CYBER HERO FRONTLINE

magazine by THREATLOCKER®

Fighting on the AI frontier
New AI models are disrupting the cybersecurity landscape

The hit list
Why Allowlisting is the key to fighting unknown threats

The patient machine
Tokyo's craft and philosophy shape a unique tech culture

Even your best employees make mistakes.

ThreatLocker® ensures those mistakes don’t cost you.

Human error is a major cybersecurity risk. Intentional or unintentional threats pose the same danger. So, stop trusting and start controlling. Decide exactly what runs, when, where, and how. Deploy in hours to days and stop threats before they start

THREATLOCKER®
ZERO TRUST PLATFORM

Ready to take control?
Schedule a demo today to see how to transform your security.

INSIDE THREATLOCKER®

Hello and welcome to another edition of Cyber Hero® Frontline.

With attack vectors continuing to expand through the development of AI-gifted capabilities, our objective here at ThreatLocker remains the same: streamlining Zero Trust controls so your organization can prevent attacks from executing, no matter who or what the culprit is.

This philosophy continues to drive everything we do.

The cybersecurity conversation often focuses on detecting these increasingly overwhelming threats, but our belief is firm: The most effective defense is greatly restricting what attackers are allowed to do in the first place.

Away from the AI conversation, but not entirely unrelated, robotics is starting to play a major role in boosting efficiency across industries that rely on physical labor (which will be discussed in this issue). These developments mean physical infrastructure is regularly meshing with IT environments, so maintaining clear security boundaries and air-gapping systems is vital.

There are tremendous opportunities for innovation, but also a requirement for organizations to rethink how permission is established across increasingly connected environments. Whether attacks are initiated by a human or an autonomous AI agent, denying unnecessary execution, limiting application behavior, and enforcing least privilege remain the strongest mitigation.

Technology will continue to evolve and threat actors will surely take advantage, but the fundamentals of Zero Trust remain the same. That's why, beyond the mission to provide the most comprehensive Zero Trust solutions on the market, our commitment to education continues.

Every webinar, customer event, training session, and new edition of Cyber Hero Frontline is developed with the same goal: providing practical guidance that your organization can immediately apply to strengthen its security posture.

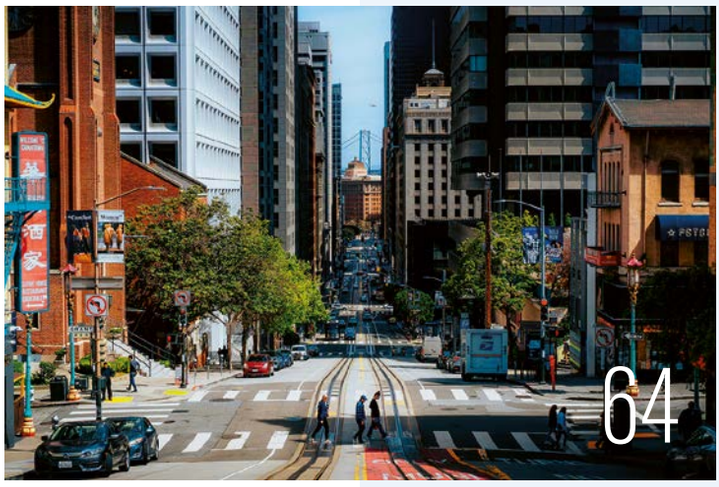
Thank you for joining us. These pages will provide valuable insights to build a stronger, more resilient security posture.

Yours,

Sami Jenkins,
ThreatLocker Co-founder
and Chief Operating Officer

Danny Jenkins,
ThreatLocker Co-founder
and Chief Executive Officer

TABLE OF **CONTENTS**



LIFESTYLE

CYBER HERO HIGHLIGHT

SECURITY INSIGHTS

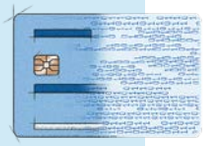
EXPERT INTERVIEW

GLOBAL POLICY

TRENDING

PEACE OF MIND

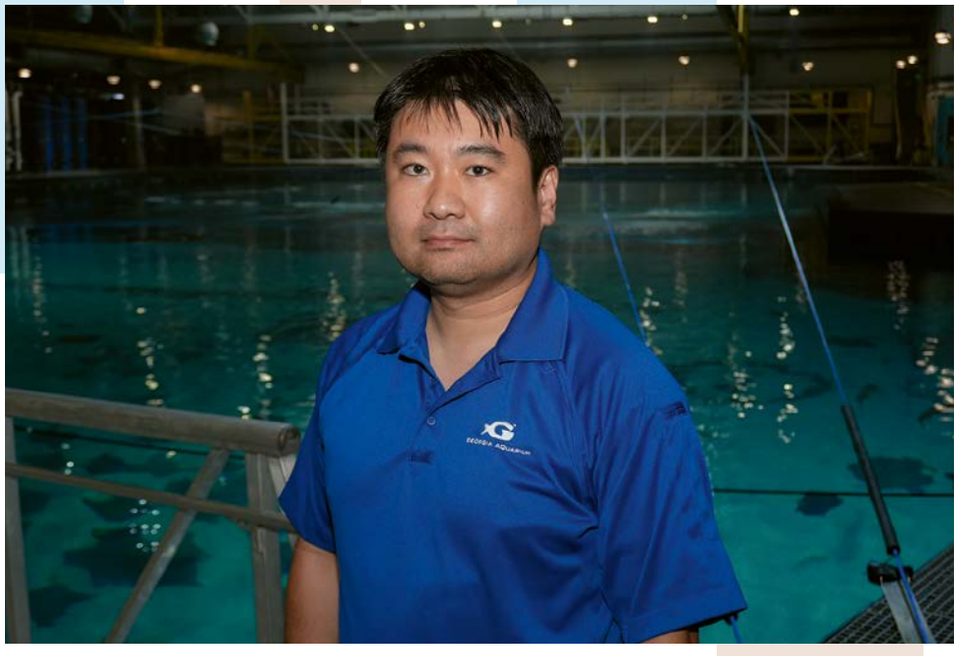
CYBER TOOLS



“

Tokyo is the best city in the world to witness human-machine coexistence as a lived daily reality

Sara Filipčić, Founder and Institute Director, Humane AI & Robotics
P. 38



“

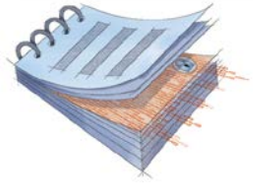
We are sometimes very focused on the tools and solutions. I want to know how people work first

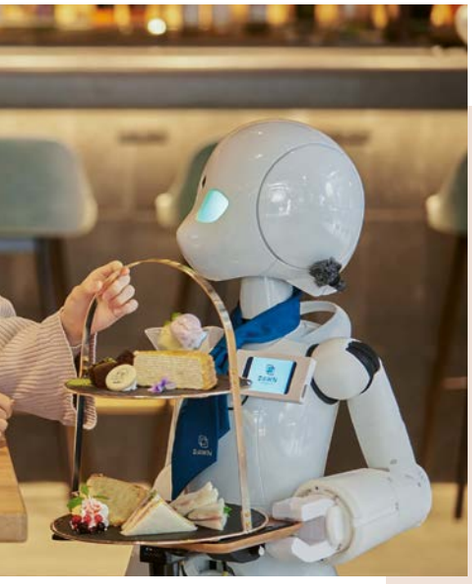
Yuta Kitabayashi, Senior Manager of IT and Systems Engineer, Georgia Aquarium
P. 30

“

Make sure one bad endpoint can’t become a bad day for the whole organization

Kieran Human, Lead Cybersecurity Engineer, ThreatLocker®
P. 24





“

If all I need is publicly available information, I can probably get into anybody’s account

Rachel Tobac, CEO and Co-founder, SocialProof Security
P. 56

“

We don’t start with a trend or a product category. We start with a problem we want to solve

Rob Allen, Chief Product Officer, ThreatLocker
P. 44

“

Each innovation brings enormous opportunity, and also creates another place where trust can be abused if it isn’t properly controlled

Danny Jenkins, CEO and Co-founder, ThreatLocker
P. 6

ON MY **MIND**

BY DANNY JENKINS

Artificial intelligence continues to generate concerns that dominate the conversation in cybersecurity circles, but I don't believe AI itself is the biggest issue.

What concerns me more is how rapidly trust boundaries are changing. Organizations are connecting operational technology to cloud platforms, adopting increasingly autonomous AI systems, granting long-lived access tokens to third-party services, and relying on open-source models that anyone can download and modify.

I do not intend to frame it in any other way: Innovation can bring enormous opportunity, but it also creates another place where trust can be abused if it isn't properly controlled.

Here are a few topics that have been on my mind recently.

1 AI is accelerating faster than security controls

Initially, the most capable frontier AI models remained largely under the control of a handful of organizations. That is quickly changing as capable open-source models become available to anyone with sufficient computing resources. Democratization will fuel opportunity. It will also make advanced offensive capabilities significantly more accessible.

We are already seeing AI automate reconnaissance, vulnerability research, phishing campaigns, malware development, and social engineering. As models continue improving, those capabilities become faster, cheaper, and easier to scale.

Attackers now have fewer barriers preventing them from using increasingly sophisticated AI. Trying to detect every new technique these systems generate simply isn't sustainable. Instead, organizations need to reduce what AI-powered attacks can achieve after they reach an environment.

If all executions on the application and user level hinge on explicit approvals, AI-generated attacks will lose much of the freedom they depend on. So rather than trying to predict what the next development is, the conversation should be about ensuring your environment only allows what is required and nothing more.

2 Guidance from the Five Eyes intelligence group reinforces something we've known for years

One of the most encouraging developments recently wasn't a new security product or breakthrough technology. It was seeing all the cyber agencies within the Five eyes intelligence sharing alliance (U.S., U.K., Australia, Canada, and New Zealand) reinforce a principle that many security professionals have advocated for years: Implicit trust should not be acceptable.

Their latest guidance acknowledges that AI will continue accelerating both defensive and offensive capabilities. Rather than attempting to predict every possible attack, organizations should focus on reducing opportunities through stronger architectural controls.

As AI systems become more autonomous, this mindset should be a base requirement. Autonomous software can make decisions extraordinarily quickly. If those systems inherit excessive permissions, broad network access, or unrestricted application behavior, they can amplify mistakes just as quickly as malicious activity.

Whether those actions originate from a human attacker or an AI agent ultimately becomes less important. Control remains a constant requirement.

3 Connecting IT and OT needs to be strictly monitored

We're seeing more organizations merge traditional IT systems with operational technology. Manufacturing, healthcare, utilities, and transportation are all benefiting from better connectivity, automation, and data visibility.

Done correctly, this transformation creates enormous operational advantages. Unfortunately, it also removes many of the usual barriers that historically separate business systems from physical operations.

An attacker who compromises a user workstation should never be able to freely interact with production equipment, industrial controllers, or operational networks simply because they're connected. Yet many environments still contain far more trust between IT and OT than they probably realize.

Organizations should focus on maintaining clear security boundaries as these environments become increasingly connected: segmenting networks, limiting communications between systems, ensuring applications can only perform approved actions, and applying least privilege consistently across both operational and corporate environments.

Physical systems may now share data with cloud services and AI platforms, but they shouldn't automatically share trust. That is the difference between modernization and unnecessary risk.

4 Token-based access should be time-sensitive

The recent discussion surrounding the compromise involving long-lived OAuth tokens highlights an issue that doesn't receive enough attention.

Authentication is only part of the security equation.

Many organizations invest heavily in strong passwords, phishing-resistant MFA, and secure login processes, but once access has been granted, tokens often remain valid for weeks or even months. If those tokens are stolen, attackers may continue accessing cloud services without ever needing to authenticate again.

This is an architectural challenge that affects many SaaS environments. Ensure your tokens don't remain valuable for longer than required.

Organizations should regularly expire tokens, continuously tie devices to a specific and approved IP address, limit where tokens can be used, and remove standing trust wherever possible. Shorter token lifetimes dramatically reduce the window of opportunity available to attackers if credentials or sessions are compromised.

Zero Trust is defining whether a user, device, or application should be able to take an action. I don't believe I am alone in the expectation that we will see it become more important as AI services and third-party integrations continue in prominence.

> "Organizations should regularly expire tokens, continuously validate devices alongside identities, limit where tokens can be used, and remove standing trust wherever possible

THE HIT LIST

Fighting unknown threats needs the confidence to allow what is required and block the rest

450,000

new malicious and potentially unwanted programs are identified each day‡

Every organization has applications that are as close as possible to being trusted. They also have applications that are tolerated and, in many cases, those that fly under the radar, remaining completely unknown to IT until they surface. Few would call that an acceptable level of risk, but it is an easy mode to slip into. It stems from users installing free tools to solve a problem, temporary browser extensions slipping into daily use, or legitimate applications carrying flaws nobody has accounted for.

Under a security regime that relies entirely on blocklisting, administrators attempt to manage that risk by stopping bad applications from running. Even describing the theory of blocklisting highlights its flaw: It means defining the applications, file types, or behaviors that should be blocked, then allowing everything else to run. Blocklisting is useful for off-the-cuff business action because it is essentially friction-free for users, but it struggles when a threat has not yet been identified.

Given the rapidly evolving nature of often AI-driven malware, the growth of zero-day exploits, and the weaponization of legitimate tools, blocklisting leaves too much room for unknown threats to execute, regardless of how comprehensive and well-maintained the underlying list may be.

LOOKING IN THE MIRROR

Allowlisting reverses the control model, following a deny-by-default Zero Trust approach. Administrators decide what is allowed, and everything outside that list is blocked by default. Applications, scripts, and libraries must be approved before they can execute. If the software is unknown, unauthorized, or simply unnecessary, it cannot run. In the case of malicious applications, allowlisting means the attack chain can be stopped before it starts.

Blocklisting does not fit neatly with the principles of a Zero Trust environment or any environment that requires a high level of application control. A new ransomware payload that does not trigger a block can execute freely. Users have free rein to install extensions or run scripts that have not yet caught the attention of IT. Blocklisting is not chaos by default, but its effectiveness depends on an incredibly noisy and flawed management process.

MAKING MANAGEMENT EASIER

Some see allowlisting as problematic because of the management precision it requires. Blocklisting is blunt: It swings at known problems. Allowlisting is more precise and, at the outset, means making difficult decisions and putting clear rules into action. In older allowlisting systems, allowlists had to be built manually and updated whenever an approved application was updated. Tight control meant more friction, creating the opposite of the blocklisting problem: rather than being allowed to do too much, users struggled to do anything at all.

ThreatLocker **Allowlisting** is designed to facilitate the tight control that it demands, removing much of the operational burden. When the ThreatLocker agent is deployed, it can build a picture of its environment by automatically cataloging existing applications and dependencies. Allowlisting already recognizes more than 15,000 applications, giving administrators immediate visibility into what is present and offering advice on potential policy decisions in their operating environment—without having to build everything from scratch.

ONGOING CONTROL WITHOUT PAIN

Allowlisting makes the process simple once the environment has been cataloged, giving IT teams a clear view of the approved software estate. This is a direct way to prevent shadow IT issues, as ThreatLocker prevents unauthorized tools from appearing on endpoints and running in the background. If a new tool is needed, users can request access directly from the endpoint, giving IT a chance to review internally, offer an alternative, or work with the Cyber Hero® Team to understand the potential impact of allowlisting the application in question.

31%

of breaches start with software vulnerabilities‡

Through the ThreatLocker **User Store**, end-users can install pre-allowed applications as and when they need them, reducing the friction on IT teams.

Without the pain of setup and ongoing administration, Allowlisting reveals itself as a governance tool that doubles up as a very effective defense mechanism. Admins can determine what runs, when, where, and by whom, so Allowlisting is well suited to maintaining a deny-by-default stance that supports regulatory compliance with key governmental frameworks or to satisfying auditors by clearly demonstrating what is allowed and why.

A PRACTICAL APPROACH TO A FAMILIAR GOAL

For endpoint security within the unique environment that every organization presents, the strongest tools are usually the clearest.

Allowlisting allows for all the nuance a business needs, but boils application execution down to a binary rule: If it is on the list, it can run; otherwise, it cannot. Instead of constantly chasing down the bad, organizations get to define what good looks like. If that definition changes, ThreatLocker allows them to modify their stance cleanly and quickly. In an environment where software can appear silently and attackers can move quickly, that kind of control makes all the difference.

THE WORKING SANDBOX

AI tools are powerful, but their power must be contained—if they run on the endpoint, they need a safe place to play

Many of the most useful endpoint tools are local, and that is what makes them useful. For a developer under pressure, agentic AI—delivered through tools such as Claude Code, Copilot, and Cursor—can deliver a strong productivity lift. But installed tools come with inherent risk, as their use may grant them broad filesystem access when run in a normal local workflow.

A laptop used for development is rarely employed for one single task. It likely also contains sensitive information that AI engines should not be able to access. Does AI running on a user's desktop need to reach Secure Shell (SSH) keys, API tokens, or environment files? Should it be able to work with unrelated client repositories, cached credentials, or personal files? The answer is inevitably no. The user needs these, but if an AI coding tool runs with the same permissions, the boundary between the intended project and everything else can become dangerously thin.

AI coding agents can and do read untrusted content and act on it. A poisoned or somehow ambiguous project instruction can become part of the tool's working context. A single one of the tool's inadvertent hallucinations can become part of the project in perpetuity. If the agent is also allowed to run commands, call other tools, or make network requests, a bad instruction—malicious or accidental—can inadvertently move from text into action.

ESCAPING CONTAINMENT

AI tools are a subtle threat even in everyday operation. Attacked directly, they can be particularly damaging. Security researchers have repeatedly shown how indirect prompt injection can push agents toward unauthorized behavior. Recent work on development tools running on the AI-standard Model Context Protocol (MCP) found uneven protection across real-world clients, with risks including hidden parameter abuse and unauthorized tool invocation.

Another recent study of AI agents working in integration and delivery pipelines described prompt-injection attacks against live GitHub workflows, including credential exfiltration, configuration manipulation, and availability attacks. Not every coding assistant is unsafe, but when powerful assistants inherit permissions, they also inherit risk.

AI-enabled tools do not behave consistently by design, so building security around them means determining what they can do and what they should never be allowed to do

In 2026, PocketOS Founder Jer Crane said an AI coding agent working through Cursor deleted a production database and its backups after misreading its environment and using an overly broad token. There was no malice involved—large language models (LLMs) and other AI tools do not currently have any real sense of their intentions, and they are notoriously easy to poison—and in this case the agent was not actively trying to disrupt operations. It simply used its available access, which was created by unsafe human assumptions, and exploited through unwitting AI assumptions.

BEYOND TRADITIONAL APPROVAL

A human might ask an AI assistant to "check the config" or "look at the keys for this project." The tool may then infer the wrong location, follow a misleading instruction, or extend its search more widely than the user intended or expected. If it can read everything the user can read, an unintended misdirection may allow it to reach sensitive material. If it is granted permission to write or delete, it becomes significantly more difficult to contain.

Traditional approval processes are poorly suited to this problem. Knowing whether an application is reputable, whether the publisher is recognized, or whether the binary appears clean does not cover this potential activity. Those are valid questions, but it is not enough. AI-enabled tools do not behave consistently by design, so building security around them means determining what they can do and what they should never be allowed to do. They must be placed within a sandbox, where they can play safely.

FREEDOM WITHIN LIMITS

AI agents do not behave like standard software. They are variable in their nature, so while it might seem reasonable to determine their limits with ThreatLocker® **Controlled Application Testing Environment**, those limits may change depending on their prompts. It is reasonable to suggest that AI agents only be allowed to run within virtual machines, to keep them airgapped from production data, but this vastly limits their usefulness and almost defeats the object of using them.

Given that agents are volatile and cannot offer the reliable visibility of flatter software designs, security teams must approach them with the concept of a sandbox in mind—if not the classic sandbox structure. They must, in effect, build a sandbox from scratch, using the tools and materials at hand. If a code assistant or AI agent cannot be trusted to stay in its lane, administrators must ensure that lane is the only place it can be.

This begins with **Allowlisting**, to ensure that only authorized software can run on the endpoint; if an agent heads off to try to download and run an unknown or unapproved tool, it will be stopped in its tracks.

Ringfencing™ restricts the agent's third-party reach, allowing it to interact only in the way administrators feel is safe. Even if it is employing pre-approved tools, the blast radius of any odd or damaging behavior is limited in the same way that it would be at the user level.

Data Storage Access Control is vital to keep data safe, preventing a tool working on one task from roaming into unrelated client data, credential stores, or shared folders. Limiting an agent's privilege—giving it its own permissions, away from the user level—helps stop administrative convenience from becoming unchecked power.

This DIY sandboxing approach bucks tradition, but it suits the reality of AI adoption. Businesses cannot be expected to stop experimenting with new endpoint-based tools, because the productivity case is obvious. But security teams must be able to trust these tools, so technical boundaries and enforceable controls must be put in place to make the endpoint a safe place for AI agents to play.

DANGER IN PLAIN TEXT

Even the most familiar operating system utilities can become execution paths when their functionality expands beyond old security assumptions

Windows' built-in Notepad application was never supposed to be exciting. For decades, that was almost the point. It opened text. It saved text. That was it. Security teams had plenty of other, more obviously capable Windows components to worry about before they began losing sleep over a bare-bones note-taking application. That all changed when Notepad changed.

There has been plenty of criticism of Microsoft's decision to add new features to Notepad in Windows 11—including an understandable furor over what some consider unnecessary functionality, notably AI writing support. And the existence of CVE-2026-20841, patched by Microsoft in February 2026, suggests that the application's upgrade came with certain unintended side effects.

The vulnerability relates to how the application handled Markdown links. In simple terms, an attacker could craft a malicious Markdown file containing a specially formed link. If a user opened the file in Notepad and clicked the link, Notepad could forward the request to Windows in a way that allowed command execution in that user's context.

Markdown itself is not inherently unsafe and was not the root cause of this issue. The problem was the meeting point between Markdown rendering and Windows behavior. A link in a document is just that, until an application comes to hand it over to the OS. Depending on the protocol used, it may open a remote resource, call another application, invoke a local file path, or trigger another handler installed on the machine.

