records.

Recommendations for policymakers

To ensure the CVE Program remains a trusted and effective foundation for global vulnerability management, policymakers and ecosystem leaders should pursue a targeted set of reforms focused on sustainability, governance, technical modernization, and international alignment.

Recognize the CVE Program as essential shared infrastructure and resource it sustainably. This provides a starting point for ensuring its long-term stability by supporting more predictable, appropriately scaled funding models and continuity planning.

Support innovation through continued investment in APIs, tooling, and automation. As vulnerability discovery continues to grow, the CVE Program will benefit from investment in modern technical capabilities. Scalable APIs, structured and machine-readable data, and greater use of automation can improve efficiency, consistency, and timeliness. Stakeholders should direct funding and technical support toward modernizing CVE infrastructure, especially APIs, CNA tooling, and automation capabilities.

Leverage public-private partnerships and international participation to advance shared goals. Strengthening the CVE Program presents an opportunity for closer collaboration between governments, industry, researchers, and the open-source community that can help accelerate improvements in infrastructure, tooling, and ecosystem support, translate shared policy objectives into scalable outcomes, and contribute to CVE modernization efforts.

Subsea cables and global connectivity: Building resilience through a comprehensive approach to security

Subsea fiber optic cables are the critical backbone of the global communications and internet infrastructure, transporting over 99% of intercontinental data traffic.

They underpin intergovernmental communications, financial transactions, and defense communications, making them essential to national and international security. However, this critical infrastructure is physically vulnerable. Many cables traverse vast oceanic spans that include volcanic zones, tectonic fault lines, and polar seas. Natural disasters like earthquakes or undersea landslides, fishing activity and ship anchors, and deliberate sabotage or tampering can all damage these cables.

Most subsea cables are owned and operated by private consortia. As a result, primary responsibility for securing these systems rests with private cable operators, who must invest in robust physical, logical, and supply-chain security measures. At the same time, there is an important supporting role for governments to play in enhancing cable resilience and security. This is consistent with historic practice: critical infrastructure has long been protected by a combination of private and public actors.

Recommendations

Governments today can take the following steps to support subsea cable security:

Promote redundancy as the most critical component of resilience by supporting regulatory frameworks that make it easier to build and operate subsea cables—enabling multiple routes between key regions and ensuring that data can continue to flow even if one or more cable routes are temporarily compromised.

Encourage geographic diversity of cable landings to reduce concentration risk and avoid single points of failure.

Prepare for large-scale disruptive events through public-private coordination, advance planning, tabletop readiness exercises, and prioritization for restoration of service.

Reform cabotage rules and other maritime regulations to enable efficient use of specialized repair ships and faster restoration following cable damage.

Support data-driven planning for future repair capacity, including assessments of repair ship retirement, anticipated demand, and repair ship market stability considerations.

Leverage national security and defense capabilities, where appropriate, to support protection and repair activities, including the use of military assets to assist or escort commercial repair operations.



Preparing for the next *security transition*



Quantum-safe readiness for organizations: Start preparing today

Quantum computing threatens the cryptography that protects an organization's data, identities, software, and digital services.

That risk is not limited to a future point when quantum computers can break today's public-key algorithms: threat actors can store encrypted information now for later decryption, a capability known as harvest now, decrypt later (HNDL). That means sensitive data captured today may still need protection years from now.

The transition to post-quantum cryptography (PQC) requires organizations to modernize the cryptographic foundations they depend on. Sensitive data, identity systems, signing infrastructure, software updates, and service-to-service trust pathways all depend on these cryptographic foundations. Organizations need to discover, modernize, and make them agile before the transition becomes urgent. PQC standards and some implementations are already available, while support varies across products and vendors.

For practical guidance on accelerating quantum-safe readiness, see the Quantum Safe Program resources (https://aka.ms/QSP/accelerate and https://aka.ms/QSP/inventory).

Quantum safety is a migration challenge, not a single replacement project

In the future, sufficiently capable quantum computers are expected to break widely deployed algorithms such as RSA and elliptic-curve cryptography (ECC), threatening web security, software integrity, and identity verification. Moving to PQC will require a multi-year transformation across protocols, certificates, core libraries, and the systems that depend on them. Done well, it is also an opportunity to reduce exposure from legacy cryptography, standardize on modern primitives, and build crypto-agility so future cryptographic changes are less disruptive.