EARNING SUSPICION

CVE-2026-20841 earned its "high severity" rating. Yes, it required user interaction, which is an important limitation, but not the kind of interaction that is especially far-fetched or difficult to achieve. Getting someone to open a document and click a link is the raw material of ordinary phishing. The attack complexity was low, no privileges were required, and the potential impact was high across confidentiality, integrity, and availability metrics. In the right conditions, successful exploitation of the vulnerability could allow attackers to execute malware, steal data, establish persistence, perform lateral movement, or deliver further payloads.

Could this be considered living-off-the-land (LOTL)? Not quite, though it sits close to the boundary. LOTL attacks usually describe adversaries abusing legitimate tools already present on the system to carry out malicious activity during or after compromise. PowerShell, Windows Management Instrumentation (WMI), mshta, certutil, and other native binaries are common examples. The attacker uses what is already there because it blends into normal activity and avoids the need to drop obvious malware. CVE-2026-20841 is better understood as an exploitation of a trusted application vulnerability. A flaw in Notepad's handling of Markdown links created an unintended execution path.

The distinction is there, but it is not exactly comforting. The same trust problem is present in both cases. Defenders often assume that a familiar binary is safe because it belongs to the OS—what could be more benign than Notepad? But attackers do not care whether the route comes from intended functionality or a vulnerability, only whether their attack works.

CONTROLS THAT WORK

Security controls must, therefore, look to behavior rather than reputation. Proper cyber controls would not make the vulnerability irrelevant; they would limit its impact. In this case, patching with a ThreatLocker® capability like **Patch Management** is the first step. Microsoft issued a fix, and organizations with disciplined patch management reduced exposure quickly. But patching alone is a race once a vulnerability has been disclosed. It does not solve the problem of zero-day exploits or machines that fall outside normal policy.

Controls must assume that some systems will remain vulnerable longer than planned. Application control through **Allowlisting** is the next layer: If even a simple application like Notepad could become an attack vector through a malicious Markdown link, deny-by-default enforcement prevents further action. The attacker may exploit the click, but they do not automatically get to run whatever they like. Unknown executables, unapproved scripts, and untrusted child processes should not be granted free movement simply because the initial interaction came through a Microsoft application.

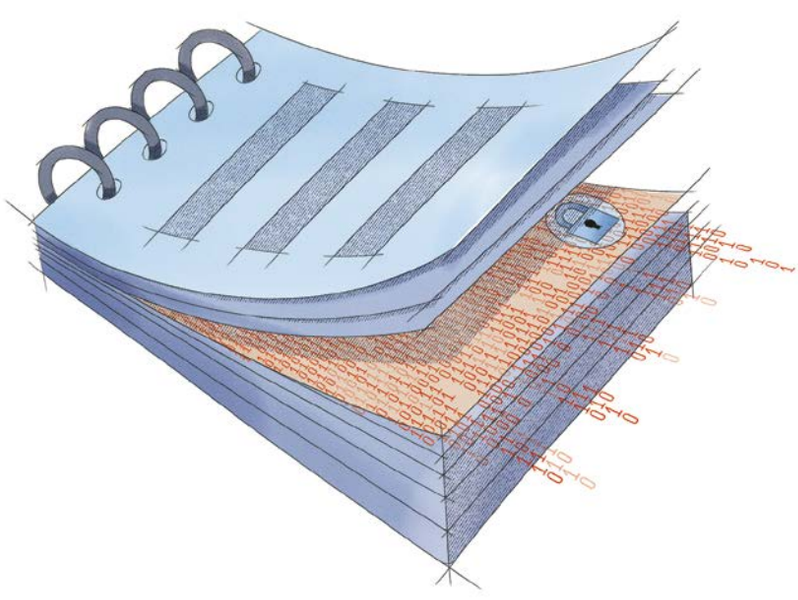
When common applications are constrained to expected behavior, abuse becomes harder and abnormal activity becomes clearer

The rule of least privilege, enforced through **Privileged Access Management,** also matters. The exploit runs in the user's context, so the user's rights define its potential blast radius. A standard user with tightly controlled write access offers a less attractive foothold than a user with local admin rights, broad file permissions, and access to shared locations. Permission boundaries can turn a serious incident into a contained event.

Ringfencing™ adds another practical layer of defense. Trusted applications should not have unlimited freedom to call other tools, access sensitive locations, or communicate wherever they like. Notepad may need to open text files. It does not need broad authority to launch scripts, reach administrative paths, or start unexpected processes. When common applications are constrained to expected behavior, abuse becomes harder and abnormal activity becomes clearer.

LESSONS FROM THE ORDINARY

CVE-2026-20841 is a useful warning because it came from the least dramatic place possible. Notepad is, or at least once was, relatively ordinary. Modern security cannot reserve suspicion for obviously risky software while granting permanent trust to old favorite applications. Everything must be treated with a proportionate level of respect because the attack surface extends beyond where the dangerous tools live. It is also where safe tools become more capable, more connected, and more trusted than they deserve.



THE HUMAN **DEDUCTIBLE**

The insurance industry is built on trust—and criminals have learned where that trust is thinnest

There is a core promise at the heart of the insurance business. If a customer makes a claim, their insurer essentially pledges to recognize them, know—or find out—everything that matters, and respond appropriately. Data swings every claim decision in the right direction, toward a legitimate claimant or away from a fraudster. But the more digital and data-heavy the insurance model becomes, the more chances there are for criminals to test the strength of the sector's defenses. Insurers do not want to break their promise, but keeping it is by no means a given.

There is high value in an attack on an insurer. The dark web offers up handsome bounties for identity information, financial details, health data, family relationships, property records, and so forth—essentially, everything insurance companies need to do their job, and are trusted to hold. Some of that information might also prove promising for more direct fraud or extortion. And since most of it cannot be changed once exposed, a breach results in long-lasting repercussions.

SECTOR IN THE SIGHTS

The sector is firmly in the sights of cybercriminals. Aflac disclosed suspicious activity on its U.S. network in June 2025 and later said personal information connected to about 22.65 million individuals had been involved. Allianz Life confirmed that a malicious actor had used social engineering techniques to gain access to a third-party cloud customer relationship management (CRM) system, affecting the majority of its 1.4 million U.S. customers. Erie Insurance also suffered a cyber incident in June 2025, disrupting customer and agent services before full operations were restored in July.

Not identical attacks, but all symptomatic of the same problem: Insurance is a connected business that often depends on people and systems outside the core organization. That makes the boundary harder to defend, and the consequences harder to contain. And the breach rarely begins where the security team is watching.

Scattered Spider—the loosely organized criminal network that Google's Threat Intelligence Group (GTIG) flagged in June 2025 as actively targeting insurance companies—does not typically arrive by crashing through a firewall. It is more likely that its operatives will use social engineering techniques, calling a help desk to convincingly claim they need a password reset and that they are locked out of their authenticator. If the person on the other end follows habit rather than protocol, the attacker has all the rights and privileges of the account they have breached.

THE HUMAN IN THE BREACH

The insurance industry has spent years hardening its technical perimeter. Human weakness threatens that, and attackers know it. Scattered Spider—also tracked as UNC3944—built its reputation targeting telecoms and retail before turning its attention to financial services. The group also applies techniques like SIM-swapping and multi-factor authentication (MFA) fatigue attacks, not technically complex, but patient and, indeed, scattered. Whatever the weapon, an attacker who can simply talk their way into a device enrollment has bypassed everything the security budget was spent on.

The service desk is therefore a serious vulnerability in any insurance operation. As is a busy contractor onboarding flow, a support queue running at peak volume, or any process where speed and helpfulness are rewarded over caution. Identity, wherever it is being verified under pressure, is the point at which this kind of campaign looks for its opening.

Broker and agent access creates a related problem. Insurance distribution depends on external relationships, and these require live system access. Improperly managed, that access may be broader than the role requires, outlast the commercial relationship itself, or lack appropriate monitoring. Should an attacker obtain a compromised broker account, they may go undetected for longer than usual, since such an account generates activity that already appears to be normal business.

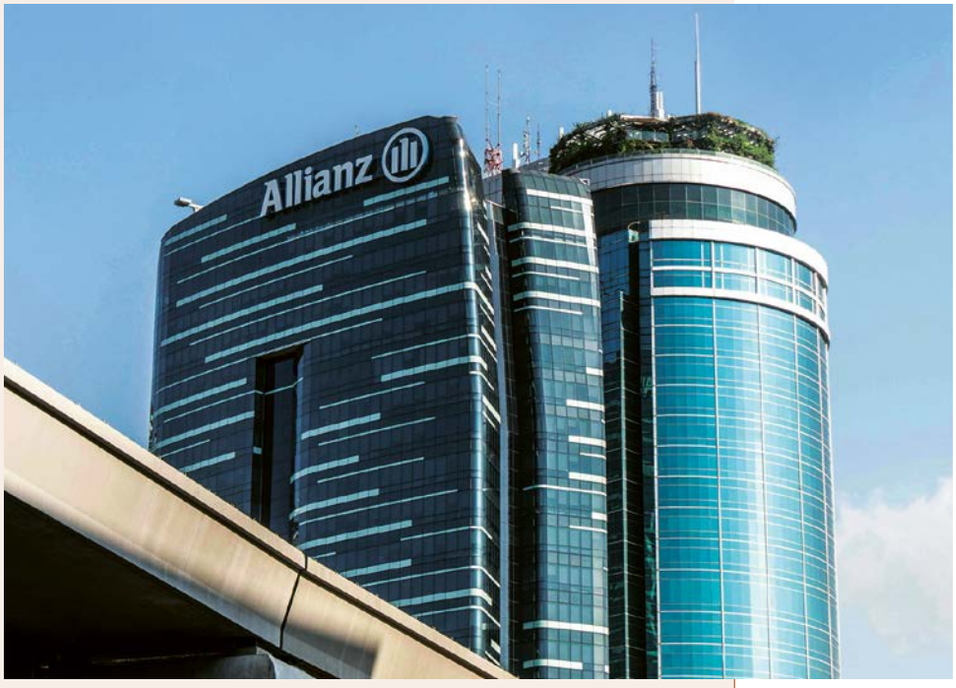
USD 40 million

—the ransom CNA Financial is reported to have paid following a 2021 ransomware attack‡

SUPPLY CHAIN EXPOSURE

Insurers rarely deliver the full customer journey entirely from within their own walls. Many core functions beyond brokering pass through suppliers, and each relationship is a potential entry point. The U.K.'s Financial Conduct Authority (FCA) reported that in 2025, more than 40% of the cyber incidents it received involved a third party.

Governance in this area often relies too heavily on the contract stage and the assumption that relationships remain static over time. A supplier assessment completed at onboarding does not protect data during an incident two years later. The insurer needs to understand, in practice rather than on paper, what each supplier can access, how that access is controlled in real time, and how quickly the relationship can be isolated if something goes wrong.



Insurers such as Allianz hold vast stores of personal, financial, and health data, placing the sector firmly among the highest-value targets for cybercriminals

The same discipline should apply to cloud platforms—a growing concern for distributed, connected insurers and the underlying cause of the Allianz data breach. A software-as-a-service (SaaS) application that holds sensitive policyholder data does not become lower risk simply because the infrastructure is hosted by a third party. The data is still the insurer's responsibility and must be treated with the same scrutiny as any other data store. Reliance on a third party means partial reliance on that party's security procedures.

40%

of cyber incidents reported to the U.K. FCA in 2025 involved a third party

RANSOMWARE AND THE QUIETER LOSSES

CNA Financial is reported to have made a USD 40 million payment following a 2021 ransomware attack[‡], which went on to become one of the most cited figures in insurance cyber risk. Tokio Marine Life Insurance Singapore confirmed that year that a number of its internal Windows servers had been hit, though it said core systems were unaffected. Fidelity National Financial's 2023 ransomware attack disrupted title insurance and related transaction services across the U.S.

While ransomware can act as a rather sharp weapon, the danger is that its dramatic components—encryption, downtime, money demands—mean it dominates the security conversation over quieter threats. Business email compromise (BEC) and payment fraud remain stubbornly effective because they move through trusted workflows rather than against them. Insurers may be left paying claims for policyholders hit by these techniques while facing exactly the same methods in their own finance and operations teams. Whether through a compromised claims handler, finance team, or broker account, the target is often the same—any trusted process that moves money, data, or authority.

THREATLOCKER® TIP

Learn how our Zero Trust platform helps build enterprise integrity in financial institutions

COSTS BEYOND CASH

The Medibank breach in Australia remains one of the clearest illustrations of what data exposure can cost in the insurance sector, beyond the immediate incident. The 2022 attack affected 9.7 million current and former customers, with personal information later released on the dark web[‡]. Australia's privacy regulator has alleged that Medibank failed to take reasonable steps to protect the personal information it held.

The lesson has relevance well beyond health insurance. Any insurer holding large volumes of personal data needs to think seriously about what data it retains and why. Old records, dormant accounts, and forgotten systems can create genuine regulatory and reputational exposure. Retention has nothing to do with housekeeping and everything to do with risk.

Aviva said it detected a record level of claims fraud in 2025, identifying more than 18,400 suspect claims across its brands worth GBP 233 million (USD 309 million), with AI being used to fabricate accident scenes, alter documents, and inflate damage. The tools available to a fraudster today are cheap, accessible, and improving. Claims integrity is, therefore, a cybersecurity concern like any other, even if no network is compromised.

Any insurer holding large volumes of personal data needs to think seriously about what data it retains and why

A convincingly altered document may not be a technical intrusion, but it attacks the insurer's ability to trust digital evidence. A cloned voice on a vishing call can compromise a decision, and a synthetic identity can move through onboarding if the controls depend too heavily on manipulable documentation. The industry's instinct may be to divide these problems by department—ransomware dealt with by IT, fraud by claims departments, and so on—but this defense approach does not match the method of attack. Attackers do not organize themselves politely, so insurers must take a holistic approach to security that covers the business as a whole.

The breach rarely begins where the security team is watching

NEXT STEPS FOR INSURANCE SECURITY

The key now is for insurers to look carefully at their operating model across the entire map of the business. Where is trust being granted, and who has the authority to change it? Place controls into policies, and ensure they hold under real working conditions. Recovery planning deserves the same realism. Give operational questions operational answers, focusing on what should happen in the event of a breach before moving on to the technical question of how.

The insurers best positioned to weather the storm are those that treat cyber risk as an integrated feature of the insurance operating model rather than a parallel concern managed by a separate team. They will know which data carries the greatest exposure, which processes are most accessible to external pressure, and which third-party relationships could cause the most damage if compromised. They will test the human routes into the business with the same seriousness they apply to the technical ones. And they will understand that trust, once extended, must be continually verified. Zero Trust controls help them to avoid outside threats and prevent internal attacks—in this way, they keep their promise of integrity to their policyholders.

NEXT PAGE

Find out about how ThreatLocker® helps customers in the insurance industry to secure customer data:

Dan Bourdeau, Director of IT and Cybersecurity at the Michigan Municipal Risk Management Authority (MMRMA), details the role of Zero Trust security in protecting the policies of its growing membership of public entities.

MICHIGAN'S **MUNICIPAL SAFETY NET**

Dan Bourdeau, Director of IT and Cybersecurity at the Michigan Municipal Risk Management Authority (MMRMA), explains how local government risk, cyber resilience, and practical Zero Trust meet inside the public sector

Often found poring over his iPad, late at night, Dan Bourdeau does not stop. What Bourdeau does outside of working hours says a great deal about how he sees cybersecurity. "It's not just work," he said. "It's a hobby and a passion, and it is what lights my creativity. I can't sing, I don't dance, I can't play an instrument—so this is where I get to be creative."

That creative outlet might seem unusual: Bourdeau is the cyber risk and technology leader at the Michigan Municipal Risk Management Authority (MMRMA). A governmental insurance pool with 463 members across Michigan, MMRMA allows public bodies to pool their money and self-insure, sharing risk rather than buying coverage in the ordinary commercial market.

But Bourdeau's role is not simply to sit behind a screen and protect one organization. As he explains it, he was brought in to "specifically tackle" cyber risk across the pool: to drive education, encourage knowledge sharing, support innovation, and surface tools that can make MMRMA's members more resilient in a threat landscape that continues to evolve.

SHARED RISK, SHARED RESPONSIBILITY

For Bourdeau, the most powerful aspect of MMRMA's shared model is its ability to turn one solution into many. "When we tackle risk as a pool, as a group, what works for one oftentimes works for others," he said. If a better training model can reduce excessive-force claims related to law enforcement, for example, the entire pool benefits. Fewer claims are paid, more money stays in the pool, and reinsurance partners can price the risk more favorably.

The authority's approach to cybersecurity now follows the same logic. Bourdeau's role was created four years ago after MMRMA's board recognized a clear change in the loss picture. Traditional claims still represented a major part of the organization's exposure, but cybersecurity had begun to move in a way that made insurance people pay attention.

"Cyber was a bit of a dark horse in public sector insurance," he said. "People obviously knew that the technology was out there, we're all using it, but cyber really began to pivot from private-sector kinds of attacks into public-sector targeting. When you go from zero to USD 100,000 in a blink, and you do that in an insurance company that is built on looking at loss trends and risks, it rings an alarm bell."

> "One slip-up is all it takes for an attacker to find their way in. We somehow have to get cybersecurity right all the time

FROM DIGITAL TO PHYSICAL

Local government is complex and rarely enjoys the luxury of cyber risks that stay neatly inside the IT domain. “When people can’t pay their water bills or their tax bills, or they can’t get building permits issued, all of that has a real physical consequence,” Bourdeau said. “Cyberthreats can stop commerce. They can impact privacy and confidentiality.”

MMRMA’s board saw this pressure coming. The pool works with reinsurance partners worldwide to ensure it can withstand catastrophic claims. In effect, the insurance company buys insurance, and the price of that backing depends on how well it can control risk. “One slip-up is all it takes for an attacker to find their way in,” explained Bourdeau. “We somehow have to get cybersecurity right all the time.”

Bourdeau came to MMRMA already familiar with local government pressure; he previously spent nine years as Chief Innovation Officer at the City of Westland, a member community of around 86,000 people. There, he had to protect residents’ data while explaining to elected officials why invisible technology deserved visible public money. “It’s much easier for a mayor to trot out a brand-new shiny fire truck, or a new dump truck, or a cute canine officer, and have residents understand that that’s where their tax dollars are going,” he said. “It’s much harder to say, we just bought a USD 50,000 firewall to protect your data.”

That experience helps Bourdeau speak to MMRMA’s full range of members. At one end are large cities with dedicated cyber leadership. At the other end are communities with almost no formal IT resources. “We have large cities like Grand Rapids, Michigan, that have a full-time CISO,” he said. “Then, we have a tiny little hamlet in the Upper Peninsula of Michigan; their grasscutter happens to be a technologist as a hobby, and that’s their IT guy.”

↓ Grand Rapids is a growing economic center in western Michigan, where thousands of businesses rely on resilient cybersecurity

→ Bourdeau’s strong security expertise and setup help communities stay resilient when it matters most

THE POWER OF WORKING TOGETHER

Bourdeau recognizes the pressure on cybersecurity teams, especially the sense that a missed signal can define a career. “There is a sense of isolation and maybe even a lack of hope sometimes,” he said. “I have to be right every day, every minute. The bad guy has to be right once, and it could ruin my career. It could cost me my job. That’s a difficult place to live.”

His advice is simple: collaborate. In local government, information sharing is often encouraged. In more competitive sectors, the same openness may be harder to achieve, but Bourdeau says nobody has to work in isolation. “Even something as simple as finding a podcast or some kind of online journal that is germane to your industry can be a source of inspiration,” he said. “Nobody is in this business alone.”

The point is to learn from the experience of others and keep some perspective. “There are a ton of resources for creative people to share what they’ve done, what they’ve discovered, and what they’ve realized you should never, ever do. Why repeat pains and sins?”

His own approach is to keep the subject useful without making it joyless. “Don’t just drone on with acronyms and dire warnings,” he said. “We get that en masse all the time. You’ve got to have some fun. You’ve got to realize that security is important, but there’s more to life.”

PROGRESSION THROUGH EDUCATION

When Bourdeau started, even multi-factor authentication (MFA) had to be explained, funded, and implemented. Some members lacked the license. Others lacked the skills. But within 18 months, most had transitioned from having no MFA to having it in place. “What was once kind of our ceiling, our goal, became our floor,” he said. “If you don’t have this, you’re behind.”

> “Without the visibility that ThreatLocker® provides, you have to rely on luck. You’re lucky to get through the day and not have a threat actor find you

The same pattern is now being applied to phishing-resistant MFA and the rollout of Zero Trust. The large communities may move first, but Bourdeau measures success by whether the whole pool can follow. “Not everybody can do Zero Trust the same way, nor should they. The times when I feel we’ve really accomplished something for the pool are when we’ve made it accessible to every member, big, small, complex, simple, you name it.”

PRACTICAL APPLICATIONS

MMRMA’s cyber work is not limited to advice or encouragement. Bourdeau also has access to a practical solution to help members act on what they learn. “Our board of directors runs an annual USD 2 million plus grant program,” he said. “I get the opportunity, the real privilege, to sit across from my colleagues and say, hey, I have a USD 2 million grant dollars bucket that can help you tap into some professional and consultative services and technology devices. Let me help.”

With a financial mechanism, MMRMA's education program gains weight, bridging the gap between awareness and implementation for communities that might not otherwise be able to reach their cyber goals. Bourdeau is careful, however, about the direction of support. His role is not to issue blanket endorsements. "While I can't actively endorse any particular product, I can at least share my professional experience," he said. Even inside the shared structure of the pool, each member has its own environment, constraints, and politics. There is no single answer.

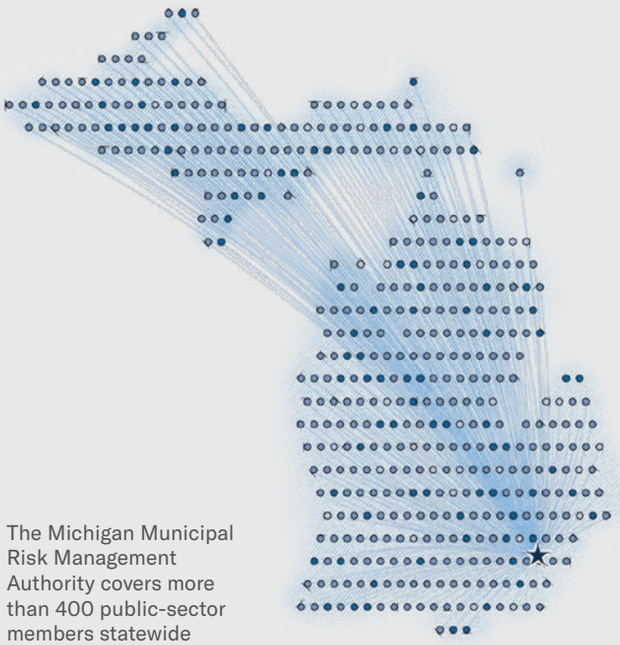
Even so, the similarities across local government create useful patterns. "We're all in the same kind of business," Bourdeau said. "We're all doing local government kinds of things. We all have very similar types of assets and infrastructure to protect." One tested approach can become a lesson for many others—and in many cases, Zero Trust is the perfect fit.

REACHING THE SOLUTION

ThreatLocker entered the picture through Bourdeau's headphones. "I'm a huge consumer of podcasts," he said. "That's where I get a lot of my daily cyber intelligence. In that sense, having heard about it on so many podcasts, ThreatLocker found me." The discovery led Bourdeau back to a Zero Trust idea he had previously found too difficult to apply. "Every time I looked at Zero Trust in the early days, when it was all just buzzy and flashy, I'd wonder if it was just the next trend. It was really complex, clunky, and hard to effectively implement."

> "I wish I had it in 2009, when protecting a community of **86,000** people required a lot of grit, a lot of effort, and a whole lot of luck

ThreatLocker made the practice feasible. Bourdeau brought it into MMRMA, tested it inside his own operation, and saw the difference immediately. "Finding ThreatLocker two years ago was a game changer," he said. "I wish I had it in 2009, when protecting a community of 86,000 people required a lot of grit, a lot of effort, and a whole lot of luck. Without the visibility that ThreatLocker provides, you have to rely on luck. You're lucky to get through the day and not have a threat actor find you."



The Michigan Municipal Risk Management Authority covers more than 400 public-sector members statewide from a single pool

THREATLOCKER TIP

Scan here to discover how the immediate visibility of **Unified Audit** streamlines **Allowlisting**

Iron Mountain's industrial legacy and spirit of innovation highlight the need for strong cybersecurity foundations

> "What was once kind of our ceiling, our goal, became our floor. If you don't have this, you're behind

"Luck no longer comes into the picture," explained Bourdeau. "You think ThreatLocker is meant to lock down and control everything, but in reality, what it's done for our operation has empowered me to give more access more confidently than I ever would have before."

The capability he comes back to most often is **Unified Audit**. "For our specific operation, Unified Audit is magic," he said. Earlier in his career, the same work meant combing through machine logs, removing noise, and trying to build enough context to act. Now, multiple sources of information are brought together to form a single view of activity and status. "It's 30 seconds of really good, focused information. If I can do that 100 times a day instead of just five or six times a day because it takes me 30 minutes, it's a force multiplier."

BUILDING THE FUTURE

Bourdeau does not spend much time generating reports because he sees those as records of the past. "Reports are reactive. ThreatLocker allows me to live in the moment. It finds me in my inbox and says, 'I think you should look at this.'" He is careful, though, not to present any platform as a substitute for the basics. Patching, vulnerability management, and cyber hygiene are still vital. "It doesn't take away the need to do the work," he said. "It visualizes it, and it streamlines it."

In the end, the work is primarily focused on education. Bourdeau speaks, hosts a podcast, runs workshops, and uses any available platform to make cyber risk easier to understand. "Digital isn't just digital; there are real-world consequences and ramifications," he said. That mission is what keeps him engaged after hours, in the chair with the iPad. "If I can help every city in Michigan find the way to Zero Trust, from major metropolitan areas all the way to smaller places like Iron Mountain, I feel like I've made a difference," he said. "That's the mission that drives me. Cybersecurity is a passion; it's so much more than a career."

IN THIS ISSUE

Dive deeper into cybersecurity in the insurance industry and discover why trust is the sector's greatest asset on page 14

HOW TO: STOP REMOTE ENCRYPTION

Ransomware does not need to compromise the file server to encrypt files; it just needs to find a route. Kieran Human, Lead Cybersecurity Engineer at ThreatLocker, explains how to create the controls that protect files in shared space

Although ransomware is often characterized as an attack on a machine where files are stored, it can run on one device while rewriting files somewhere else. “Remote encryption is not always a file server problem,” explained Human. “It’s often a workstation issue that becomes a file server problem. If a workstation can reach a share and write to the files, it follows that it can damage them, too.”

A file server may itself be secured, but that does not help enough if a compromised workstation has been granted write access to its folders. All an attacker needs is a route to the data. “When I’m looking at this kind of thing, the question I always ask is: Why did that device have access to those files in the first place?” Human said. “Ransomware does not need unlimited power if the environment gives it unlimited access.”

THE PROBLEM WITH ALWAYS-ON ACCESS

From a ransomware point of view, ordinary write access can be enough to cause a problem. Human says storage access has to be reviewed with the same discipline as application control or privileged access. Some users only need read access. Some devices should never connect to certain shares. Some support work needs only a short window of access before rights are revoked. “Read access and write access should not be treated as the same thing,” said Human. “Ransomware needs the ability to change files. Reducing unnecessary write access can make a major difference.”

THREE STEPS TO REDUCING REMOTE ENCRYPTION RISK

1. Find the shares that matter

Identify the file shares that would cause serious disruption if encrypted. For each one, ask who needs access, what type of access they need, which devices should connect, when, and for how long. Allow this to drive secure policy decisions.

2. Restrict the route

Use **Zero Trust Endpoint Firewall** to control device-to-resource access. Avoid broad internal connectivity by default. Permit approved devices to communicate with protected file servers, then narrow that access by purpose, role, and time window.

3. Limit what access can do

Separate read and write permissions. Remove standing access where timed access would work. Use **Allowlisting** to block unapproved software and **Ringfencing™** to limit what trusted applications can do if they are misused or otherwise compromised.

The ThreatLocker **Zero Trust Endpoint Firewall** capability is, bluntly, designed to control which devices can communicate with which resources. Applied to remote encryption, the principle is equally blunt. Being on the network should not automatically give a device a route to the file server. It should be allowed to reach the resources it needs, not every resource on a network.

"Under the old model, once someone is inside, they get insider access," Human said. "With ThreatLocker, the network constantly looks for proof that a device should be talking to this resource, for this reason, right now. If there's no proof, or if the justification is lost, there's no connection. Moving the firewall decision closer to the endpoint is key. It's not enough to protect the edge of the network if an attack is moving between devices and shares internally."

ACCESS WITH AN EXPIRY DATE

"It's really a question of policy management," explained Human. "Zero Trust Endpoint Firewall gives you a lot of options, and you just need to use them sensibly. If someone needs access in working hours, for example, give them that access, and nothing more. Don't give them permanent unrestricted access just because that was easier on the day."

> "Ransomware does not need unlimited power if the environment gives it unlimited access

"A lot of ransomware damage happens because access is always on," he continued. "If a share is reachable all day, every day, from devices that do not constantly need it, that's an unnecessary risk."

OTHER LAYERS STILL COUNT

Human is clear that restricting access to shares is not the whole answer. "You still want **Allowlisting** and enforceable application control. You still want other access restrictions enforced through **Ringfencing.** But when the specific problem is remote encryption, the first question is always whether the compromised device should have been able to reach the share in the first place."

For Human, the useful test is whether every route to a share can be defended in plain English. If the answer is vague, the rule probably needs work. "I don't think this has to be complicated," he said. "Begin by picking the shares that would really hurt if they were encrypted. Then ask who needs to write to them, from which devices, at what times, and why? If you can't answer those questions clearly, you've probably found access that needs to be removed, narrowed, or time-limited."

"You'll spend forever trying to make every endpoint perfect," concluded Human. "So, you have to make sure one bad endpoint can't become a bad day for the whole organization. That means controlling the route, the permissions, and closing access when it's no longer needed."

HOW TO:
PATCH PORTABLE APPLICATIONS

Portable applications can sit outside the places some patching tools expect to look. Rob Allen, Chief Product Officer at ThreatLocker®, explains how technical teams can find applications, patch them, and prove the risk has been reduced

Patch management often hinges on the assumption that software has been installed in a predictable way. An application goes through an installer, appears in inventory, registers itself properly, and can then be checked against a known update path. Really, administrators may not even be aware that they have made that assumption, at least until the patching process reaches portable applications and gets stuck.

Portable applications do not follow the rules. They may be copied into or launched from a user folder or USB drive, kept by a user who needs a quick tool to get a job done. There may be no conventional installation record, but the code is still there—and if that code is vulnerable and left that way, the endpoint carries the risk.

"The risk is the application, not the way it's been installed," said Allen. "If vulnerable code is sitting on the endpoint, it needs to be found. If you're presuming that an application is safe just because it's gone through a formal installation process, you've granted implicit trust to both the application and installer, which makes no sense. Ignoring portable applications doesn't work either. You have to know what's in your estate, however it happens to be there, and make it safe."

Patch Management is designed to address that problem by identifying applications using signatures, rather than relying solely on installation records. "We're looking for the application itself," Allen said. "It doesn't matter if it's installed, portable, or sitting somewhere unusual; the patching decision is based on the software we find."

PATCH WITHOUT LOSING CONTROL

Security teams want exposure reduced quickly, but patching still has to align with the flow and philosophy of the business. Some organizations need to work within maintenance windows or to stick to a defined release level to ensure continuity. Some need time to test updates before they are pushed widely. Allen suggests that patching can go wrong if it is treated as a race rather than a controlled process. "The point of patching is to reduce risk, not create disruption," he said. "You still need to decide how much change the business can absorb at once."

Portable applications may not have gone through standard IT approval, but once they are present, they still need to be patched from a trusted source. Updating software means allowing new code onto an endpoint, which makes the source of the patch as important as the decision to patch in the first place.

ThreatLocker capabilities currently cover around

15,000

applications

ThreatLocker capabilities currently cover around 15,000 applications, but the scale is not the only point. Those patches are managed and curated by a human team, rather than blindly scraped from the internet and pushed directly into customer environments. “You don’t want a patching system grabbing files from wherever it can find them,” Allen said. “That’s a supply chain risk.”

“If you’re asking an endpoint to accept new code, you need to know where that code came from and who has checked it,” continued Allen. “There’s a reason we have people in the loop. Never trust, always verify. An endpoint is going to run that code. We don’t blindly trust a patch—we verify it comes from a known-good source and put a controlled process behind it.”

PROVE WHAT CHANGED

“Patching isn’t finished when the update runs,” Allen said. “You also need to show what happened. What is patched, what is still unpatched, and where does the remaining exposure sit?” Reporting is a vital part of the verification process, and **DAC (Defense Against Configurations)** can help teams build reports that paint a picture of the state of the estate. For internal security teams, it supports prioritization. For leadership, clients, or auditors, it helps show that patching is being managed as a measurable process rather than a best-efforts task.

“Patch Management closes known gaps,” Allen said. “DAC (Defense Against Configurations) helps prove where those gaps still exist. And then **EDR Real-Time Threat-Detection** is there for the moment when an endpoint starts behaving like something is wrong. If you can detect everything, portable applications stop being a blind spot and become part of the same managed endpoint policy as everything else.”

> “
>
> The risk is the application, not the way it’s been installed

THREE STEPS TO PATCHING PORTABLE APPLICATIONS

1. Look beyond installed software
Do not rely only on the standard installed applications list. Portable executables, copied tools, and applications in unusual locations can still carry vulnerabilities. Build the patching view around what is present on the endpoint.

2. Patch from a trusted catalog
Use **Patch Management** to identify vulnerable applications by signature, whether they are installed normally or running as portable applications. Apply approved patches from a controlled catalog, on a schedule and at a release level the business can support.

3. Prove what is patched
Use **DAC (Defense Against Configurations)** to report which endpoints are patched, which still need attention, and where exceptions remain. Use **EDR Real-Time Threat-Detection** as the active monitoring layer for suspicious behavior while known exposure is being reduced.

DEFENDER, ATTACKER

A leaked zero-day exploit reveals the fragile nature of cybersecurity research and the need for Zero Trust in a world where trusted tools can go rogue

What happens when the trust between a cyber researcher and a vendor breaks down completely? The Microsoft Security Response Center (MSRC) recently found out. On March 26, 2026, a disgruntled researcher using the alias “Nightmare Eclipse” launched a public blog with a post attacking Microsoft for reneging on an agreement that “left me homeless with nothing.” Despite warning the company of what would happen, the researcher wrote, “they still stabbed me in the back anyways [sic].”

One week later, on April 2, 2026, a new post titled “Public Disclosure” was published, linking to a (now-deleted) GitHub account that revealed public details of what Nightmare Eclipse dubbed the “BlueHammer” vulnerability. This was a zero-day exploit that affected Windows’ own built-in security tool, Microsoft Defender. The vulnerability had not previously been disclosed to Microsoft, so there was no accompanying common vulnerabilities and exposures (CVE) disclosure or patch.

The exploit enabled anyone with access to any Windows PC running Defender to gain elevated privileges on a limited-access account. This was achieved by chaining three individually legitimate processes via Defender during certain update and remediation workflows, in which a temporary Volume Shadow Copy snapshot was created. By pausing Defender at the right moment, an attacker would be left with access to key Registry hives, which would give them the information required to both hijack a local administrator account and spawn a SYSTEM-level shell—all without being detected.

Despite clearly revealing the steps to perform the attack, no initial patch or CVE was immediately forthcoming. It was not until 12 days later, once

external researchers confirmed the attack path used, that CVE-2026-33825 was published, accompanied by a patch for Microsoft Defender. Within 24 hours, Nightmare Eclipse had published details of a second Defender exploit, named "RedSun." That exploit took even longer to recognize; it was not until May 19, 2026, that a CVE and patch were deployed.

THE DANGERS OF LOTL ATTACKS

BlueHammer is a textbook example of a living-off-the-land (LOTL) attack. No actual malware or phishing is involved or required. Instead, local privilege escalation via native tools can be used to penetrate systems by chaining low-privileged accounts to high-privileged ones.

This type of targeting is nothing new. Trusted and signed native Windows functionality has been exploited for years, but this particular exploit gained notoriety for two reasons. First, it exposed the fragility of the working relationship between vendors and those they employ to hunt down vulnerabilities in their products. And second, there is a certain irony in that the exploit took advantage of weaknesses in Microsoft Defender, which is built into all versions of Windows and, ostensibly, a tool designed to protect.

It is worth noting that while vulnerabilities like BlueHammer require access to an affected computer, that access need not be physical. That means organizations are not simply at risk from on-site employees and contractors, but from anyone with any kind of foothold—including remote ones—on a vulnerable system.

Patching remains vital, but it is clearly not enough. Microsoft may have been slow to respond to Nightmare Eclipse's actions, but it eventually produced both CVEs and patches for Microsoft Defender that close the vulnerabilities. Keeping systems patched, therefore, remains a vital part of an organization's security policy, and ThreatLocker tools like **Patch Management** can ensure patching stays up to date.

However, the reactive nature of patching makes it a bit like closing the barn door after the horse has bolted. In the time it takes to close the door, the system has been vulnerable for days or even weeks. LOTL attacks are particularly dangerous because they employ trusted tools to perform what, on the surface, appear to be perfectly legitimate actions. So even security tools that rely on heuristic techniques may find it difficult—or even impossible—to detect malicious behavior. What is more, once an attacker gains access, they can bypass normal account-based restrictions—both BlueHammer and RedSun gave attackers enough access to potentially reach everything from file shares, admin tools, and credentials to backups, cloud drives, and sensitive datasets.

When examining Zero Trust security tools, look for solutions that employ a multi-layered approach

THE MOST EFFECTIVE PROTECTION REMAINS ZERO TRUST

The best way to protect an organization from zero-day exploits—or at least limit the damage an incursion can cause—is to deploy security policies which follow the Zero Trust philosophy. These employ a deny-by-default model that assumes nothing is safe, even native, signed, and patched components. As a result, only expressly permitted actions are allowed, helping limit the damage an attacker can do even if they get past an organization's first line of defense.

When examining Zero Trust security tools, look for solutions that employ a multi-layered approach. For example, **Ringfencing™** sets strict limits on what processes and applications can and cannot do or access, limiting an attacker's ability to weaponize even compromised components. **Privileged Access Management** removes the broad access privileges granted to administrator-level accounts in favor of a tightly controlled least-privilege approach that confers only the minimum access required at any time. Used in tandem, these tools ensure that any attacker who gains even administrator-level access to a system will still be curtailed by its Zero Trust policies, limiting the damage they might otherwise cause.

As for Nightmare Eclipse, the disgruntled researcher continues to drip-feed discovered Windows vulnerabilities into the public domain—on May 29, 2026, they claimed that other researchers were now handing them "free vulnerabilities," the first of which is said to bypass BitLocker. It seems this story is set to run and run.

BEHIND THE GLASS

At Georgia Aquarium, cybersecurity is part of a wider duty of care: protecting the people, systems, and resources that support animal care, education, conservation, and research

Visitors to Georgia Aquarium come for the animals and the experience, both shaped by the careful choreography of a major public attraction. Yuta Kitabayashi sees all of that through the lens of endpoints, permissions, and security decisions that help keep the organization running.

Kitabayashi is the Senior Manager of IT and Systems Engineer at Georgia Aquarium, though he describes the role in broader terms. He sees himself as an "IT orchestrator": part engineer, part systems administrator, part project manager, and part technical lead. That mix is useful in an environment that is often misunderstood from the outside.

"When I first started, I just thought about the animals, the habitats, and the shows," he said. "But over time, I obviously started seeing more of the people behind it. Georgia Aquarium is unique, but it is also a typical business. It requires a lot of common business and IT operations as well."

A NONPROFIT WITH A WIDER MISSION

Kitabayashi is keen to stress that the aquarium operates as an independent nonprofit. "We are not operated by the government as a public facility nor backed by large corporations," he explained. "Our money goes to research and education, which means we have to be good stewards of our finances and our resource allocation."

Kitabayashi has found common ground with scientists across the organization, as both worlds depend on records, logs, analysis, and action. "I love working with the scientists," he said. "They have so many cool stories, and IT is part of computer science, so we actually speak the same language. Also, a good proportion of our guests are field trip or summer camp groups, not just there to enjoy the habitats, so I work with a whole education group on site that teaches our local children."

> "Zero Trust really means not doubting, but making everything vouch for itself

The aquarium's wider purpose has changed the way he thinks about the job. "There's something contagious about being in this environment and working with passionate people. This work feels different," said Kitabayashi. "Every day I have more appreciation for the education team, the conservation team, and the scientists, as well as their passion for studying and working with animals."

BUILDING INTERNAL CONTROL

Before joining Georgia Aquarium, Kitabayashi worked in the managed service provider (MSP) world. "Generally at an MSP, as much as you want to guide clients toward long-term structure and comprehensive cybersecurity, the decision is often not yours," he said. "Understandably many clients limit your involvement to break-fix type work, rather than holistic consultation." Kitabayashi is careful, though, not to present a dramatic before-and-after story. The aquarium had a functioning environment, but his concern was what might happen if that environment came under pressure.

ThreatLocker® came into the picture through a familiar route: people Kitabayashi already trusted. One former colleague from his MSP days had moved to Georgia Aquarium before him and helped point him toward the vacant role. Another had moved from the same MSP to ThreatLocker. Kitabayashi explains that Georgia Aquarium's adoption of ThreatLocker was natural given the IT estate he inherited.

"It's not just because I knew someone that worked at ThreatLocker," he said. "I knew what ThreatLocker did, and we needed it. As soon as I got here, I felt like there was a lot of opportunity for lateral movement on our network. I also realized that local administration rights needed improvement at the end-user level. And we needed **Allowlisting,** because a lot of people were just installing whatever they wanted."

EVERY EXECUTION IS A DECISION

Kitabayashi talks about Zero Trust without grand language. For him, it is a working discipline. “Zero Trust really means not doubting, but making everything vouch for itself,” he said. “Every transaction that happens at every layer, from user to devices to each mailbox to each email.”

He has strengthened multi-factor authentication (MFA), used Group Policy to reduce local administrative privileges, introduced tiered administrator access in Active Directory (AD), leveraged Microsoft 365 and Defender capabilities, and worked on device management and network segmentation. ThreatLocker plays its part in that broader effort by controlling what runs on endpoints.

“People focus on devices and endpoint detection and response (EDR) and user logins,” he said. “But some people forget about the application as a resource. Every execution within the computer is a resource. A lot of the time, an application that a scientist wants to run is something niche, written by another scientist, with no certificates and no backing of a major software company.”

> “Every day means discovery, and every day means reinforcement

That left Kitabayashi asking the questions any security-minded IT lead would ask: “Where did this come from? Where did you download this? And often the answer is ‘oh, it’s a research tool.’ And who knows who wrote it?” A blanket refusal would be a tidy solution from the IT side, but that does not support the work. “For our research efforts, finding a way to run this software safely is very important,” he said. “There’s a balance to be struck there, and capabilities like Allowlisting and **Ringfencing™** allow me to do that.”

“DOLPHINS ARE NOT ON WI-FI”

Cybersecurity at an aquarium invites one obvious question, given the infamous 2017 hack of a Las Vegas casino through a network-connected temperature sensor installed in one of its fish tanks. Kitabayashi is, however, quick to separate the animal systems from ordinary business IT.

“We try to air-gap everything that’s related to life-support systems,” he said. “That equipment does not even touch the internet at all. We have it segmented off in a very strategic way.” Then, he put it more simply: “Dolphins are not on Wi-Fi.”

The systems around animal care are treated as a separate environment, but Kitabayashi’s day-to-day work still leaves him plenty to protect. Georgia Aquarium has roughly 700 endpoints with a wide spread of users. “Cybersecurity, Zero Trust enforcement, it’s never complete,” he said. “It just keeps on continuing. Every day means discovery, and every day means reinforcement.”

Just as the exhibits are contained away from the reach of digital intruders, application control gives him internal containment. If a user accidentally downloads something malicious, or if ransomware tries to distribute itself, he has more confidence that it cannot simply run and spread unchecked. “Knowing that there’s a containment policy there is really nice,” he said. “It does help me sleep better at night.”