Organizations can build a cryptographic inventory in four steps

Becoming quantum-safe is a large-scale software and infrastructure modernization effort that will eventually touch nearly every device, service, application, and dependency in an organization. The journey starts with discovering where cryptography is used, how it is deployed, who owns it, and which business processes depend on it. Certificates embedded in appliances, custom implementations in legacy applications, hard-coded keys, outdated libraries, and undocumented protocol choices are common sources of hidden risk. Without an accurate, continuously updated cryptographic inventory, PQC migration plans quickly become guesswork.

A practical quantum-safe plan usually follows four stages:

- **Discover** where cryptography is used across your enterprise, including the systems, services, protocols, certificates, keys, and secrets that underpin trust.
- **Evaluate** which assets rely on quantum-vulnerable algorithms, which data has long confidentiality requirements, and which vendors or platforms have upgrade paths available.
- **Plan** remediation based on business criticality, exposure, and operational feasibility, with particular focus on identity systems, signing services, update mechanisms, and other high-impact trust pathways.
- **Implement** the required updates in a controlled way.

This work also pays off immediately for AI security

Many weaknesses that attackers exploit against models and agents are rooted in identity controls, exposed secrets, and brittle trust chains. Closing those gaps helps prevent credential theft and token replay, limits misuse of service principals and API keys that grant access to AI endpoints and tools, and hardens integrity paths for signing, updates, model artifacts, and plugins. It also reduces the impact of prompt injection-to-action scenarios by tightening least-privilege access, key scope, and approval boundaries for high-impact actions.

By understanding which cryptographic assets protect the organization's most important trust pathways, security teams can use that visibility to drive prioritized modernization. Effective inventories go beyond a spreadsheet. They capture certificates and their chains, key types and sizes, protocols in use, cryptographic libraries and versions, and where keys live. They also track ownership, change windows, and dependencies so teams can coordinate upgrades. As environments evolve, inventories must stay "living" through automation and continuous scanning, enabling security teams to measure progress, validate policy, and prevent regression.

Quantum-safe readiness for organizations: Start preparing today (continued)

Preparing for 2029 and beyond: From readiness to action

The next few years will determine whether the transition to PQC is orderly or disruptive. The remainder of this decade is likely to be a critical readiness period as quantum-safe capabilities become broadly available and organizations begin planning large-scale adoption. The priority now is to convert readiness into execution: modernize infrastructure by adopting current operating systems with built-in PQC capabilities, complete cryptographic inventories, assign ownership for critical systems, test hybrid PQC, and engage suppliers early so adoption can accelerate as PQC standards and products mature.

Securing the quantum future through policy

Governments are moving from recognizing the risk posed by PQC to accelerating the transition to quantum-safe cryptography.

As of June 2026, more than 25 governments have issued PQC guidance or mandates. While these actions demonstrate that PQC has become a policy priority across regions, the challenge now is translating this momentum into implementation at scale.

Recent policy developments reflect growing international alignment around the need for quantum-safe migration. This includes planning to migrate high-value government assets and high-impact systems to PQC-based cryptography and mitigate as much quantum risk as feasible, securing critical infrastructure and pursuing phased approaches that align with international standards. This momentum is also evident beyond G7 economies, with countries such as India, Singapore, and Malaysia developing guidelines to help both the public and private sector.

Recommendations for governments

Governments play a critical role in enabling a quantum-safe future through strong collaboration with industry and effective policies. To accelerate readiness, we recommend governments take the following actions:

Establish quantum safety as a national cybersecurity priority. Position quantum-safe cryptography as a strategic imperative and embed it into national cybersecurity frameworks.

Align quantum-safe strategies across jurisdictions, including by adopting international standards. Harmonize public policies, standards, and transition timelines. The G7 should lead by expanding its financial sector post-quantum cryptography workstream to align G7 members' broader quantum-safe strategies. Global standards development will help overcome fragmented, region-specific approaches that hinder interoperability, innovation, and security.