> Good cybersecurity understands the everyday operation of people and complements them

Yuta Kitabayashi
Senior Manager of IT and Systems Engineer at Georgia Aquarium

A graduate of Southern Polytechnic State University, where he earned a Bachelor of Science, Kitabayashi came to the aquarium after building his career in IT engineering and managed services. That background has led him to be more interested in what will work than what sounds impressive. At Georgia Aquarium, his remit is wider than servers and tickets. Kitabayashi supports an organization where technology enables specialist scientific and educational work across a complex public venue.

SECURITY THAT UNDERSTANDS THE WORK

For Kitabayashi, good cybersecurity starts with understanding how people actually operate, how the business works, where people need freedom, and how security can strengthen the environment without impeding the aquarium's purpose. "I personally feel that whoever's leading cybersecurity needs to understand the business well," he said. "We are sometimes very focused on the tools and solutions."

That means seeing the reality of Georgia Aquarium's various systems and protecting them in ways that serve the mission rather than competing with it. "I want to know how people work first," he said. "Then we implement, find the weaknesses and strengths, and implement again. Good cybersecurity understands the everyday operation of people and complements them."

The result is a practical, careful version of security. It is not cyber theater, and it is not a simple story about protecting habitats. It is the daily work of reducing risk in a place where technology supports animal care, education, research, and public engagement.

LIE, ROBOT

If a robot becomes compromised, it can no longer be trusted to tell the truth about what it is doing. But in many cases, operators will not even know there is a problem until it is too late

Robots are the building blocks of modern industry. Once isolated entities, caged on factory floors, and divorced from the messy reality of enterprise IT, they are now an inextricable part of it. Robotics has become part of the expanding operational technology (OT) landscape. They are routinely connected to production systems, remote support tools, and increasingly intelligent workflows. There is no longer any absolute separation. Robots and IT are one.

Robotics is escaping its niche and becoming standard industrial infrastructure. According to the International Federation of Robotics' World Robotics 2025 report, 542,000 industrial robots were installed globally in 2024, more than double the number installed 10 years earlier. Annual installations have now topped 500,000 units for four consecutive years, with Asia accounting for 74% of new deployments, Europe 16%, and the Americas 9%.

Such growth has brought the cybersecurity of robotics into focus. The concern is not necessarily robots going Hollywood haywire; since the earliest shop floor accidents, the institution of physical safety controls has been of utmost importance in the field. The burning question is whether those controls can be trusted and whether robots will reliably do as instructed in a connected environment. A robot may move with perfect mechanical precision, but if its programming has been compromised, that precision is compromised with it.

THE ROBOT IN THE NETWORK

A recent vulnerability in Universal Robots' PolyScope 5 software shows that this is no idle concern. CVE-2026-8153, disclosed in May 2026, affected the Dashboard Server interface used by the company's collaborative robots. In a statement, Universal Robots said the vulnerability allowed "an unauthenticated attacker who can reach the Dashboard Server network port to craft commands that are executed on the robot's operating system." Essentially, access to the dashboard—through a local network or, if somehow configured to accept it, through the internet directly—would be enough for an intruder to take control of the associated robot.

The flaw received a score of 10 from the Common Vulnerability Scoring System (CVSS) 3.1, placing it in the critical category, and was quickly patched, but it makes clear that a collaborative robot is a networked endpoint. Like any other endpoint, it runs an operating system, and on that runs software which may be vulnerable. It must keep maintenance pathways open, and assumptions must be made about the validity and security of its environment. Universal Robots' own guidance following the discovery of the vulnerability made the lesson plain: The security of the network is essential to the security of the robot.

Industrial robots sit in environments where IT and OT are increasingly interdependent. When OT and IT meet, every piece of software and hardware that boosts productivity also becomes part of the attack surface. Robots are no different.

PRECISION, FAILING QUIETLY

Trend Micro and Politecnico di Milano's Rogue Robots research, published in 2017, was a landmark contribution to the understanding of cybersecurity in robotics. It demonstrated that attacks against robotic systems could alter operational parameters, introduce defects, disrupt production, and undermine safety assumptions.

It is the possibility of subtle, undetected change that makes robotics risk especially uncomfortable. A cyberattack does not have to make a machine behave wildly to damage a business. In high-volume manufacturing, a small change in a robot's movement, calibration, timing, or tooling path can create

defective products at speed. In food, pharmaceuticals, automotive, or electronics, those defects may not be immediately visible, but they tend to become clear once quality control, warranty claims, or safety checks expose the problem—or when human life is affected down the line.

This is a world of indirect threat, where compromised systems elsewhere on the network may open the door to digital industrial espionage, and where that espionage may trickle down to an unsuspecting—perhaps unintended—victim.

542,000

industrial robots were installed globally in 2024, more than double the number installed 10 years earlier

THE WIDER OT LESSON

Robotics' connectivity means it inherits the wider cybersecurity pressures affecting industrial operations. The FBI's Internet Crime Complaint Center (IC3) recorded more than 2,100 ransomware incidents directed at U.S. critical infrastructure in 2025, with critical manufacturing among the three most heavily targeted sectors. The FBI attributed this concentration to threat actors deliberately targeting organizations with low tolerance for operational downtime; those manufacturers operating within just-in-time supply chains face enormous pressure to pay quickly and restore operations.

Recent manufacturing incidents tell a clear story of the impact of such attacks. In May 2025, steel manufacturer Nucor halted certain production operations after identifying unauthorized third-party access to some of its IT systems. The company took affected systems offline, worked with external cybersecurity experts, and moved through containment, remediation, and recovery—the manufacturing endpoints remained unaffected, but system intrusion eroded confidence in them.

Toyota's 2022 production shutdown in Japan was not even the result of a problem at Toyota. Supplier Kojima Industries was hit by a massive cyberattack, bringing production at each of Toyota's 14 Japanese factories to a halt. The attack did not hit the carmaker, but the loss of confidence in a key dependency was enough to cause it to hit the emergency stop.

EXPOSURE IS STILL VISIBLE

Some of the direct risks are also more basic than the language of advanced robotics suggests. A 2025 analysis of publicly accessible OT found nearly 70,000 exposed OT devices globally[‡], including systems using protocols such as Modbus TCP, EtherNet/IP, and S7. The researchers also identified exposed information, outdated firmware with known critical vulnerabilities, and accessible graphical interfaces for human-machine interface (HMI) and supervisory control and data acquisition (SCADA) systems.

Various studies count exposure differently, but the problem is clear enough: Connections to industrial systems continue to appear in places they should not. Not every exposed device is a robot controller, and not every reachable service is immediately exploitable, but the assumptions that once protected OT are much weaker in a connected world. Isolation, obscurity, and specialist knowledge are no longer enough—and just about every IT professor will gladly educate first-day students about the dangers of making assumptions.

MEDICAL ROBOTICS RAISES THE STAKES

As robotics moves closer to humans—in the field of medical robotics, for instance—these devices raise the stakes, losing the air of separation that comes from industrial assets placed in a different room. They must comply with specific robotics regulations such as ANSI/A3 R15.06-2025, but also align with those regarding patient safety, data protection, and beyond. They must be sufficiently protected to foster assurance and trust between clinicians and technology.

Research into teleoperated surgical robots has already shown why cybersecurity must be part of that conversation. University of Washington researchers examined the Raven II research surgical robot and demonstrated attacks including command manipulation, denial-of-service (DoS), and abuse of emergency-stop behavior. The work was conducted on a research platform, not a deployed clinical product, but it was nonetheless significant: If robotics depends on command integrity, timing, telemetry, and operator control, then cyber risk is indisputably part of the safety and reliability picture.

Medical robots are not inherently suspicious; quite the opposite, in fact. In microsurgery, movement occurs at a scale far beyond ordinary manual dexterity, which is exactly where robotics can create new possibilities. But the more valuable and precise the system, the more important it becomes to protect the environment around it. It is part of product trust, clinical confidence, and operational maturity.

The more valuable and precise the system, the more important it becomes to protect the environment around it

The way to build trust in such a disparate environment, rife with legacy hardware, unpatchable systems, and mixed protocols, is through a strict Zero Trust stance

Securing OT is not only a technical task—the people who operate, maintain, and access robot systems are central to keeping them safe

THE SAFETY CONVERSATION IS CHANGING

As the boundary between safety and cybersecurity narrows, robot controller advisories show the criticality of that link. Japanese automation expert FANUC has previously disclosed vulnerabilities in robot controllers that, if exploited, could cause system software to stop working correctly, requiring a restart. Not as dramatic as a remote takeover, sure, but in a production environment, a controller crash is far from trivial. Issues like this and the PolyScope 5 bug prove the importance of instituting a regular patching regime alongside more direct cybersecurity measures.

That said, OT environments cannot always be treated like conventional IT. A laptop can usually be safely rebooted during a maintenance window, but a robot cell may be part of a validated production process, a continuous manufacturing schedule, or a clinical environment. Changes—and patches—must be tested, documented, and coordinated, a less-than-simple process that makes prevention, containment, least privilege, and controlled execution more important.

THREATLOCKER® TIP

Learn more about protecting OT environments with enterprise-grade Zero Trust from ThreatLocker

TOWARD CONTROLLED AUTONOMY

As robots and their software become more adaptive and connected, the need for robust cybersecurity practices and solutions will only increase. Software is now an integral part of the robotics equation, powering functions like computer vision, AI-enabled inspection, autonomous mobility, digital twins, remote operation, fleet management, and machine-to-machine decision-making. This is an increasingly complex—and, in turn, vulnerable—sector.

That extra difficulty and potential vulnerability are no reason to slow innovation, but rather proof that security has to be built into the operating model by default. Cybersecurity is essentially about governing behavior, deciding what can run and communicate, and how. The future of robotics depends on protecting the trust that comes from getting this right; counterintuitively, the way to build trust in such a disparate environment, rife with legacy hardware, unpatchable systems, and mixed protocols, is through a strict Zero Trust stance.

Robots are not becoming dangerous just because they are robots, but because attackers know they are high-consequence endpoints inside complex, connected environments. The industry's focus now must be on building confidence—both inside and outside of the organization—in the systems that protect the machines at the end of the line.

THE PATIENT MACHINE

Inside Tokyo, the city where craft, patience, and philosophy shaped a different kind of tech culture

Nestled in Tokyo's historic Nihonbashi district, the DAWN Avatar Robot Café is a permanent café where robotic servers—OriHime and OriHime-D units—are teleoperated by people with amyotrophic lateral sclerosis (ALS), spinal muscular atrophy (SMA), or conditions preventing them from leaving home. Operators control their robots remotely and interact with customers via the robotic server. One operator works from 435 miles away.

The idea was developed by inventor Kentaro Yoshifuji, who refers to this technology as a form of teleportation, granting disabled people greater access to the workplace while also enabling older individuals to stay active and address loneliness.

For Sara Filipčić, Founder and Institute Director of Humane AI & Robotics, the café reframes what robots are for—and reflects Tokyo's distinctive approach to technology. Her institute focuses on human-centered systems for families, education, caregiving, and cross-cultural research.

← Tokyo, the city where robotics and human-machine coexistence have grown from a deep industrial and philosophical tradition

Japan produces

38%

of the world's industrial robots—more than any other country on earth[‡]

"Tokyo is genuinely the best city in the world to witness human-machine coexistence as a lived daily reality rather than a tech demo," said Filipčić, who moved to Japan in 2025. Rather than being embedded in an industrial or military context, teleoperation here "restores dignity, income, and social connection to bedridden people. That's the philosophy in action."

A PHILOSOPHY OF MEANING OVER PROFIT

This concept addresses the major economic challenges—labor shortages, an aging population, and a declining birth rate—that directly affect Japan's workers and carers. In 2021, the café won Japan's Good Design Grand Award. In response, the government launched the Moonshot R&D Program to pursue meaningful human-robot coexistence, with a 2030 target of ensuring that over 90% of people feel comfortable with AI robots in daily life.

Tokyo is a city where precision manufacturing, robotics, and human-machine coexistence have evolved from a deep industrial and philosophical tradition. "There's a spiritual orientation in Shinto and Japanese Buddhism that sees life—or something like life, something deserving of care and respect—in everything," said Filipčić. "When Japanese engineers build a robot, they are not just solving a functional problem. They are, in some sense, creating an entity that will live among people—and that carries moral weight from the beginning."

This is reflected in the philosophy of *monozukuri*, meaning "the making of things," which unites technical prowess with a spirit of sincerity, craft, and dedication to perfection. "You can trace a direct line from the Edo-era *karakuri* mechanical dolls, intricate clockwork figures built by master artisans, to Honda's ASIMO to Groove X's LOVOT. The craft sensibility never left; it just migrated from wood to silicon," said Filipčić.

Building on that is *kaizen*, meaning the commitment to continuous, incremental improvement at every level of an organization, which translates into a culture of never declaring a product "finished." Instead, a product is viewed through a lens of iteration and refinement, based on how humans experience the technology.

"When you combine these—the moral obligation to the user embedded in animism, the craft ethic of *monozukuri*, and the continuous improvement discipline of *kaizen*—you get something the West doesn't have a clean word for: technology designed as a long-term relationship, not a transaction," said Filipčić.

By 2030, more than

90%

of people should feel comfortable living and working alongside AI robots

> Tokyo is genuinely the best city in the world to witness human-machine coexistence as a lived daily reality rather than a tech demo

That framing resonates beyond the private sector. Professor Takahiro Ueyama, Chief Executive Member of the Council for Science, Technology, and Innovation in the Cabinet Office, sees it as a national orientation: "We raise the slogan that nobody should be left behind, whether that's the aging population or people with disabilities, and people expect that the development of science and technology can improve these kinds of people's welfare." It is a philosophy that shapes what gets built and who it is built for—a distinction that is nowhere more visible than in how Japanese robotics companies bring products to market.

↑ Japan's tradition of craftsmanship lives on in modern robotics, where precision engineering meets the pursuit of creating more intuitive, human-centered interactions

← The Nextage robots, which act as telebaristas at the DAWN Avatar Robot Café, offer disabled operators a new lease on life

TRULY HUMAN-LED TECHNOLOGY

What is unique in these instances is that Japan's robotics ecosystem is guided by the user, not by a market segment to be captured, Filipčić said. "You're building something that will live in someone's home, that their child or parent will interact with daily, and that responsibility is felt throughout the process." Deployment follows strict safety procedures that go beyond technical assurance to address psychosocial safety, demonstrating that the technology is beneficial to its users.

"Most of Japan's leading robotics companies emerged from research labs, which means by the time something reaches consumers, it has often been through years of academic publication and peer review," said Filipčić. "Groove X, which makes the LOVOT companion robot, is a perfect example." Before bringing LOVOT to market, Groove X published research demonstrating measurable benefits for well-being, including effects on dementia prevention. The robot is intentionally expensive and has been designed to last, with longevity in mind rather than obsolescence.

Shinjuku has been home to numerous robot restaurants, and currently hosts tourist hotspot Samurai Restaurant Time

"That's an entirely different philosophy than the iterative-launch, move-fast model that dominates Silicon Valley," said Filipčić. "I think of it as 'compliance as design'—where safety, care, and long-term human well-being aren't constraints layered on top of a product, but the actual design objectives from the start. The West often treats ethics and regulation as friction. Japan treats them as the engineering brief. This unique philosophy is most visible in the ordinary moments of usage, such as the care built into vending machines designed for the elderly or hospital service robots that move through corridors with genuine attentiveness—the technology isn't performing for you. It's serving you, quietly, the way a skilled artisan would. That's *monozukuri* in motion."

Japan has several institutions and centers where this work is focused, such as Odaiba and Miraikan—the National Museum of Emerging Science and Innovation—which rigorously examines what the relationship between humans and machines truly means.

LOOKING WEST

The closing human-robot divide is not a philosophy frozen in the past. In May 2026, the Robotics Innovation Center at the University of Tokyo's Institute of Science established a laboratory at its Yushima campus where robots conduct medical experiments previously performed by human researchers—only this time they are performed autonomously, without human operators. The craft tradition and the cutting edge are not in tension here. They are the same project.

"The fundamental lesson is that the relationship between a society and its technology reflects the values of that society—and if you haven't been intentional about those values, you've outsourced that decision to whoever gets to market first," said Filipčić. "Japan made a choice, culturally and institutionally, to treat technology as something you are in relationship with, not something you deploy at people."

Furthermore, Japan's response to its labor crisis was more considered than just automating roles; rather, it called upon the power of technology to increase inclusivity in the potential workforce, a framing that produces very different, more resilient systems.

There are direct implications here for security professionals. The instinct in Western technology culture is to treat compliance, ethics, and long-term safety as constraints—things that slow the build, where Japan's model inverts that logic.

> "
>
> Japan made a choice, culturally and institutionally, to treat technology as something you are in relationship with, not something you deploy at people

"What I find most hopeful, spending time in this ecosystem, is that the Japanese approach is not culturally untransferable," said Filipčić. That is a case worth making to any organization building or procuring technology today; it is vital to assess whether the values embedded in your systems—by design or by default—are the ones you actually intended.

In Japan, museums and research centers examine what the human-machine relationship means

BUILDING SECURITY **FOR THE FUTURE**

As organizations navigate the threat landscape, cybersecurity leaders face a familiar challenge: protecting their environments while adopting new technologies and maintaining operational efficiency.

Cyber Hero® Frontline spoke with Rob Allen, Chief Product Officer at ThreatLocker®, about the experiences that shaped his approach to cybersecurity, the shift toward prevention-first security, and the role of Zero Trust in securing the future of AI-driven environments.

You've spent much of your career working directly with customers before becoming Chief Product Officer at ThreatLocker. What first drew you into IT, and what ultimately led you here?

What first drew me into IT was a love of computers. I got my first computer, a Sinclair ZX Spectrum, when I was seven or eight years old (which shows my age). I remember "hacking" games by interrupting them as they loaded, getting into the BASIC code, to give myself more lives or experiment with how they worked. That curiosity continued into the early days of the internet, when I ran up enormous dial-up charges on my poor parents' phone bill chatting with people in other countries on Internet Relay Chat (IRC). That curiosity has stayed with me ever since.

I consider myself enormously fortunate to have found a career that allows me to do what I love every day. If I weren't breaking, fixing, and playing with computers for work, I'd be doing it in my spare time.

What led me to ThreatLocker was Danny Jenkins. I've known him for the best part of 15 years, and from the first time I met him, he impressed me with both his knowledge and his drive. When he approached me about ThreatLocker and said, to paraphrase, "This is what we do, it's amazing, and I'd like you to be part of it," it didn't take me long to decide to come on board.

Before joining ThreatLocker, you spent years helping businesses recover from cyberattacks. What experiences still influence the way you think about cybersecurity today?

For much of my time working with these businesses, particularly before 2020, cyberattacks were primarily disruptive. They were costly, but generally we would help our customers recover, and then everyone would get on with their day. They were rarely business-ending.

> "Ransomware was no longer something you could 'just' recover from. Once data had been stolen, restoring from backups or shadow copies could not solve the whole problem

Whenever an attack happened, though, I was always left with nagging questions: How did this happen? What could we have done differently? Was there something else we could have done to prevent it? At the time, I didn't know that something like ThreatLocker existed, so it was frustrating to feel that there was nothing more I could do.

One event in particular stands out, and, ironically, it happened the week before I joined ThreatLocker. A customer called me early on a Monday morning to say they couldn't access their files. I logged in, worked out what was happening, stopped it, and began restoring their data. What was different this time was the ransom note. Unlike any I had seen before, it included a message that read: "We have uploaded 1.6 GB of your data. If you don't pay, we will begin releasing it on this day at this time." It completely changed my perspective. Ransomware was no longer something you could "just" recover from. Once data had been stolen, restoring from backups or shadow copies could not solve the whole problem. It was something that had to be stopped before it happened.

Fortunately, the following week I arrived at exactly the kind of company that could protect businesses like the ones I had been helping.

As Chief Product Officer, you're responsible for deciding where ThreatLocker invests its time and engineering effort. How do you identify which cybersecurity trends are worth acting on?

We don't start with a trend or a product category. We start with a problem we want to solve. A cybersecurity trend is worth acting on when it exposes a real, recurring problem; the potential consequences are significant; and existing solutions don't address it effectively.

Our recent **Zero Trust Network Access** and **Zero Trust Cloud Access** offerings are good examples. Development didn't begin with someone saying, "We need to build a Zero Trust access product."

> Attackers can steal encrypted data today and retain it until the technology to decrypt it exists. For sensitive information with a long lifespan, that means the risk may already exist

It began with the problems created by traditional remote access. Every port exposed to the internet increases an organization's attack surface, and VPN ports have proven to be a particularly attractive target because of the access they protect. We asked how we could provide the access people need without exposing that same entry point, and our **Zero Trust Network Access** solution grew from solving that problem.

Zero Trust Cloud Access developed in much the same way. Phishing remains a serious problem, and the theft of credentials and session tokens can allow attackers to bypass traditional security controls. We started by asking how we could prevent those attacks and make stolen credentials or tokens less useful.

That is ultimately how we decide where to invest. We don't act on a trend because it is receiving attention. We look at the underlying problem and ask whether we can solve it in a meaningful way.

Where do the best product ideas come from? Are they driven by customer feedback, emerging threats, your engineering teams, or a combination of all three?

They come from all those sources, as well as from our advisory boards and our solution engineers, who live and breathe the ThreatLocker portal every day. The best ideas often emerge where those different perspectives intersect. Customers tell us where they are experiencing friction, emerging threats reveal new problems to solve, and our engineering teams help us understand what is technically possible.

Direct customer feedback is invaluable. I'm fortunate to attend many industry events where I can meet and speak directly with ThreatLocker customers. I also still enjoy working with customers on a day-to-day basis, helping them implement our solutions and hearing their thoughts and feedback firsthand.

Before moving into my current role, the feedback I most enjoyed hearing when speaking to customers was, "You guys are awesome. You help me sleep at night knowing that ThreatLocker is protecting us."

The feedback I value most now is when customers tell me what we're not doing well, what isn't working for them, or what we could improve. It isn't always comfortable to hear, but that kind of feedback is a goldmine. It helps us identify real problems and make the product better for everyone.

AI is reshaping cybersecurity at an incredible pace. Which changes do you think will have the biggest impact on defenders over the next few years?

The biggest impact of AI has been, and will continue to be, speed and scale. AI will allow attackers to create more convincing phishing campaigns, research their targets, identify weaknesses, and develop or modify malicious code much faster. It may not create entirely new types of attack, but it will make existing attacks easier to launch while lowering the barrier to entry for attackers, allowing more people to carry them out.

For defenders, AI will help analyze enormous amounts of data, identify patterns, and automate routine investigation and response. I don't believe it will replace security professionals, but those who learn to use it effectively will be able to work far more efficiently.

AI agents are already creating a new attack surface, one that many organizations are unprepared for. As they give agents access to sensitive data and the ability to take actions, they will need to be controlled in much the same way as users and applications. They should only have access to what they need, their actions should be auditable, and organizations must be able to shut them down when they behave unexpectedly.

AI may be changing the pace of cybersecurity, but the fundamentals remain the same. Reducing the attack surface, applying least privilege, and preventing unauthorized activity will become even more important in the future.

Looking beyond AI, are there any emerging threats or industry trends you believe organizations aren't paying enough attention to yet?

Quantum computing is one area I don't believe organizations are paying enough attention to. The immediate threat is not that quantum computers can break current encryption, but that organizations may be underestimating how long it will take to prepare for that eventuality.

Attackers can steal encrypted data today and retain it until the technology to decrypt it exists. For sensitive information with a long lifespan, that means the risk may already exist.

Preparing for quantum computing requires understanding where and how

cryptography is used, identifying the systems and data that will be affected, and beginning to plan the transition to post-quantum standards. That process could take years. If organizations wait until quantum computing becomes an immediate threat, it will probably already be too late.

Security teams are under constant pressure to do more with less. How do you see the role of cybersecurity evolving over the next five years, and what skills will be most valuable?

Over the next five years, cybersecurity will need to focus more on prevention than detection if organizations are going to do more with less. Security teams cannot keep responding to increasing volumes of threats by hiring more people or adding more tools and dashboards. Security products need to reduce workload rather than generate more alerts.

Most routine investigation and response will likely be automated, allowing people to concentrate on policy, exceptions, and the decisions that carry the greatest risk. AI will play an important role in that change, but accountability will remain with people. Security professionals will need to know how to question AI-generated conclusions, validate the information they receive, and recognize when important context is missing.

The most valuable skills will still begin with a strong understanding of fundamentals. Technologies and tools will change, but networking, operating systems, and application behavior will remain essential. Curiosity, critical thinking, and the ability to explain complex technical concepts and risks in terms the business can understand will continue to be incredibly valuable.

Where do you see ThreatLocker and Zero Trust fitting into that future? What role do you want the platform to play as organizations rethink their security strategies?

I believe that, in some ways, the more things change, the more they stay the same. New and more advanced malware, the risks presented by agentic AI, and the growing number and severity of vulnerabilities all bring us back to the same basic security principles.

You don't need to identify every piece of malware in existence if unapproved software is denied by default. Organizations can allow AI agents to operate in their environments if appropriate guardrails control what they can access and what actions they can take. As frontier models such as Claude Mythos discover more vulnerabilities, strong controls can significantly limit the value of even a successful exploit by restricting what a compromised application can do and reducing the potential blast radius.

That is where I see ThreatLocker and Zero Trust fitting into the future. I want ThreatLocker to give organizations the control they need to adopt new technologies without giving those technologies unrestricted access. The platform should make deny-by-default security, least privilege, and containment practical to operate at scale, helping organizations remain secure as threats continue to evolve.

Finally, if you could leave CISOs and IT leaders with one piece of advice as they prepare for the future of cybersecurity, what would it be?

Don't build your security strategy around trying to predict the next attack. Build an environment where the next attack has as little opportunity to succeed as possible.

You cannot identify every piece of malware, anticipate every vulnerability, or guarantee that a credential will never be compromised. You can control what is allowed to run, what users and applications can access, and what actions they are permitted to take.

Focus on reducing the attack surface, enforcing least privilege, and containing anything that does get through. If you get those fundamentals right, you will be better prepared not only for the threats you know about today, but also for those nobody has seen yet.

> "Don't build your security strategy around trying to predict the next attack. Build an environment where the next attack has as little opportunity to succeed as possible

TOO MANY PHISH

AI-generated voices, fake executives, deepfake video calls, QR-code scams, and poisoned search results have turned phishing into something far more difficult to spot

Old-school phishing emails have become very familiar. Bad grammar, fake bank warnings, odd attachments, and demands to act now or lose access to something important. For years, most security advice boiled down to teaching people how to spot those obvious red flags before doing something they would regret. That version of phishing still exists in volume, but it now sits alongside scams that look far more convincing.

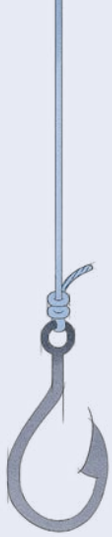
Modern phishing attacks increasingly resemble normal business communication because they are built on the same tools, platforms, and workflows people already trust every day. AI-generated content is accelerating this trend by enabling attackers to produce more convincing emails, cloned voices, fake video calls, and highly personalized scams at scale. The result is a phishing industry that now stretches far beyond inboxes.

Workplaces operate in an environment laden with messages, from team chats to shared documents, and employees are expected to respond quickly across all of them. Attackers have adapted accordingly, building scams that blend far more naturally into the pace and appearance of ordinary working life. IBM's 2025 Cost of a Data Breach Report found phishing was the most common initial attack vector among studied breaches, accounting for 16% of attacks, with an average breach cost of USD 4.8 million.

Phishing breaches cost organizations an average of

USD 4.8 million

in 2025



SPEAR PHISHING AND BUSINESS EMAIL COMPROMISE

Classic email phishing still works well, particularly when attackers narrow their focus. Spear phishing attacks, for example, have become far more tailored than the generic scams that used to clog inboxes. Instead of sending identical messages to thousands of random users, attackers now spend time gathering information about specific employees, teams, suppliers, or executives before making contact. A phishing email referencing a real invoice, an ongoing project, or an actual coworker is far more likely to slip past someone who is already busy and expects dozens of legitimate requests throughout the day.

Business email compromise (BEC) scams often follow the same pattern. Rather than relying on malware, attackers simply impersonate trusted contacts or take over genuine accounts for long enough to insert fake bank details or urgent payment requests into existing conversations without immediately attracting attention. In many cases, there is no malware involved at all, just a message that is believable enough for someone to follow the instructions without questioning them.

Whaling takes the same approach further by targeting executives or senior decision-makers directly. Rather than blasting thousands of generic emails, attackers focus on a handful of high-value targets where a single successful compromise could expose sensitive systems or authorize large financial transfers.

Business email compromise cost U.S. organizations

USD 2.77 billion

across 21,000+ incidents in 2024[‡]

SMISHING, QUISHING, AND POISONED SEARCH RESULTS

Phishing has also spread into places many users still do not associate with cybercrime. Smishing attacks use text messages instead of emails and often impersonate delivery firms, banks, or government agencies. Fake package notifications, unpaid parking fines, and suspicious account warnings are all commonly used to pressure victims into clicking malicious links without a second thought.

QR-code scams, known as quishing, have also become more common in recent years. A malicious code printed on a fake parking notice, delivery slip, invoice, or office poster can send victims to credential-harvesting sites without ever showing a suspicious-looking URL first. By the time someone realizes where the link leads, they may already have entered passwords or payment details.

Search engine optimization (SEO) poisoning means search engines have also become useful hunting grounds for phishing campaigns. A fake login page for a bank, payroll platform, cloud service, or cryptocurrency exchange only needs to look convincing for a few seconds to succeed. Users searching for familiar services can easily end up on cloned sites without immediately realizing they clicked the wrong result.

Clone phishing uses emails that feel familiar, like genuine-looking document notifications, invoices, password resets, and software alerts. They alter the links or attachments while leaving the rest largely intact. During a busy workday, when people are skimming dozens of messages at speed, those changes can be easy to miss.

THE RISE OF VISHING AND DEEPFAKE SCAMS

Voice phishing, or vishing, has become far more convincing thanks to advances in AI-generated audio. A few years ago, most phone scams still sounded exactly like phone scams. The caller would stumble through a script, mispronounce names, or deliver lines with all the emotional realism of a parking meter. That has started changing as cheap voice-generation tools become easier to access and increasingly believable.

That changes the psychology of the attack considerably. An employee receiving a phone call from someone who sounds like their manager may be far less likely to stop and verify instructions, particularly if the request appears urgent or financially sensitive. Criminals have already used cloned voices to impersonate executives requesting payments and family members claiming to be injured or arrested.

Deepfake video scams are creating even stranger situations. Early in 2024, fraudsters used AI-generated versions of senior executives during a video conference call to persuade an employee at engineering firm Arup to transfer roughly USD 25 million[‡]. Nobody hacked their way through a firewall or deployed malware. The employee believed they were sitting in an ordinary meeting with familiar company leadership.

Security teams spent years teaching employees not to trust suspicious emails. Very few organizations prepared workers for fake Zoom calls involving apparently recognizable executives or trusted colleagues. The Cybersecurity and Infrastructure Security Agency (CISA) has warned that Scattered Spider actors use techniques including push bombing, SIM swapping, and help desk impersonation to obtain credentials or account access. The aim is to persuade the people around security technology to reset multi-factor authentication (MFA), approve a login, or treat the attacker as a legitimate employee locked out of an account.

In that sense, the reach of phishing has grown, and it has become a way of turning support processes, identity checks, and recovery workflows into attack paths.

MFA FATIGUE AND AUTHENTICATION ATTACKS

Attackers are also adapting to the widespread adoption of MFA. Fatigue attacks bombard users with repeated authentication prompts until someone eventually accepts one out of frustration, confusion, or distraction. Other phishing kits perform authentication token interception in real time, allowing attackers to bypass certain MFA protections while maintaining the appearance of legitimate logins.

Nobody hacked their way through a firewall or deployed malware. The employee believed they were sitting in an ordinary meeting with familiar company leadership

For many attackers, getting hold of legitimate logins is now more useful than dropping malware onto a device. A stolen password or active session can open the door to email accounts, collaboration tools, cloud platforms, and internal systems without creating the sort of obvious disruption normally associated with malicious software. Once inside, attackers can often move around quietly because the activity looks much like ordinary day-to-day work. That makes phishing particularly difficult to defend against because the attack often appears to be normal user activity after the initial compromise.

WHY PHISHING STILL WORKS

Despite years of awareness training and expensive security tooling, phishing remains effective because it targets people rather than software vulnerabilities. Most organizations have become reasonably good at blocking known malware, filtering suspicious attachments, and detecting unusual network activity. But humans are much harder to secure consistently.

THREATLOCKER® TIP

PLAN FOR THE CLICK

Modern phishing does not announce itself. A user may approve a login, follow a convincing link, hand over credentials, or trust a fake support request. At that point, an attacker is flush with options as to what to do next.

ThreatLocker helps organizations reduce the blast radius of successful phishing by enforcing a full-spectrum deny-by-default approach across endpoints and access. **Zero Trust Cloud Access** helps neutralize phishing and token theft, ensuring that access can only happen from an approved, cataloged device. **Zero Trust Network Access** protects within the perimeter, with robust policy management capabilities that ensure users and devices can access only what they genuinely need. **Allowlisting** can stop untrusted executables, scripts, and tools from running, even if a user has been fooled into downloading them. **Ringfencing™** can limit what approved applications are allowed to access, helping prevent trusted tools from being abused.

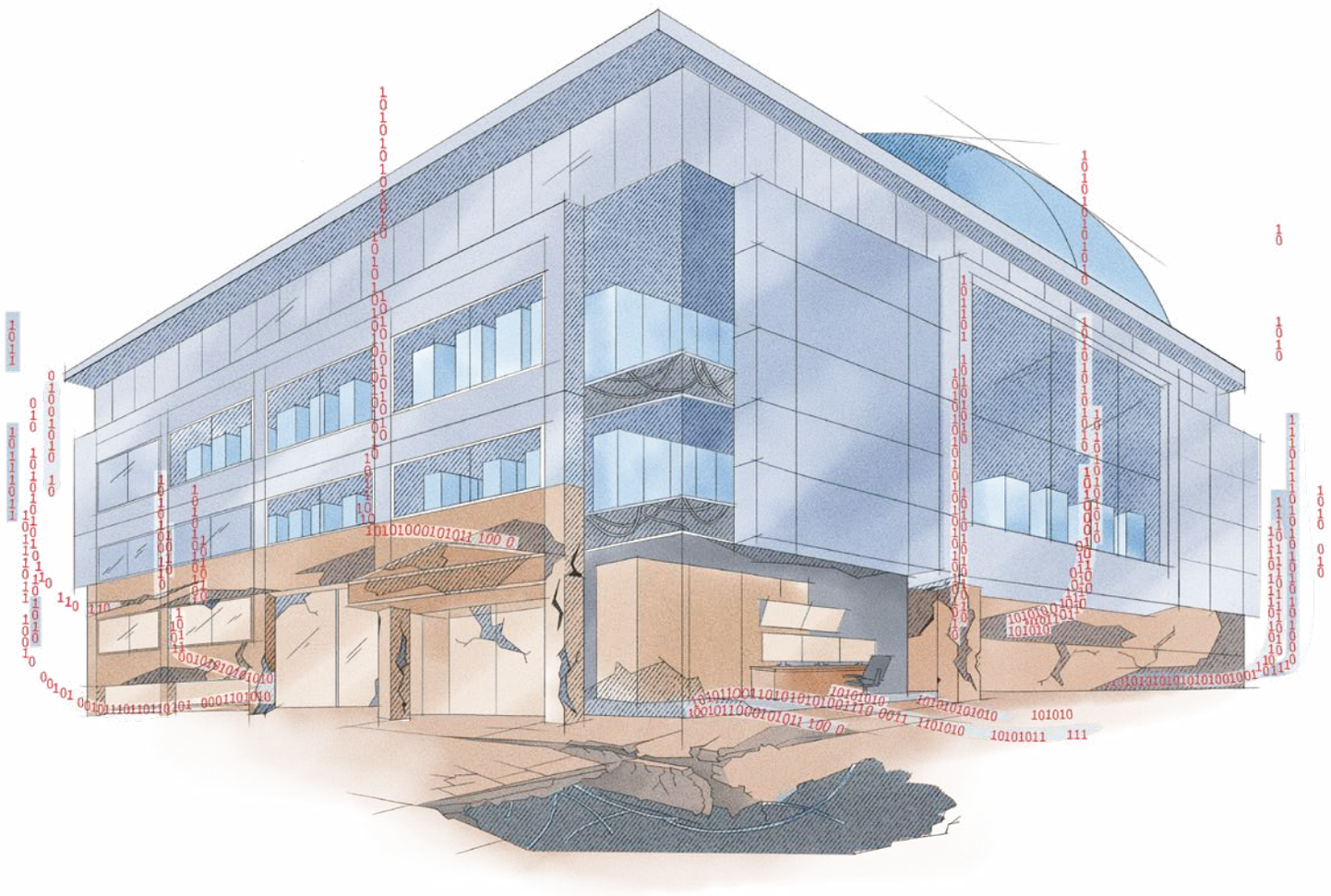
Phishing is designed to exploit trust. ThreatLocker capabilities are built under the assumption that trust has already been exploited. Nothing should run or connect unless it has been explicitly allowed—even if a phishing attempt is successful, an attacker must not find anything of value at the end of the line.

Many successful phishing attacks do not rely on technical exploits at all. They succeed because someone is tired, distracted, multitasking, or under pressure to respond quickly. AI-generated content simply makes the deception more convincing and easier to scale.

More organizations are responding by tightening how access works across their networks and internal systems. Security teams are placing greater emphasis on additional verification for sensitive actions, limiting how much access individual accounts receive, and watching more closely for unusual behavior after login. The thinking behind those measures is fairly straightforward: Phishing attacks are becoming difficult to distinguish from ordinary communication, so companies are planning around the assumption that some scams will eventually work.

An attempt no longer needs to look obviously fake to succeed. In many cases, the scams only need to look normal long enough for someone to trust them.

GOVERNMENTS UNDER FIRE



When the pace of digitalization does not match the speed of attack development, the developing world stands to fall behind and leave itself open to cyberattacks

Balancing the books in a fast-growing, rapidly changing city like Johannesburg, South Africa, is a tough job for municipal managers at the best of times. Emerging markets are some of the most dynamic and exciting places to do business today, and as the richest city in sub-Saharan Africa, Joburg—as it is affectionately called by residents—is brimming with both opportunity and challenge. For city leaders, financial worries are never far away, as the city's expanding geographical and demographic footprints strain services to the breaking point.

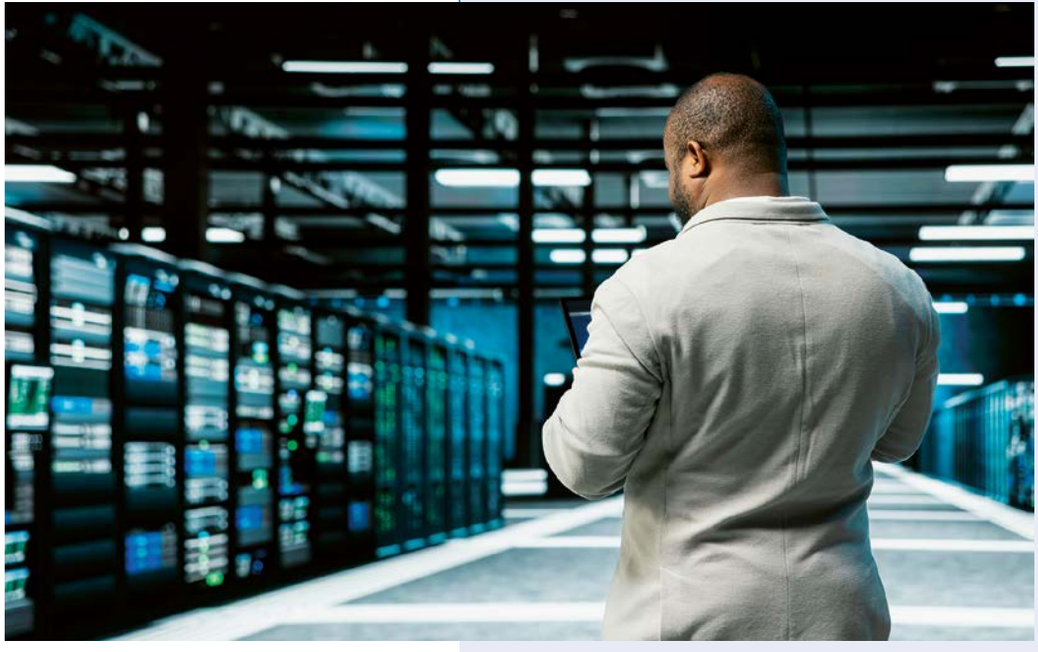
So, when it emerged that cybercriminals had managed to defraud the Ekurhuleni Metropolitan Municipality on the city's east side of an estimated ZAR 2 billion[‡] (USD 123 million), it was a big deal. Using a combination of attacks, including the deployment of spyware, identity theft, and physical access to the metro's network, thieves were able to manipulate billing accounts without leaving an audit trail. The attack would have been devastating enough on its own, except that it was just one of multiple high-profile, successful security breaches against state-run IT infrastructure that have been reported in the nation recently.

In the first half of 2026, the Gauteng provincial government lost almost 4TB of residents' data in a ransomware attack[‡]; Statistics South Africa (Stats SA) acknowledged that the same criminal group had accessed its HR database; and the records of two million members of the African National Congress—the country's largest political party—were published online. In May, distributed denial-of-service (DDoS) attacks successfully overwhelmed several major internet service providers and disrupted connectivity across the nation[‡], and over the last five years, South Africa's Department of Defence (DoD) has suffered three breaches, including the loss of contract data and sensitive service records[‡].

POOR SOLACE

The Ekurhuleni incident highlights all the issues that Zero Trust security is designed to mitigate: It was the result of unsecured accounts, elevated levels of access, poor audit logging, and no warning systems to flag suspicious behavior. South Africa, too, has an unusual risk profile for cybersecurity. It is not just the largest economy on the continent. The country occupies an almost singular diplomatic position, being close to the U.S. and Europe economically and politically, but also a member of the BRICS bloc alongside Russia.

Using a combination of attacks, including the deployment of spyware, identity theft, and physical access to the metro's network, thieves were able to manipulate billing accounts without leaving an audit trail



As one of Africa's leading business centers, Johannesburg highlights the increasing need for strong cyber resilience in an interconnected world

State-sponsored actors and activist groups have targeted South African government IT systems over a range of geopolitical and domestic issues. But it would be a mistake to view South Africa as the exception rather than the rule for emerging markets.

Elsewhere in Africa, an attack on B'Odogwu, the Nigeria Customs Service's core trade-facilitation platform, caused chaos across ports in the country in August 2025[‡]. Late last year, a coordinated attack defaced government websites for multiple services across Kenya[‡]. A breach at Morocco's National Social Security Fund yielded financial details of two million people to attackers[‡].

OUT OF AFRICA

Meanwhile, the Puerto Rican government was forced to cancel all public appointments at Centros de Servicios al Conductor (CESCO), the agency responsible for issuing driver's licenses and vehicle registrations, following a cyberattack. In Mexico, multiple government departments were attacked over a two-month period, during which it has been claimed that 195 million taxpayer records were stolen[‡]. A World Bank report, Cybersecurity for Emerging Markets, recently identified Latin America as the fastest-growing region for cybercrime, and in countries such as Venezuela, Peru, Guatemala, and El Salvador, most reported incidents are in the public sector.

The African Development Bank has invested nearly **USD 3 billion** in information and communications technology and connectivity infrastructure across the continent

Interpol also warns of dangers in Southeast Asia, where countries such as the Philippines, Sri Lanka, and Cambodia are home to a USD 40 billion industry of "scam centers." These criminal syndicates conduct fraudulent activities on a global scale; for example, a phishing attack through a fake dating application profile targeting a New York romance seeker might originate from one of these groups. But they produce serious threats in their home countries.