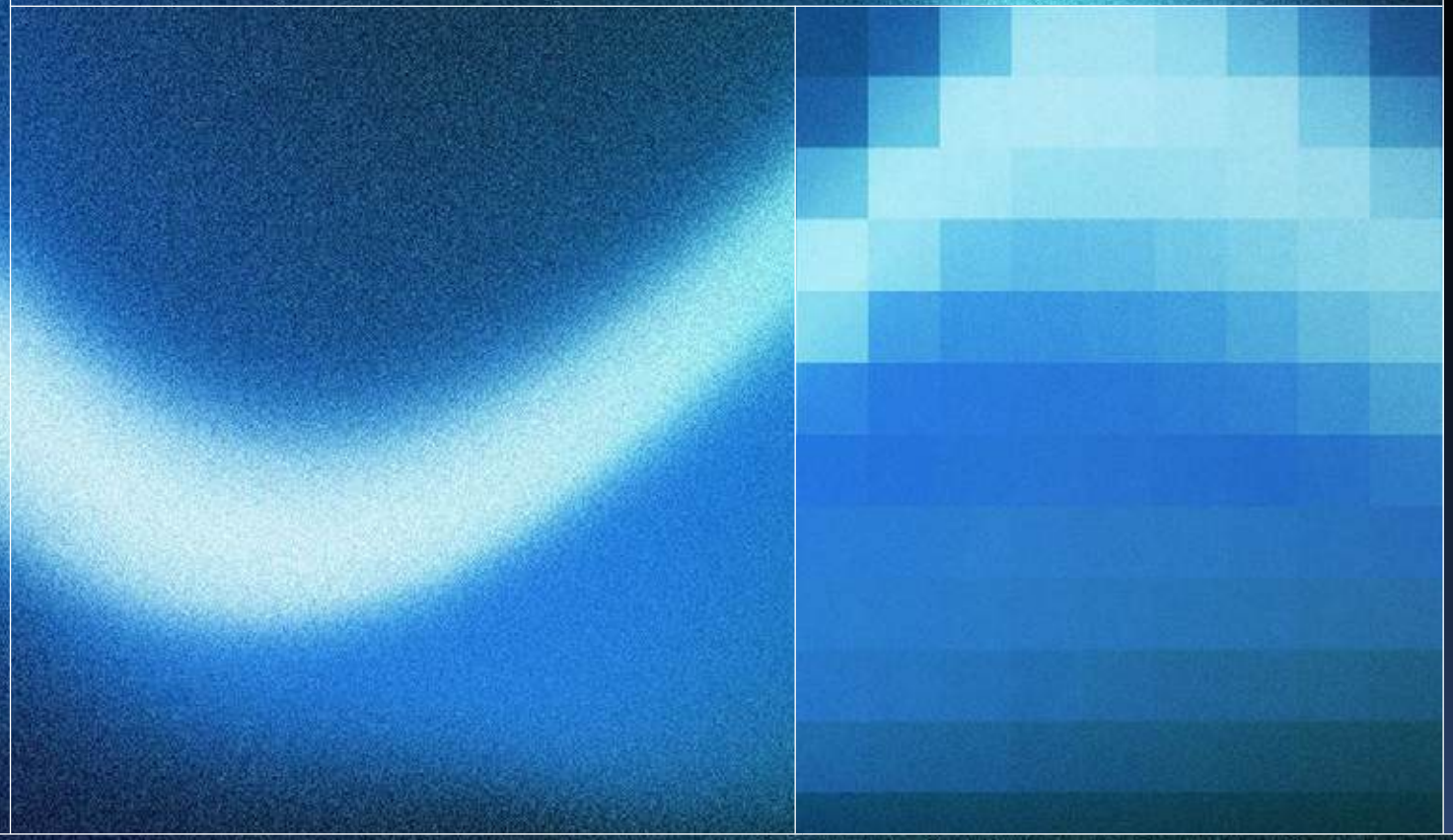
Set early and progressive timelines. For instance, US policy for national security systems mandates starting in 2027 that products should not be procured if they are not upgradeable to become quantum-safe by 2030.

Lead by example with transparent transition plans. Publish and regularly update government transition roadmaps—including timelines, milestones, and budgets—to foster knowledge sharing and best practices.

Raise awareness and build workforce capacity. Educate the public and critical infrastructure sectors on quantum risks and readiness. Invest in skilling programs to equip the workforce for a quantum-safe transition.



Appendix



Glossary 100

References 102

Glossary

Adversary-in-the-middle (AitM) attack Occurs when an attacker reroutes communication between two users through the attacker's computer without the knowledge of the two communicating users. Each user in the communication unknowingly sends traffic to and receives traffic from the attacker, all the while thinking they are communicating only with the intended user.

Attack surface An information system's characteristics that permit an adversary to probe, attack, or maintain presence in the information system.

Authentication The process of verifying the identity of an entity (user, process, or device).

Critical infrastructure The systems and assets, whether physical or virtual, so vital to society that the incapacity or destruction of such may have a debilitating impact on the security, economy, public health or safety, environment, or any combination of these areas.