Nearby, it has been well documented that while border tensions between Pakistan and India sometimes result in physical incidents, hacktivists aligned to both sides are also reported to routinely target government services and critical infrastructure.

According to the World Bank report, government bodies across emerging markets are both particularly at risk and increasingly seeing that risk exploited. The World Economic Forum's (WEF) 2026 Global Cybersecurity Outlook provides more evidence: Business leaders surveyed reported the greatest threat to cybersecurity as being geopolitically motivated attacks, and those from Latin America and sub-Saharan Africa were far less likely to have confidence in their government's ability to successfully protect infrastructure.

MODERNIZING THE CIVIL SERVICE

As in any sector, the historical weaknesses of public administration networks are exacerbated by the increasing number and constant evolution of phishing attacks and zero-day exploits powered by large language model (LLM) tools.

There are other driving factors for the high level of risk, rooted in historical challenges for the public sector. On the one hand, there is enormous pressure to digitize public services, both for government delivery and to support the private sector as it rapidly digitalizes (online food delivery is a market worth more than USD 500 million in Kenya alone)[‡].

The African Digital Compact (ADC), an initiative of the African Union, urges shared frameworks, interoperability, and resources to drive digital transformation across the continent. Digital public infrastructure (DPI) for the continent was a key talking point at last year's G20 conference in Johannesburg, and the African Development Bank has invested nearly USD 3 billion in information and communications technology and connectivity infrastructure across the continent over the past decade[‡].

This digitalization inevitably increases the attack surface for the public sector. State-backed Chinese hackers, for example, infiltrated multiple government networks across Asia using a zero-day exploit that had been discovered in a popular videoconferencing application.

HIGH COMPETITION

While policies and regulations are catching up with environmental realities, competition for resources on the ground is higher than ever. A scoping review of electronic health records in sub-Saharan Africa over a 10-year period showed their unquestionable benefits in enhancing patient care, but highlighted recurring issues with infrastructure, data management, and training[‡].

There is difficulty with recruitment: Salaries for cybersecurity professionals in emerging markets rival those elsewhere in the world, as their skills are in such high demand. The largest private-sector companies can afford to attract the best local talent, while public-sector organizations struggle to offer competitive salaries, working conditions, and a cybersecurity-focused internal culture to support IT professionals' work.

According to the WEF report, 57% of public-sector organizations say they lack the skills and resources needed for effective cybersecurity, compared with just 29% of large corporations. The World Bank estimates that government spending on cybersecurity is USD 30 per capita in Canada, compared with just USD 1 per capita in Mexico or India.

DEALING WITH COMPLEX ENVIRONMENTS

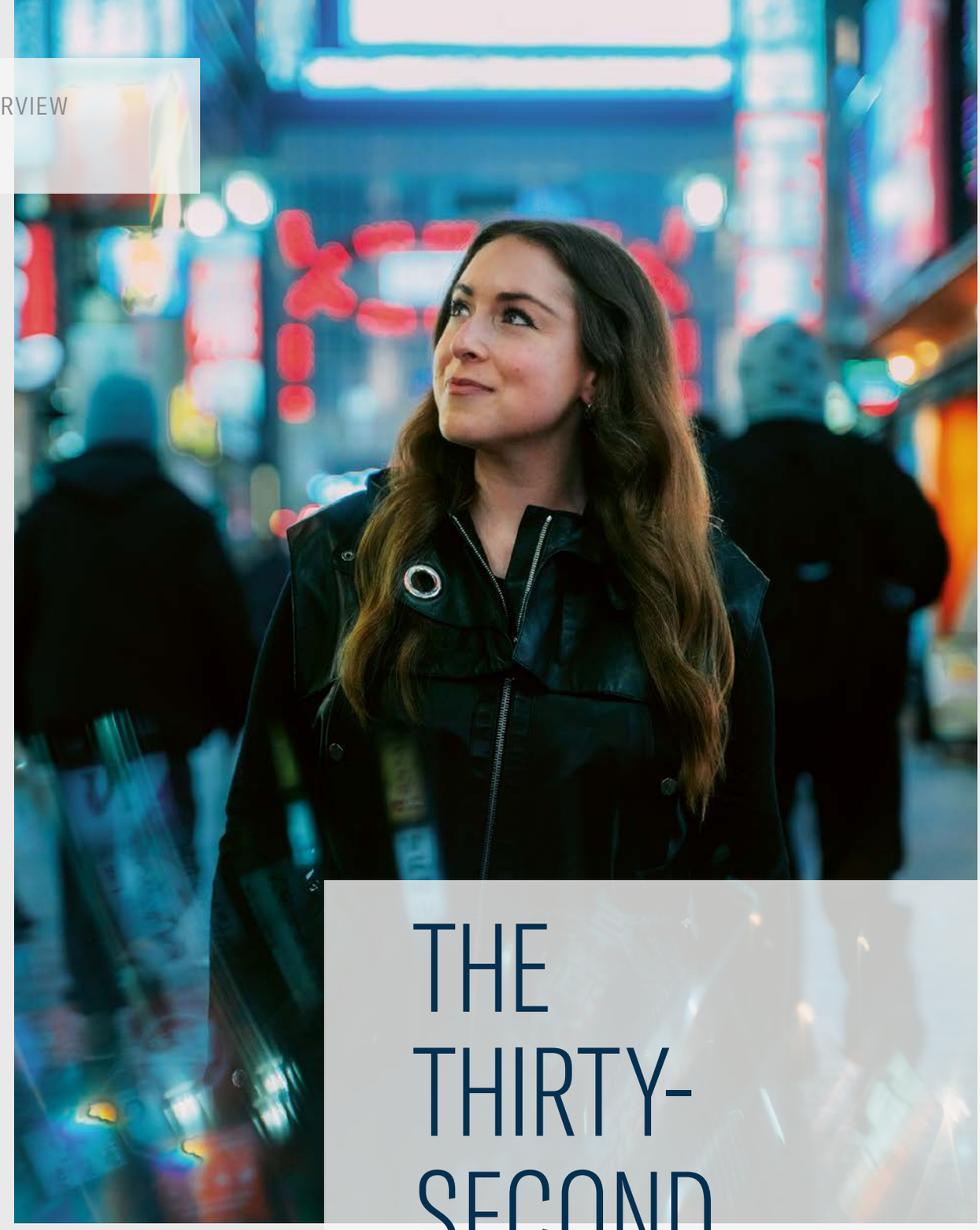
Emerging markets must continue to digitize their public services and build out DPI to meet the needs of connected citizens and businesses, but investments in cybersecurity need to be increased and made in strategically sensible ways.

The principles of Zero Trust security are essential to securing complex environments in the most cost-efficient manner. Indeed, the attack on Ekurhuleni municipality is a prime example of when Zero Trust could have prevented the problem from occurring: It could potentially have been thwarted by blocking unauthorized or unknown access to critical systems, or by using **Ringfencing™**, which restricts application actions. Effective security also requires keeping audit logs to identify suspicious behavior. Deny-by-default is the best way to prevent AI-spawned ransomware attacks on databases of citizen information.

There is a clear lesson to be learned here. Governments worldwide must use the tools and techniques required to create country-level resilience to cyberthreats, and act fast. To quote the World Bank Group's cyber resilience team, "Building secure, resilient, and inclusive digital futures is one of the defining development challenges of our time."

THREATLOCKER® TIP

Learn more about the way ThreatLocker capabilities protect governments and contractors



THE THIRTY-SECOND BREACH

Rachel Tobac, CEO and Co-founder of SocialProof Security, reveals that the most damaging attack methods are often the least complex—and businesses are ill-equipped to cope

The first mistake security professionals make is expecting social engineering to look clever. "People think that social engineering is going to be this really sophisticated, elaborate scam," said Rachel Tobac. "Most of the time someone just sends an invoice pretending to be a customer, and then the company pays it." Or it may be a customer-support call from someone pretending to be a real customer: "Hey, I changed my email address. Can you change it on your end? Simple phone call. Takes less than 30 seconds."

Tobac's work in testing and training against social engineering means she can often tell when a live test is going to work. Imagine she calls a help desk pretending to be an employee called Jack Smith. The story behind the ticket is usually familiar: The employee's phone has stopped working, multi-factor authentication (MFA) is broken, the account needs a new device, and the password must be reset.

"If they ask me to verify that I am who I say I am, perhaps by asking for a phone number or email address, they're usually asking for information that is widely available on the internet. We can find it within a few seconds of a Google search, because data brokerage sites publish this information. If that happens, then I know for the most part it's going to be over for this company."

"Jack Smith might, for instance, have admin access," Tobac continued. "If all I need to be able to take over his account is publicly available information, easily found with basic open-source intelligence (OSINT) skills, then I can probably get into anybody's account and steal data, money, and access."

"My expectation for attackers is that they're just going to continue running the same playbook, because it works. Attackers only do what works because they run like a business. They have money to make

SPEED IS WEAKNESS

Tobac sees the same weak logic shown repeatedly: organizations operating under pressure, fighting against tight service level agreements which give them a strong motivation to move quickly, yet with no serious way to verify identity beyond information an attacker can collect online. A help desk may ask for a date of birth, or HR may ask for a manager's name before changing a direct-deposit bank account. "Okay, I found that on RocketReach within 10 seconds," she said. "That doesn't mean anything to me."

She calls this "OSINT-able knowledge-based authentication." It covers the small facts organizations still treat as if they are private, like date of birth, email address, phone number, boss' name, or length of employment. "Looking for that kind of data is not sufficient," she explained. "It's like saying, 'We have locks on our doors, they're fantastic,' while leaving the windows wide open."

Asked what the next major breach caused by human error will look like, Tobac suggests things are unlikely to change. "I'd probably say that the next attack is going to look like all the attacks we've been seeing from the past year," she said. A help desk call, compromised credentials, reset MFA, and then theft of data or a request for a wire transfer to an attacker-controlled bank account. "That's the sad thing," she continued. "My expectation for attackers is that they're just going to continue running the same playbook, because it works. Attackers only do what works because they run like a business. They have money to make."

ABOUT

Rachel Tobac is CEO and Co-founder of SocialProof Security, a social engineering prevention company that trains and tests organizations against the tactics attackers use to exploit trust, urgency, and routine. The company describes its team as "friendly hackers," using live social engineering exercises to help strengthen what Tobac calls an organization's first line of defense: its people.

Tobac is an ethical hacker and social engineering specialist known for turning attacker psychology into plain language. She placed second three years running in DEF CON's Social Engineering Capture the Flag competition and has served on CISA's Technical Advisory Council. Her work spans phishing, vishing, human-layer security, and security awareness training, including attacks shaped by AI.

SocialProof Security's stated goal is to work itself out of a job by helping companies and individuals become what the company calls "Politely Paranoid": calm enough to keep working, but skeptical enough to verify before trust becomes a breach.

EASY ACCESS

Tobac points to ShinyHunters and Scattered Spider as groups leaning on base-level social engineering for access. "The vector that they've been using the most successfully is the phone. If they don't have to do a bunch of work to set up domains, email phishing infrastructure, text messages, or spoofing—if they can just run a playbook where they place a phone call and they gain access to an account—then they're going to do the easy thing."

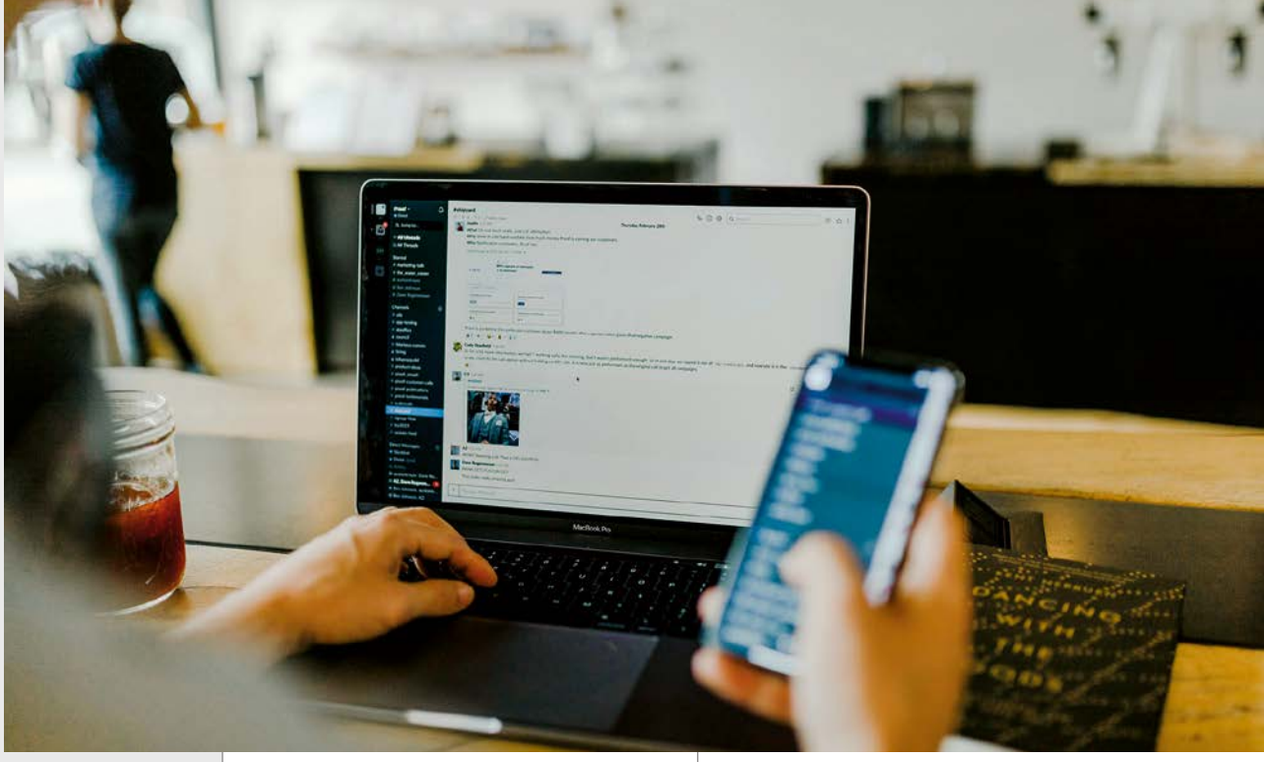
That creates a problem for awareness programs built primarily around email. "Okay, now I know what a phishing message looks like, but I wasn't actually told what I should do when I see one," she said. "Or they told me what to do when I see a phishing email, but most messages that I get that seem like phishing are text messages and phone calls. What do I do about a phone call? There's no button to press in Outlook when I'm on the phone, right?"

> "If organizations can [verify me through another channel] when I'm testing their defenses, they're going to catch me nine times out of 10

REIMAGINING VERIFICATION

Staff can understand social engineering perfectly well and still be set up to fail. "I may be a trained worker, say, but I am still expected to respond quickly to my executive's request," suggested Tobac, "and I am not expected to verify that that person is who they say they are, and I have to do it quickly, or else I'll get fired." A better program gives people a script, a protocol, and permission. If an executive assistant receives a call, text, or email from an executive asking for a password before a board meeting, the response should be cautious by default. "I can help you, but first let me verify your request."

Tobac says that verification should happen through a separate channel. If the request arrives by phone, confirm through Slack, or internal email. If it arrives through chat, use another route. "When we, as attackers, are gaining access at a company, we gain access to one channel at a time," she said. "We don't have access to your phone line, your internal email thread with your executive, your Slack, and your WhatsApp at the same time. That's precisely why we're attacking, because we don't have that level of access yet."



For HR, that might mean responding to a payroll change request by sending a word through Slack and asking the person to read it out. If the caller says Slack is broken, HR can use an approved fallback. "It's as simple as that," Tobac said. "Another method of communication to verify I am who I say I am. If organizations can do that when I'm testing their defenses, they're going to catch me nine times out of 10."

FIGHTING FRICTION

"You're always going to have a trade-off between security and convenience," she explained. "There is no way around that. But if an attacker steals all your data and leaks it online, you're going to have 150 hours of work just dealing with that versus the 30 seconds that it takes to look at your authentication app." Her advice for lowering friction is to not verify every human interaction, but instead tailor verification to the request.

"It's for sensitive actions," said Tobac. Passwords, wire transfers, data movement, and access changes should trigger verification, as should location details in a higher-risk threat model. "You have to evaluate the threat model," she says. "Who is this person? Why do we have to protect them? What are they currently dealing with?"

THE BLAME GAME

After a social engineering attack, the blame often lands on the CISO or the employee who fell for it. Tobac is cautious about allowing such reflexive finger-pointing. "You can do everything right and somebody can still fall for a scam," she says. "Most human beings will make a mistake at some point in their life. We need to take a softer approach when it comes to the blame game. It just doesn't really make sense."

The change she would make to most company training is the one she comes back to repeatedly: "In addition to education, we need to update our protocols to verify people are who they say they are." Awareness tells people what social engineering looks like, but protocol tells them what to do when the phone rings.

NO MORE EXCUSES

CISA's latest guidance makes the case for Zero Trust as a practical response to modern cyber risk—particularly in environments that do not seem to suit it

The increasing presence of Zero Trust in government guidance makes it clear that it has become part of mainstream security thinking. Though sometimes softened in practice by presumptions surrounding legacy infrastructure, operational complexity, and the difficulty of organizational change, Zero Trust is now a respected security model embedded in serious cyber guidance.

The latest guidance from the U.S. Cybersecurity and Infrastructure Security Agency (CISA) offers organizations the context needed to overcome those presumptions. It frames Zero Trust as a practical way to protect what organizations actually run: messy, connected, high-consequence environments made up of old systems, new platforms, remote users, connected suppliers, and uneven visibility.

CISA's new guidance also shifts the discussion away from conventional IT estates and toward operational technology (OT). Its April 2026 guide, Adapting Zero Trust Principles to Operational Technology, transposes Zero Trust values onto factory floors and into the plant rooms of critical industries, where constraints are sharper and the consequences of failure can be more serious.

OT systems are inherently difficult to secure. They may include legacy devices, safety-critical processes, and fragile protocols, while uptime requirements can make standard security change procedures difficult. Even in those environments, CISA and its government partners argue that Zero Trust principles can be adapted, prioritized, and applied.

Perimeter-based models and implicit trust are no longer sufficient when attackers can use valid accounts, trusted suppliers, remote access routes, and living-off-the-land (LOTL) techniques to gain and maintain access

NO CLEAN SLATES

Ordinary business environments should take note: OT poses many of the hardest conditions for security change, yet the guidance still pushes toward Zero Trust adoption. CISA recognizes that many organizations do not have the luxury of starting a security rollout from a clean slate. They run old and difficult-to-patch systems as part of mixed estates. Suppliers are often granted remote access, sometimes through poorly governed pathways. Incomplete asset inventories can leave devices forgotten and vulnerable.

The biggest takeaway from the guidance is that Zero Trust principles must be applied in a way to secure real-world systems, rather than demand the creation of idealized ones. A security model that only fits clean, modern, cloud-first environments would, after all, struggle to help the sectors where resilience matters most.

CISA and its government partners, including the FBI, the Department of Energy, and the Department of State, are joined in presenting Zero Trust as

CISA'S OT ZERO TRUST GUIDANCE: KEY RECOMMENDATIONS

CISA is clear that the blanket application of IT-focused Zero Trust capabilities to OT is "neither reasonable nor feasible," so it sets out OT-specific considerations across governance, identification, protection, detection, response, and recovery, mirroring the National Institute of Standards and Technology (NIST) Cybersecurity Framework functions:

Asset visibility: Maintain a comprehensive asset inventory covering hardware, software, systems, and communications. CISA stresses that OT asset discovery may require passive monitoring, as active scanning can disrupt legacy devices.

Supplier access: Bring vendors, integrators, and support providers into the Zero Trust model. Third-party access to OT equipment should require proper authorization and oversight, with remote channels secured, limited, and monitored.

Procurement: Use buying cycles to reduce legacy risk. CISA points to newer OT components that support security logging, secure communications, and identity and access management, enabling stronger monitoring, segmentation, and enforcement. Security requirements should be part of procurement.

Identity: Tailor identity, credential, and access management to OT workflows. CISA recognizes that some OT devices cannot support modern controls, so procedures and compensating controls may be necessary.

IT/OT separation: Avoid direct trust relationships between IT and OT identity systems. Where Active Directory (AD) is used in OT, CISA recommends segmentation, ideally through a separate forest or domain.

Segmentation: Treat segmentation as a core containment control. Network boundaries should reflect operational processes, necessary communications, and the potential consequences of compromise.

Remote access: Restrict and monitor remote OT access, particularly for privileged users and third parties. Access should be granted only where it is necessary, limited, and monitored.

Resilience: Link Zero Trust to incident response and business continuity. CISA calls for OT-specific response plans, isolation procedures, tested backups, restoration checks, and recovery planning that accounts for cyber incidents.

CISA is not writing for perfect environments, nor for teams with unlimited budget, unlimited downtime, and full control over every connected asset. It is written for the awkward reality most security leaders recognize

a practical response to the way modern compromise unfolds. Perimeter-based models and implicit trust are no longer sufficient when attackers can use valid accounts, trusted suppliers, remote access routes, and living-off-the-land (LOTL) techniques to gain and maintain access. This is an argument for a security model which acknowledges the perimeter but also what is within.

SECURITY FOR REALITY

For organizations already considering the Zero Trust path, the guidance removes a familiar excuse. CISA is not writing for perfect environments, nor for teams with unlimited budget, unlimited downtime, and full control over every connected asset. This revised guidance has been written for the awkward reality most security leaders recognize: long-lived systems surrounded by imperfect knowledge, operating under significant pressure.

Risk cannot be engineered away in one pass. Nonetheless, if Zero Trust can be understood to fit environments where downtime is only possible in carefully engineered windows, the case for delay in ordinary business networks is weak at best. That is the point at the heart of the guidance: the complexity of the organization is the reason to act, not a reason to wait.

THE MATURITY BENCHMARK

The latest update to the EU's cybersecurity framework, NCAF, may be aimed at governments, but its measuring tools can make a difference to any business with a cyber supply chain

It has been some time since cybersecurity could be viewed as simply locking down an organization from within. Not only have attacks become broader and more indiscriminate, but the interconnectedness among organizations, services, and even countries means that when a threat strikes, its effects can ripple outward. The growing number of worldwide supply-chain attacks make this very clear.