Cyber espionage The unethical act of violating the privacy and security of an organization in order to leak data or disclose internal/private/confidential information. Cyber espionage can be performed by individuals, organizations, or governments for the direct purpose of causing harm to the violated entity to benefit individuals, organizations or governments.

Disinformation is deliberate, often orchestrated, attempts to confuse or manipulate people through delivering dishonest information to them.

Distributed denial of service (DDoS) attack An attack that prevents or impairs the authorized use of information system resources or services. A distributed denial of service (DDoS) is a type of attack that uses multiple networked machines to overwhelm a host connected to the internet, temporarily or permanently disrupting service or preventing access.

Exfiltration The unauthorized transfer of information from an information system.

Misinformation is misleading information, either created or disseminated without manipulative or malicious intent.

Multifactor authentication (MFA) Refers to an additional security layer beyond just username and password. One way it is described is the user name and password shows who you are based on something you know, and MFA shows who you are by something you have. The most common example of MFA is the code you get in a text on your phone when you are logging into sites like your bank or GitHub.

Nation-state actor/Nation-state threat actor A cyber threat actor that is directed, sponsored, or supported by a national government and conducts cyber operations to advance that government's strategic, political, military, economic, or intelligence objectives.

Node Package Manager (npm) A package manager for JavaScript and Node.js that allows developers to install, manage, and share software packages. Because npm packages are widely used across the software supply chain, attackers may attempt to distribute malicious code through compromised, fraudulent, or malicious packages.

NT LAN Manager (NTLM) A legacy Microsoft authentication protocol used to authenticate users and devices on a network. Because NTLM relies on older authentication mechanisms, it is commonly targeted by attackers seeking to steal credentials, conduct relay attacks, or move laterally within compromised environments.

Password spray High-volume attempts using a large number of common passwords to compromise sourced account information to authenticate and gain access to a network, often leveraging big data algorithms and extensive automation for rapid execution.

Passwordless authentication is a security method that allows users to verify an identity without using a password.

Phishing A digital form of social engineering to deceive individuals into providing sensitive information.

Phishing kit A collection of tools assembled to make it easier for individuals with little or no knowledge of phishing practices to launch a phishing exploit.

Quishing Phishing that uses malicious QR codes to direct victims to fraudulent or harmful content.

Ransomware A type of cybersecurity attack that destroys or encrypts files and folders, preventing the owner of the effected device from accessing their data. The cybercriminal can then extort money from the business owner in exchange for a key to unlock the encrypted data. But, even when paid, cybercriminals may not provide the key to return access to the business owner.

Ransomware-as-a-service (RaaS) An arrangement between an operator, who develops and maintains the tools to power extortion operations, and an affiliate, who deploys the ransomware payload. When the affiliate conducts a successful ransomware and extortion attack, both parties profit.

Remote Desktop Protocol (RDP) A protocol for remotely connecting to computers running Windows.

Resilience The ability to adapt to changing conditions and prepare for, withstand, and rapidly recover from disruption.

Security posture Refers to an organization's overall cybersecurity strength and how well it can predict, prevent, and respond to ever-changing cyber threats.

Spear phishing is a targeted phishing attack that involves highly customized lure content. Attackers typically do reconnaissance work by surveying social media and other information sources about their intended target. Spear phishing may involve tricking targets into signing in to fake sites and divulging credentials.

Glossary (continued)

Spoofing Faking the sending address of a transmission to gain illegal (unauthorized) entry into a secure system. Website spoofing involves creating a duplicate version of a website that appears to be the original. Hackers use legitimate logos, fonts, colors, and functionality to make the spoofed site look realistic. Even the URL can appear genuine.

Token replay attacks The use of tokens that were issued to a legitimate user by an attacker that has possession of said tokens. Tokens are commonly obtained via malware, for example by exfiltrating the cookies from the user's browser, or through advanced phishing methods.

Transmission control protocol (TCP) attacks Cyberattacks that exploit TCP network communications to disrupt connections, intercept or manipulate traffic, or deny access to services.

Vishing A form of phishing that uses phone or voice communications to deceive people into revealing sensitive information or taking an action.

Zero-day vulnerability A previously unknown or unaddressed software vulnerability for which a security update or comprehensive mitigation is not yet available.

Zero Trust A security model based on explicitly verifying access requests, using least-privilege access, and continuously evaluating risk while assuming that compromise is possible.

References

Section 1: AI

1. LLMs used inside malware | CERT-UA |vCommunity Response Team - Ukraine | JUL 2025
2. AI-powered malware hit 2,180 GitHub accounts in "s1ngularity" attack | Bleeping Computer | SEP 2025
3. Disrupting the first reported AI-orchestrated cyber espionage campaign | Anthropic | NOV 2025
4. Our evaluation of OpenAI's GPT-5.5 cyber capabilities | AI Security Institute | APR 2026
5. The Hugging Face incident and the road ahead | Open AI | AUG 2026
6. Measuring AI Agents' Progress on Multi-Step Cyber Attack Scenarios | arXiv Computer Science > Artificial Intelligence | MAR 2026
7. JADEPUFFER: Agentic ransomware for automated database extortion | sysdig | JUL 2026
8. Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident | Hugging Face| JUL 2026
9. Can AI Attack the Cloud? Lessons From Building an Autonomous Cloud Offensive Multi-Agent System | Unit 42 Threat Research | APR 2026
10. AI Agents Enable Adaptive Computer Worms | arXiv Computer Science > Artificial Intelligence | JUN 2026
11. NadMesh Botnet Analysis: A Product-Grade Threat for the AI Service Era | Xlab | JUL 2026
12. Frontier Governance Framework
13. Efforts to share best practices on AI measurement
14. International Network for Advanced AI Measurement, Evaluation, and Science
15. Frontier Model Forum
16. Advancing AI evaluation with the Center for AI Standards (US) and Innovation and the AI Security Institute (UK)
17. Malicious AI Assistant Extensions Harvest LLM Chat Histories | Microsoft Security Blog | MAR 2026

Section 2: Threat Landscape

18. Mitigating NTLM Relay Attacks by Default | Microsoft MSRC | DEC 2024

Section 3: Cybercrime

19. NIST 800-63B-4 requirements |NIST | National Institute of Standards and Technology |SP 800-63B
20. Defending the gates: How a global coalition disrupted Tycoon 2FA, a major driver of initial access and large-scale online impersonation | Microsoft on the Issues | MAR 2026
21. CISA et al., Unsophisticated Cyber Actor(s) Targeting Operational Technology, May 6, 2025
22. Pro-Russia Hacktivists Conduct Opportunistic Attacks Against US and Global Critical Infrastructure | America's Cyber Defense Agency | DEC 2025
23. Setting spyware standards after the Pegasus scandal PDF | European Parliament
24. CDT Europe Report: EU Export Controls Are Failing to Address the Spread of Spyware - Center for Democracy and Technology
25. Mythical Beasts Investigating the Role of Intermediaries | Atlantic Council | MAR 2026

Section 4: Resilience

26. Media Integrity and Authentication: Status, Directions, and Futures | arXiv Computer Science > Artificial Intelligence | FEB 2026
27. Media Authenticity Methods in Practice: Capabilities, Limitations, and Directions | Microsoft Research Blog | FEB 2026
28. A new study explores how AI shapes what you can trust online | Microsoft Signal | FEB 2026
29. Advancing digital content transparency and authenticity | Coalilton for Content Provenance and Authenticity
30. Unify now or pay later: New research exposes the operational cost of a fragmented SOC | Microsoft Security Blog | FEB 2026
31. NIST AI Risk Management Framework
32. EU AI Act
33. International Community of the Red Cross's (ICRC) Digital Emblem
34. Akrites
35. Microsoft announces ARC Initiative to strengthen cybersecurity in Kenya | Microsoft On the Issues | MAY 2025
36. Common Vulnerabilities and Exposures (CVE) Program
37. CVE Numbering Authority (CNA)
38. Accelerating the quantum-safe timeline | Microsoft Security Blog | JUN 2026
39. Building your cryptographic inventory: A customer strategy for cryptographic posture management | Microsoft Security Blog | APR 2026
40. Implementation of Quantum Safe Ecosystem in India - Report of the Task Force PDF | Department of Science & Technology

2026 Microsoft Digital Defense Report

For more news on cybersecurity, visit:

https://microsoft.com/corporate-responsibility/cybersecurity

For more report insights, visit:

https://microsoft.com/mddr

For more news on cybersecurity policy, follow us on LinkedIn:

https://aka.ms/MOILinkedin

For insights and trends for security leaders, visit:

https://www.microsoft.com/security/security-insider

A Microsoft Threat Intelligence report

October 2026