That thinking initially drove the European Union Agency for Cybersecurity (ENISA), the EU's cybersecurity arm, to develop the National Capabilities Assessment Framework (NCAF), a tool for measuring EU member states' national cybersecurity strategies. The EU has since unveiled the NIS 2 Directive, which requires all EU member states "to develop and adopt a national cybersecurity strategy." This, in turn, has led to the release of an updated version of the NCAF.

WHY THIS MATTERS

There is no law, audit, or product standard tied to the NCAF itself—its role is solely to provide EU member states with the tools to assess the maturity level of their strategies, identify gaps, and plan improvements while harmonizing requirements across borders. But while the NCAF may be squarely aimed at EU member states, it is a key part of a broader worldwide trend, in which most governments—including at state and federal levels in the U.S.—are developing cybersecurity policies and some form of measurement tool to assess their maturity.

While governments worldwide have their own strategies, many recognize the value of harmonizing, where possible, with others to grease the wheels of trade. Where a large supranational organization like the EU goes, other countries will likely follow—such as the U.K., which is steadily realigning itself with NIS 2 after initially diverging post-Brexit.

Therefore, it is not just EU-based organizations with direct ties to their own government (such as supplying services, holding public data, or operating infrastructure) that need to sit up and take notice. The NCAF is a useful tool for any organization, even those not headquartered in the EU. By understanding and following NCAF guidelines, businesses and other organizations should find it easier (and safer) to set up cross-border partnerships while meeting local regulatory requirements.

HOW NCAF 2.0 CAN HELP ORGANIZATIONS

In addition to meeting regulatory requirements (both now and in the future), there are good reasons for paying attention to the latest NCAF. It is a valuable document that provides a useful framework for any business looking to improve its cybersecurity strategy.

The framework provides a maturity-based method of measuring cybersecurity capabilities. There are five levels: Foundation, Developing, Established, Mature, and Advanced, with clear definitions of each in the NCAF 2.0 document and an admission that the Advanced level is more aspirational than immediately applicable. It is characterized as one that "very few countries, if any, are expected to reach … for all objectives."

The measurements are derived from 20 key strategic objectives, organized into four clusters. These clusters span capacity building and awareness, cooperation and collaboration, cybersecurity governance, and regulatory and policy frameworks.

Where a large supranational organization like the EU goes, other countries will likely follow

Some of the objectives are more relevant to nongovernmental organizations than others, such as addressing the cybersecurity skills gap and promoting cyber hygiene and cybersecurity awareness. But there is value in reviewing all objectives to see whether others can be adapted; for example, the objective of strengthening cyber resilience and hygiene in the private sector could be redefined to target the organization's suppliers, contractors, and other partners.

Each objective is assigned its own series of specific goals, and a scoring mechanism has been developed based on two parameters:

Maturity level: This is based on which capabilities and practices have been implemented.

MAPPING NCAF 2.0 TO ZERO TRUST

NCAF 2.0 concern	Zero Trust translation for business
National cybersecurity governance	Clear ownership, board visibility, and policies that are measured rather than assumed
Cyber hygiene	Multi-factor authentication (MFA), patching, least privilege, device management, and secure configuration
Digital identity	Strong identity controls, conditional access, and privileged-access management
Risk-management measures	Access based on risk, sensitivity, role, device posture, and behavior
Incident reporting	Better logs, visibility, response playbooks, and escalation routes
Supply-chain cybersecurity	Contractors and suppliers are given limited, monitored, and revocable access
Critical-sector resilience	Segmentation, data isolation, and continuity plans for essential services
Active cyber protection	Monitoring, detection, automated response, and containment

Coverage ratio: The extent to which an objective's indicators have been answered positively, regardless of maturity level.

The NCAF framework includes a complete set of questions, plus guidelines for deployment and use, providing detailed steps that organizations can adapt to measure and demonstrate their cybersecurity strategies and capabilities, both internally and with external partners and ultimately regulatory bodies.

HOW ZERO TRUST ALIGNS WITH THE NCAF

Any organization with strong cybersecurity practices should score well in key areas, particularly if it has implemented Zero Trust. Although it is mentioned only once in passing in the framework document, a good Zero Trust policy addresses a raft of the concerns raised by NCAF, including areas like cyber hygiene, digital identity, risk management, incident reporting, and supply-chain cybersecurity.

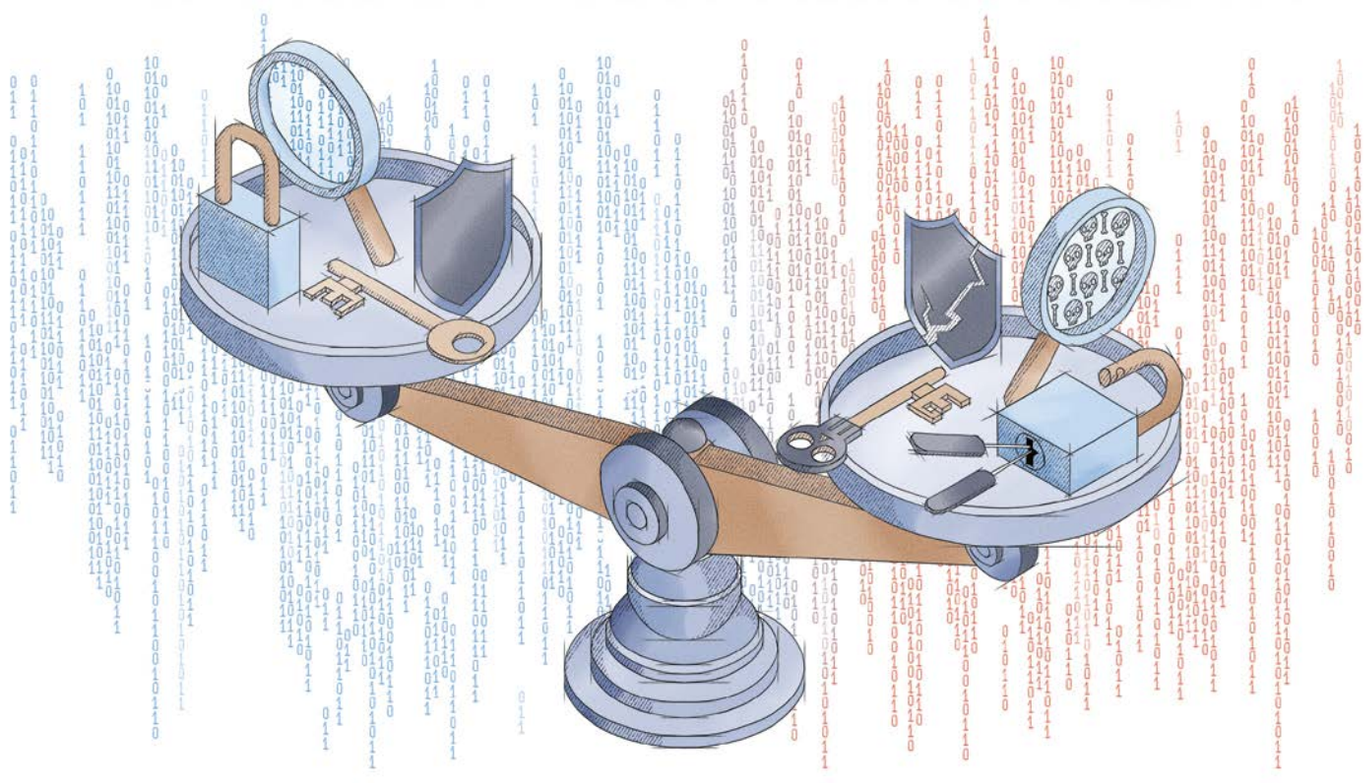
While a complete reading of the NCAF 2.0 document is not strictly necessary, it remains a useful resource for cybersecurity professionals. The entire framework can be adapted to provide a detailed cyber maturity measurement tool, but, failing that, a review of the 20 key strategic objectives could at least flag potential gaps and help determine whether the organization's current security setup is fit for purpose.

THREATLOCKER® TIP

For organizations looking to align with ENISA's latest recommendations, a strong Zero Trust solution like ThreatLocker helps to demonstrate both strong governance and consistent cyber hygiene.

By restricting access to the minimum required level with a capability like **Ringfencing™**, checking all devices before they connect to the organization's network using **Zero Trust Network Access** and **External Storage Device Control**, and isolating sensitive data from other business systems with **Data Storage Access Control**, ThreatLocker helps build a mature and robust cybersecurity management structure. This includes locking down supplier and contractor access to the organization's systems to prevent access to anything outside their strictly defined parameters.

FIGHTING ON **THE AI FRONTIER**



As next-generation AI models like Claude Mythos improve their ability to perform offensive cybersecurity tasks, the industry is worried—but are frontier models really autonomous super-hackers, or a product of the hype machine?

Cybersecurity tends to meet new technology early, and often at its point of greatest uncertainty. Cloud computing forced a rethink of the network perimeter; cryptocurrency opened new ground for fraud, theft, and financial crime; and quantum computing has long been discussed as a future threat to encryption. The conversation can sometimes move faster than the everyday security work most organizations are still trying to complete. And now the industry is meeting a new threat: the AI super-hacker.

Anthropic's Claude Mythos model has emerged as one of the clearest examples of that narrative taking hold. This cannot be entirely attributed to Anthropic's hype machine running in overdrive, though it certainly is active. Research on Mythos suggests it could identify (and, in turn, exploit) vulnerabilities across real-world software environments, which has generated headlines about frontier AI systems rapidly approaching autonomous offensive capability.

To some, the findings came across as a warning of the new capabilities of AI systems, highlighting their ability to conduct offensive cyber operations cheaply and efficiently with a lower skill demand and limited human input. Others saw something more restrained—that this was a tightly controlled demonstration that reflected how aggressively frontier AI companies are now positioning their models as cybersecurity breakthroughs. The truth, as always, is probably somewhere in the middle.

THE MARKETING PROBLEM

Part of the skepticism around Mythos stems from the difficulty of separating genuine technical progress from the increasingly theatrical marketing cycle surrounding large language models (LLMs) and agentic AI. Frontier AI companies operate in a competitive environment where claims of capability drive investment, influence regulation, and shape public perception. Security has become one of the most effective ways to demonstrate seriousness because cybersecurity threats already carry an air of urgency and inevitability. That creates a strong incentive to frame offensive AI capability in the most dramatic terms.

Some researchers have questioned how far the Mythos results can be generalized beyond tightly controlled testing environments. An evaluation by the U.K. AI Security Institute described Mythos operating inside structured cyber ranges where the model was explicitly directed and given network access to complete multi-stage attack simulations. Early industry testers, on the other hand, found that frontier cyber models relied heavily on experienced human researchers to guide workflows, validate findings, and distinguish genuine vulnerabilities from false positives.

WHEN AI FINDS THE BUG FIRST

Frontier AI can discover system weaknesses at a pace beyond human capability. But there is nothing that explicitly positions it as a black hat tool; it has clear promise for defense, for finding weaknesses before attackers do. This is not a new development; Google's Big Sleep project, developed by DeepMind and Project Zero, shows that AI has been a viable defensive tool since at least 2024, when the AI agent identified a previously unknown vulnerability in a pre-release version of SQLite. The issue was reported to developers and fixed before it reached an official release, meaning users were protected before the flaw could become a live problem.

The same ability to reason through code, test assumptions, and identify obscure flaws can be used to shrink the window of exposure. Instead of waiting for criminals to discover and weaponize vulnerabilities, defenders can use AI to hunt for them inside their own software, open-source dependencies, and critical infrastructure code before that code is exposed to the wider world.

This does not remove the need for human researchers. AI-generated findings still need validation, prioritization, and responsible disclosure. But it changes the pace of defensive work. A model that can inspect code, generate hypotheses, and surface likely weaknesses gives security teams more time to patch and developers better evidence of what needs fixing. If AI can accelerate exploitation, it can also accelerate repair.

Academic testing has also produced mixed results. One paper examining Mythos-related bug rediscovery found that advanced models often focused on vulnerabilities that sounded plausible but turned out to be incorrect, while the authors struggled to consistently identify the same flaws highlighted in Anthropic's public examples. The researchers argued that strong benchmark performance does not necessarily translate into reliable autonomous behavior in real-world environments.

WHERE AI STILL FALLS APART

Real networks rarely behave the way lab environments do. Internal tooling breaks, credentials randomly stop working, logs are incomplete, and documentation often bears little resemblance to whatever is actually running inside the environment. Under the complexity of the real world, AI systems often start getting into trouble. Models can head off in completely the wrong direction, make confident but useless suggestions, or get stuck repeating the same bad idea over and over once conditions stop matching the neat environments they were tested against.

Those limitations matter far more in offensive security work than they do when generating routine code or summarizing documents, which is why, at the moment, frontier models look less like autonomous elite hackers and more like highly capable assistants. They dramatically speed up certain parts of the process while still requiring substantial human oversight. That extra pace, however, may still be enough to reshape the threat landscape.

CYBERCRIME DOES NOT NEED AGI

Cybercriminals do not need AI systems to become sentient offensive operators for the technology to create serious problems. These tools, as they are, can reduce criminals' costs, compress their expertise, and increase their operational scale. In many ways, that process is already underway. Threat researchers have documented growing use of AI agents across phishing campaigns, malware development, reconnaissance, and social engineering operations. AI-generated phishing emails are often more convincing than the poorly translated scams that previously dominated inboxes—now targeting the richest prizes, rather than the most gullible targets. LLMs can help inexperienced attackers write scripts, debug malware, explain exploit chains, and automate repetitive research tasks that once required at least some technical knowledge.

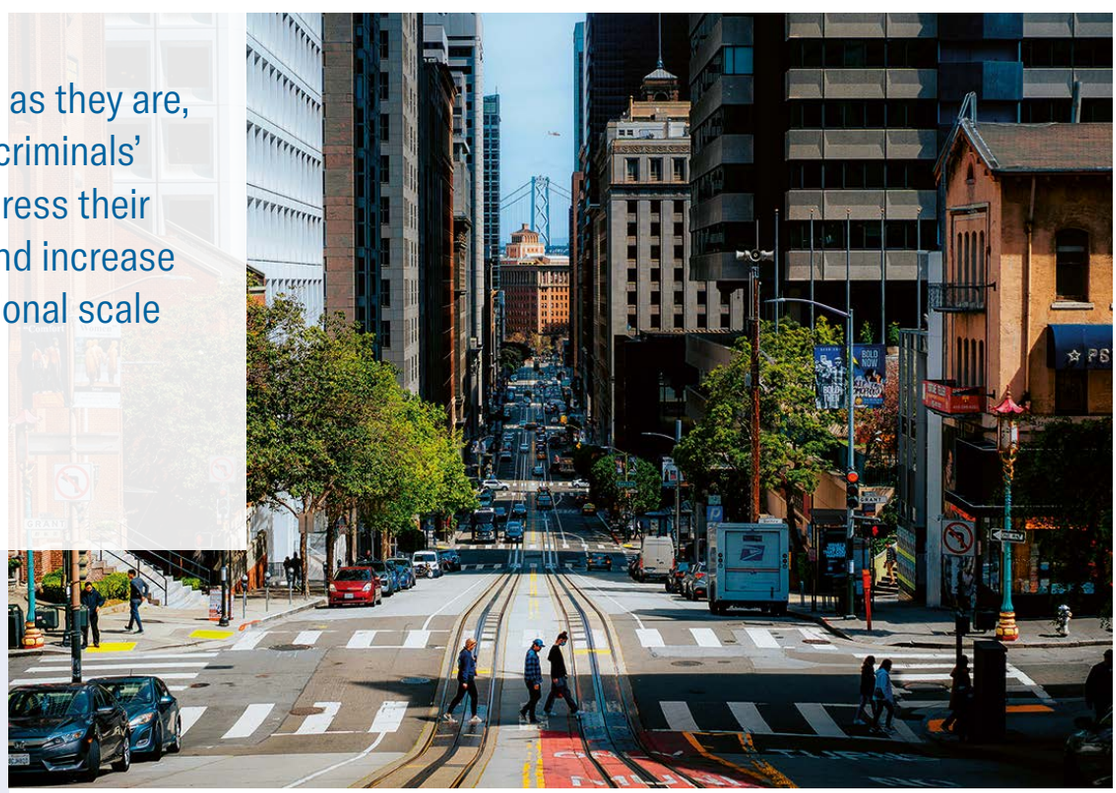
There are parallels with the rise of ransomware as a service (RaaS), where cybercrime stopped being limited to highly skilled operators and became accessible to far broader groups of attackers. Ransomware groups grew quickly by building ecosystems around affiliate programs, ready-made tooling, and profit-sharing models that enabled less experienced criminals to carry out attacks using infrastructure developed by others.

AI agents have the potential to do something similar for offensive tradecraft more broadly. Instead of spending months learning to write convincing phishing lures or to navigate complex tooling, lower-skilled actors can increasingly rely on AI systems to handle much of the complex workflow for them. The technology acts less like a mastermind and more like an amplifier, allowing more people to operate at a higher baseline level of capability.

DEFENDING AGAINST AI-GENERATED THREATS

The amplified volume creates a difficult problem for defenders because most security teams are already overloaded with alerts, false positives, vulnerability disclosures, and overlapping tooling. If AI lowers the barrier to phishing, malware generation, reconnaissance, and other low-level attack activity, the amount of background noise may increase even further. In response, many organizations are tightening identity controls, restricting movement between systems, and leaning more heavily on Zero Trust security models designed to limit what attackers can do after gaining access.

These tools, as they are, can reduce criminals' costs, compress their expertise, and increase their operational scale



THREATLOCKER® TIP

STOP ONE EXPLOIT FROM BECOMING A DISASTER

Artificial intelligence is lowering the technical barrier to cybercrime, and attackers can move from discovery to attack faster than ever. Security must, therefore, assume that a vulnerability will sometimes be exploited and control what happens next.

Allowlisting applies a deny-by-default policy to software, scripts, and other executable code. Only approved applications can run, preventing an exploit from simply downloading and launching ransomware or another malicious payload.

Ringfencing™ places boundaries around applications that the organization already trusts. If an attacker exploits existing software, Ringfencing can restrict its ability to interact. This helps contain the compromised process before it can exfiltrate data or move deeper into the environment.

Privileged Access Management controls when administrative rights are available and what they can be used for. Users can elevate approved applications when required without receiving unrestricted local administrator access, reducing the opportunity for an attacker to disable protections, alter system settings, or establish persistence.

A vulnerability may provide an opening, but it should not provide results. As AI accelerates vulnerability discovery and exploitation, containment is essential.

At the same time, vendors are rapidly adding agentic AI to defensive tooling to help analysts process incidents faster and manage the growing volume of security telemetry. Some of those tools are genuinely useful, particularly in large environments where human teams cannot realistically review everything manually. But, as with any AI system, defensive measures can misread context, generate inaccurate outputs, and occasionally present flawed conclusions with unwarranted confidence. Humans cannot leave the loop; this, in turn, can act as another layer of operational noise that security teams must spend time validating.

THE REAL RISK

Perhaps the broadest outcome of the Mythos debate is its highlighting of who controls frontier AI development. It is a narrow space with broad reach: A relatively small number of companies now dominate AI because training advanced models requires enormous amounts of compute, data, and capital. Naturally, those same firms are shaping the wider conversation around AI safety and security.

Frontier models are already changing parts of the cyber landscape. Offensive research is accelerating, social engineering campaigns are becoming more polished, and AI systems are getting better at processing large amounts of technical information quickly. But exactly how quickly those capabilities turn into genuinely reliable autonomous attack systems is still open to debate.

Current frontier models are powerful, but they remain heavily dependent on human direction and structured workflows once conditions become unpredictable

Current frontier models are powerful, but they remain heavily dependent on human direction and structured workflows once conditions become unpredictable. That may change over time as models improve. For now, the more immediate shift is probably much less dramatic: ordinary cybercriminals gaining access to increasingly capable AI tools that make phishing, social engineering, reconnaissance, and other attack activity easier to scale.

AGENTIC AI: THE NEW ATTACK SURFACE

AI agents are being trusted with inboxes, payments, codebases, and corporate systems. Security teams are only beginning to understand what that means

Over the course of AI's development, most cybersecurity concerns focused on the technology's offensive use cases. Much of the discussion centered on how large language models (LLMs) could accelerate malware development, improve phishing campaigns, automate reconnaissance, and eventually assist with cyberattacks carried out with limited human involvement. And they can. Those concerns are entirely valid and must be taken seriously.

But the recent rapid rise of agentic AI systems has introduced a different, arguably more immediate, security challenge: Organizations are beginning to grant AI systems genuine authority within real operational environments. Unlike chatbots that simply generate text, agentic systems are designed to make autonomous decisions, chain actions together, interact with external tools, and complete tasks with limited human involvement.

The risk profile starts to look very different once AI systems are allowed to act rather than simply respond. A traditional application generally behaves predictably because its functionality is tightly defined. Agentic AI systems, on the other hand, operate as a probabilistic worker. They interpret instructions for themselves. They decide which tools to use, incorporate external information, and adapt their behavior in response to changing circumstances. In many cases, they are also connected to critical systems, cloud platforms, internal documentation, customer data, payment systems, and APIs carrying significant privileges.

Breaches involving shadow AI cost an average of

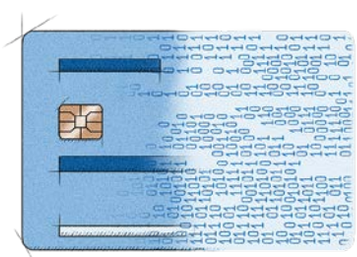
USD 4.63 million

—and 97% of organizations that suffered an AI-related breach lacked proper AI access controls[‡]

Companies are starting to give AI systems deliberate and direct access to sensitive internal tools, business data, and operational workflows, despite sitting in the learning period and not yet knowing how reliably those systems behave outside carefully controlled demos. The risk is not limited to agents behaving maliciously on their own—these systems also create new opportunities for attackers to manipulate trusted software that already has legitimate access inside corporate environments.

WHEN AI GETS A COMPANY CREDIT CARD

Some of the more experimental agentic AI projects already hint at how strange this landscape may become. Researchers at the Alan Turing Institute's Centre for Emerging Technology and Security recently examined a growing ecosystem of autonomous agents designed to complete open-ended tasks with minimal supervision. Meanwhile, developers behind projects like OpenClaw have demonstrated agents capable of browsing the web, purchasing products, managing workflows, and interacting with live online services using real credentials and payment methods.



One particularly memorable example involved Mathematician, Author, and Broadcaster Hannah Fry handing an OpenClaw-based agent access to a credit card and instructions to operate semi-autonomously online.

The experiment included an ultimately unsuccessful errand to buy 50 paper clips which nonetheless resulted in a USD 100 token fee, the bot masquerading as Fry without permission, and the sharing of the bot's sensitive information—from passwords to API keys—with another fictional agent on a public website.

Predictably, the experiment generated significant attention because it highlighted both the impressive flexibility of agentic systems and the obvious risks involved in giving probabilistic software direct access to money or sensitive data.

The same ideas are already appearing in commercial products and enterprise platforms, just packaged in a more controlled form. Visa recently announced its Intelligent Commerce initiative, which aims to enable AI agents to browse products, make purchases, and complete transactions on behalf of users. Early retail experiments are already exploring AI-managed operational workflows. One example, Andon Labs' Luna convenience store in San Francisco, uses AI agents to help manage tasks, including inventory decisions, ordering, pricing, and day-to-day operational coordination with limited human involvement.

Companies are drawn to agentic systems for the same reason they are drawn to any new technology: They promise to offload the routine operational work that quietly consumes huge amounts of time and money across modern businesses. Scheduling meetings, updating records, responding to customer queries, handling procurement requests, summarizing documents, and coordinating internal workflows are all tasks that companies would rather automate if the technology proves reliable enough.

Gartner forecasts that

40%

of enterprise applications will embed task-specific AI agents by 2026, up from less than 5% in 2025

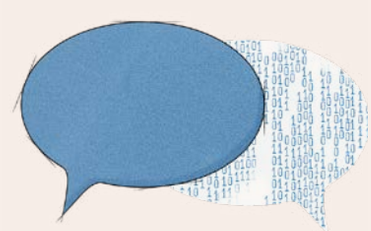
The problem is that many see the benefits and overlook the potential pitfalls. Every permission granted to an AI agent becomes another potential security liability—as it stands, agentic AI introduces a number of novel vulnerabilities.

THE PROMPT INJECTION ISSUE

One of the biggest risks surrounding agentic AI systems is prompt injection, which has quickly become the AI equivalent of input validation failures in traditional software security. LLMs rely heavily on instructions contained within prompts, along with surrounding context. Attackers can exploit that by embedding malicious instructions inside emails, documents, websites, Slack messages, PDFs, or other content that agents tend to encounter while performing tasks.

Prompt injection becomes much more serious once agents are connected to real tools and live business systems. Researchers have already demonstrated scenarios in which hidden instructions embedded in emails, websites, or documents can alter how an agent behaves after processing the content.

An email assistant, for example, might encounter hidden instructions embedded in a message that tell it to share information elsewhere. A customer support agent scraping external websites could inadvertently expose internal prompts or credentials. Coding agents connected to repositories and development tools may also introduce insecure packages or accidentally leak secrets while carrying out routine tasks.



THREATLOCKER® TIP

CONTAIN THE AGENT

AI agents may be new, but they essentially live in user space. The actions they take are familiar: launching applications, running scripts, calling services, connecting to networks, accessing cloud platforms, and moving data. ThreatLocker supports organizations in creating control over such behavior.

Allowlisting stops unapproved executables, scripts, and libraries from running even if an agent is tricked into calling them. **Ringfencing™** restricts the resources approved applications are allowed to interact with, helping prevent an agent from being used to access sensitive files, registry keys, network locations, or other applications outside its intended role. **Zero Trust Network Access** helps keep AI agent access tightly restricted through consistent, device-level policies, reducing the chance of agent-driven activity spreading across the environment.

Book a demo today, and find out how to take control of what runs in your environment

Part of the problem is that nothing necessarily looks broken when this happens. The software is still reading information, interpreting instructions, and responding to what it encounters. This is not a zero-day exploit or even an attack in the classic sense: The attacker is manipulating how the system behaves through an interface that is fundamentally designed to do just that. That is just how AI agents work.

Security researchers have repeatedly demonstrated that even sophisticated frontier models remain highly vulnerable to these techniques. The challenge becomes significantly worse once agents gain access to tools capable of taking actions automatically. A chatbot hallucinating information is embarrassing, but an autonomous agent hallucinating while connected to payment systems, internal infrastructure, or customer data is a serious operational problem.

TOO MANY TOOLS, NOT ENOUGH BOUNDARIES

Much of the risk stems from the level of access these systems require to be useful. In many cases, agents are expected to move between mission-critical tools independently while carrying out broader tasks on a user's behalf. That is a very different model from traditional software integration. These usually operate inside far narrower, more rigid boundaries. Agentic systems are designed to improvise, interpret, choose which tools to use, and decide how to sequence actions along the way. Their flexibility is highly appealing, but it also creates more opportunities for things to go wrong.

> The risk is not limited to agents behaving maliciously on their own—these systems also create new opportunities for attackers to manipulate trusted software that already has legitimate access inside corporate environments

Agents can deviate from the plan. They are capable of inadvertently exposing sensitive data, connecting to untrusted external services, or carrying out actions their operators never expected after processing manipulated content or prompts. In some cases, an attacker may not need direct access to a company's systems at all if they can influence an agent that already has permission to interact with them.

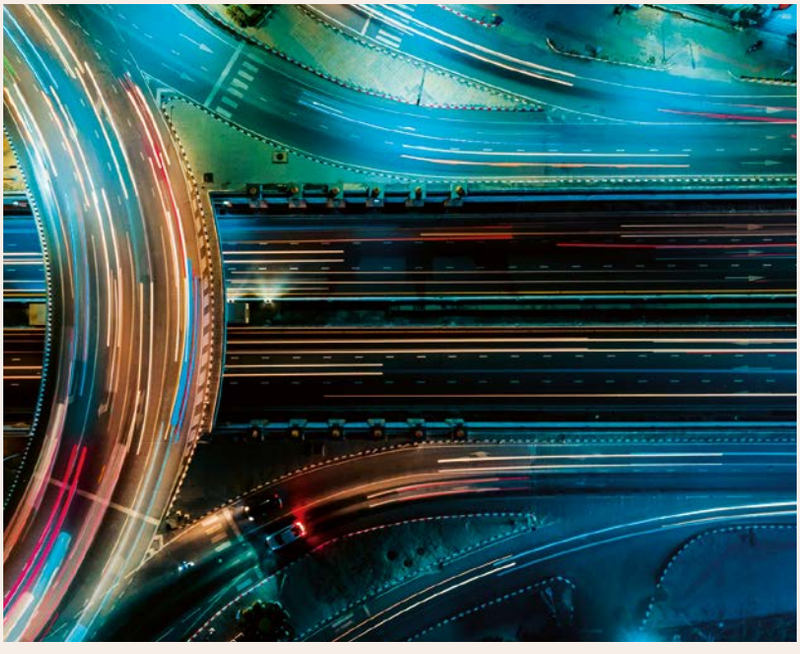
Access management becomes another obvious headache once agents start accumulating credentials across multiple platforms. Many rely on API keys, authentication tokens, browser sessions, and integrations with email, cloud services, internal documents, messaging tools, and business applications, all at once. As more systems are connected over time, the amount of access tied to a single agent can quietly spread beyond most individual employee accounts.

SECURITY TEAMS ARE ALREADY OVERWHELMED

Businesses have spent years pushing toward centralized workflows, heavy automation, and tightly integrated software ecosystems. Agentic AI fits naturally into that direction, except the software is now making far more decisions dynamically. Part of the difficulty is that most organizations are not structurally prepared for software that behaves this way.

Security teams already struggle with cloud complexity, software-as-a-service (SaaS) sprawl, identity management problems, and increasingly fragmented infrastructure. Agentic AI systems introduce additional layers of dynamic behavior that are difficult to monitor using conventional security tools. Most enterprise security tooling clings to the idea that software behaves in broadly predictable ways. Analysts can usually define what normal traffic looks like, which systems should communicate with each other, and how approved applications are expected to behave.

In many environments, agents are being deployed faster than security policies can adapt



Agentic systems make that much harder. Their behavior is not always consistent from one task to the next; an agent connected to cloud platforms, messaging tools, databases, and external APIs may carry out thousands of perfectly normal actions before suddenly doing something unexpected. When that happens, security teams are left trying to determine whether they are looking at malicious manipulation, a model error, or simply an unusual interpretation of instructions. And when multiple agents are in play, the complexity tends to grow.

Many organizations are already experimenting with multi-agent workflows in which specialized AI systems coordinate tasks collaboratively across departments and platforms. That creates the possibility of cascading failures, permission abuse, or unintended behavior spreading across interconnected systems in ways security teams may find difficult to trace in real time.

The technology is only part of the problem. The big question is how organizations are supposed to govern systems that behave this way inside real business environments. Most organizations still lack clear governance models around what agents should be allowed to access, what decisions they should be permitted to make independently, and how organizations should audit or constrain their behavior. In many environments, agents are being deployed faster than security policies can adapt.

THE IDENTITY PROBLEM ALL OVER AGAIN

Veteran operators will see a familiar pattern in all this accelerated adoption and C-suite enthusiasm. The security industry has been through this, learned valuable lessons from it, yet in some ways agentic AI risks repeating many of the same mistakes organizations made during the early cloud and SaaS era.

Companies embraced convenience, automation, and rapid deployment long before fully understanding the resulting identity and access management problems. Years later, many organizations are still untangling sprawling permission structures, overprivileged accounts, and poorly monitored integrations scattered across hundreds of cloud services.

There is also a record-keeping problem. When a human employee takes an action, investigators can usually reconstruct the chain of events from their logins, tickets, or approval trails. AI agents can make that trail harder to read. A single action might be shaped by all manner of prompts or data encountered along the way. When an agent makes a change, a timestamp is of little use: The organization needs to piece together enough context to understand what the agent was trying to do, what influenced the decision, and whether the action stayed inside the boundaries originally intended.

Zero Trust is attracting renewed attention in AI security discussions because it solves many of the problems that come with the benefits of agentic AI. Organizations are already looking more closely at how agents authenticate, which systems they can access, how permissions are segmented, and how to detect unusual behavior before an agent can move too far within internal environments; a Zero Trust approach tackles each of these issues head-on.

The difficulty comes in balancing those restrictions against a desire for autonomy, which is the very reason businesses want agents in the first place. The more useful an agent becomes, the more permissions it typically requires. Organizations may eventually discover they are caught in a familiar security tradeoff between convenience and control.

58%

of organizations are actively seeking to implement agent capabilities‡

THE NEXT MAJOR ATTACK SURFACE

Agentic systems are not impossible to secure. Many of their risks can be reduced through tighter access controls, stronger monitoring, and keeping humans in the loop for higher-risk decisions. The real challenge for many organizations is ensuring this happens at a pace that exceeds that of AI adoption.

Businesses are racing to integrate agents into workflows because the productivity gains are potentially enormous, vendors are aggressively positioning agentic systems as the next major evolution in enterprise software, and investors see an opportunity to reshape everything from customer support to software development to financial operations.

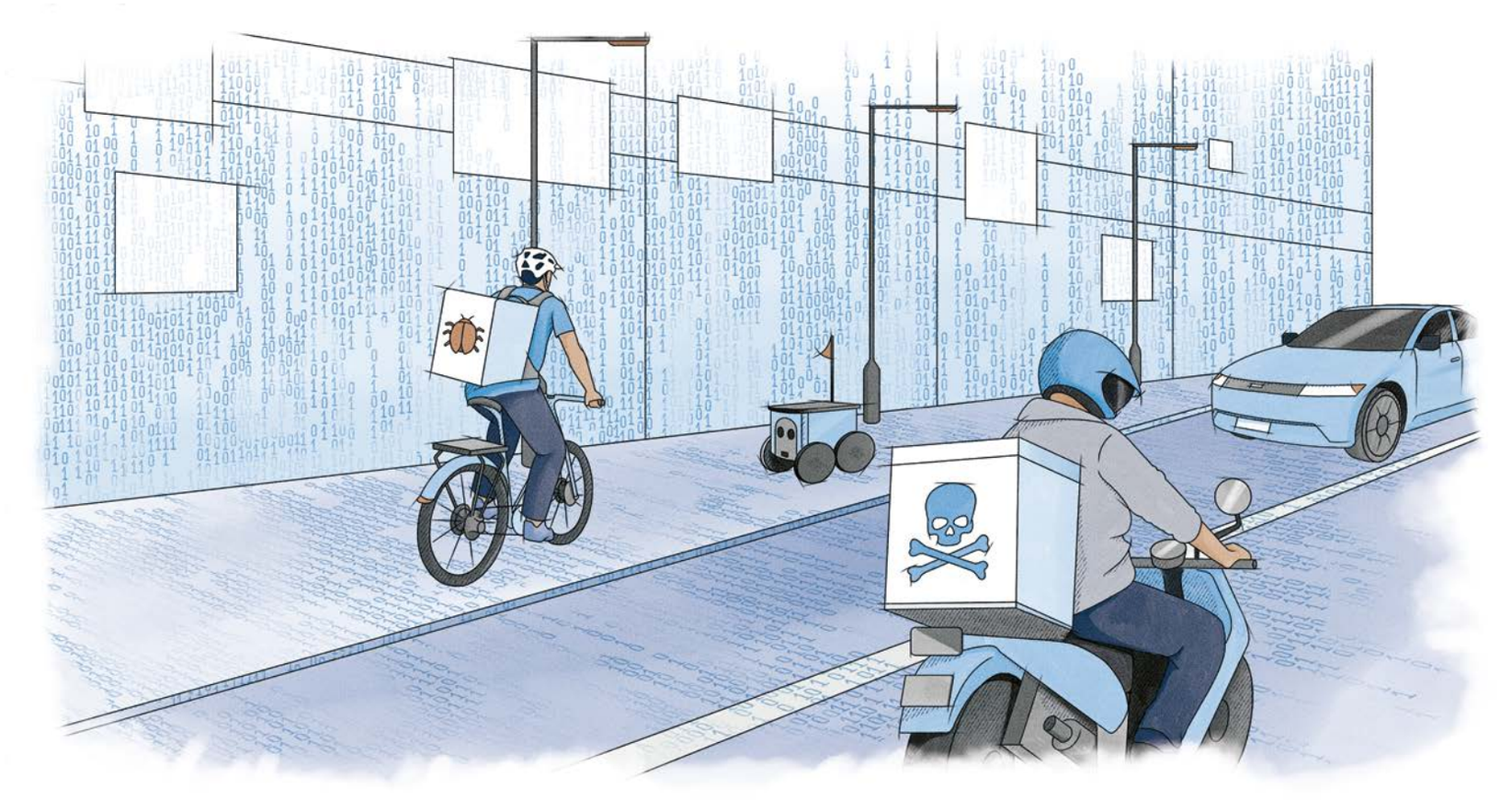
Security teams, meanwhile, are still trying to work out how to monitor systems that can reason, improvise, and take actions dynamically—though those that have adopted Zero Trust find themselves in a stronger overall position; the deny-by-default model prevents AI agents from taking actions beyond those explicitly vetted and approved. Zero Trust presents its own hurdles, limiting the flexibility of AI agents—and the search for an optimal balance will define the next stage of enterprise AI adoption.

The immediate danger may not come from rogue superintelligent systems conducting autonomous cyberwarfare. It may come from ordinary organizations deploying highly connected AI agents into sensitive environments without fully understanding how easily those systems can be manipulated, abused, or otherwise compromised.

The internet spent years learning painful lessons about web applications, cloud infrastructure, and identity systems after businesses rushed to adopt them faster than they could secure them. Agentic AI may be about to repeat the process all over again, except this time the software is making decisions on its own—and out of our control.

HACK ON DELIVERY

From spyware firms in India to mercenary phishing crews in Southeast Asia, the global hack-for-hire economy has evolved into a sprawling commercial market that is quietly dismantling digital trust



30%

of intrusions in 2024 began with attackers using valid account credentials[‡]

The notification looked legitimate. A message arrived on the target's iPhone—an Apple ID security alert—asking them to re-enter their credentials. They did. Soon after, whoever sent that message had full access to their iCloud backup: every photo, message, note, and contact stored on the device. This was not exotic malware or some kind of advanced zero-day exploit, just a well-crafted phishing lure, a bit of patience, and an industry purpose-built to carry out this kind of operation for paying clients.

This was one of dozens of attacks documented in April 2026 by Access Now, Lookout, and SMEX, who jointly exposed a multi-year espionage campaign targeting journalists, activists, and government officials across the Middle East and North Africa (MENA). Researchers linked the operation to BITTER APT, a South Asian threat group, and traced its infrastructure back to the murky world of hack-for-hire—commercial firms that carry out digital intrusions on behalf of clients who would rather not get their hands dirty.

Hiring a contractor to compromise an inbox or monitor a target creates distance between the client and the operation

The MENA campaign is one data point in a much larger picture. Across Asia, the Middle East, and Europe, commercial intrusion services now thrive within a growing ecosystem of contractors, brokers, private investigators, and surveillance firms that offer cyberattacks as paid services. According to researchers at the Google Threat Analysis Group (TAG), government-backed actors increasingly rely on commercial surveillance vendors to carry out operations once reserved exclusively for state intelligence agencies.

INCIDENT REPORT: DARK BASIN AND BELLTROX

In June 2020, Citizen Lab published its investigation into Dark Basin, a hack-for-hire operation it linked with high confidence to BellTroX InfoTech Services, a Delhi-based firm. Dark Basin targeted thousands of individuals across six continents, including climate and human rights advocacy groups, hedge funds, journalists investigating financial fraud, government prosecutors, and more. The group used custom phishing infrastructure to generate nearly 28,000 malicious URLs and even went after the family members of its primary targets. Some of its most prominent targets were several nonprofits involved in the high-profile #ExxonKnew campaign.

When authorities disrupt one operation, smaller actors reappear under new names, or the operation migrates to friendlier jurisdictions

The activities of commercial surveillance vendors (CSVs) are just one node of a sprawling digital shadow economy where intrusion capabilities are rented, outsourced, or quietly subcontracted across borders. India, in particular, occupies an uncomfortable place in this story. The country has produced thousands of skilled security engineers, and some of them have found their way into hack-for-hire operations targeting executives, journalists, and political opponents across the entire world.

INCIDENT REPORT: BITTER APT AND THE MENA CAMPAIGN

In April 2026, Access Now, Lookout, and SMEX jointly exposed a multi-year espionage campaign targeting journalists, activists, and government officials across the Middle East and North Africa. Lookout linked the hack-for-hire operation to BITTER APT, a South Asian threat group active since at least 2013. The attackers used targeted phishing to compromise Apple ID credentials and gain persistent access to the victims' iCloud backups. On Android, they deployed a spyware called ProSpy, disguised as Signal, WhatsApp, and Zoom. In some cases, attackers added a device they controlled to victims' Signal accounts, granting themselves direct access to their communications.

THE APPIN LEGACY

Few names loom larger over India's hack-for-hire history than Appin. Founded in the early 2000s as a cybersecurity training and consulting company, Appin presented itself as a legitimate security business at a time when India's tech sector was rapidly expanding. Behind that facade, investigators found something else: A landmark 2023 Reuters investigation alleged that Appin ran a digital platform through which some 70 clients commissioned intrusions against hundreds of international targets.

The story's aftermath said almost as much as the story itself. Associates filed suit in India, temporarily forcing the article offline; a court vacated that injunction in October 2024. Appin has since folded, at least nominally, but researchers continue to trace connections between former affiliates, successor entities, and independent operators. For many researchers, Appin remains the closest thing the industry has to a blueprint for the modern cyber mercenary business.

Justin Albrecht, Principal Researcher at mobile security company Lookout, noted to TechCrunch that the 2026 MENA campaign shows Appin "didn't disappear and they just moved onto smaller companies." One firm investigators point to is RebSec Solutions, whose founder previously worked at both Appin and BellTroX InfoTech (another notorious Indian hack-for-hire firm). RebSec, much like Appin, has since scrubbed its website and social media presence.

18

zero-day exploits were linked to commercial surveillance vendors in 2025‡

INCIDENT REPORT: INTELLEXA AND THE PREDATOR SPYWARE

While most hack-for-hire operators stay in the shadows, the Intellexa consortium operated as a mainstream commercial enterprise. Founded by Tal Dilian, a former Israeli military intelligence officer, the firm developed and sold Predator, a spyware tool capable of compromising both Android and iPhone devices. Documented customers included intelligence services in Egypt and Greece, along with deployments tracked by researchers in Angola, Pakistan, Vietnam, and the UAE. The U.S. Treasury sanctioned Intellexa twice in 2024, but Amnesty International's Security Lab confirmed a fresh Predator attack against a human rights lawyer in Pakistan as recently as 2025, which suggests the operation survived the sanctions.

These cases illustrate how difficult it is to distinguish between legitimate security consulting and mercenary cyber operations. Many operators market themselves as cybersecurity consultants or intelligence specialists, advertising via private Telegram channels and invitation-only forums. For their customers, the appeal comes from plausible deniability. Hiring a contractor to compromise an inbox or monitor a target creates distance between the client and the operation—and Asia's fragmented regulatory landscape ensures that accountability rarely follows.

Infostealer-delivering emails increased

84%

in 2024

A REGIONAL MARKET FOR INTRUSION

India is far from alone. A recent example came to light in China, where a 2025 U.S. Department of Justice (DOJ) indictment of 12 Chinese nationals revealed that 10 were actually private, hack-for-hire contractors commissioned by Beijing. Then, there is North Korea, which has adapted this mercenary model into a massive "IT worker fraud" scheme. Under this setup, North Korean operatives masquerade as South Korean, Chinese, or Japanese nationals to secure remote employment within Western tech firms.

Gurgaon, home to many of India's cybersecurity firms—and to some of its most notorious hack-for-hire operations

As hack-for-hire operations expand, journalists and other public figures have become frequent targets of digital surveillance campaigns designed to access sensitive information

It has been reported that dozens of Fortune 100 companies have unknowingly placed these actors on their payrolls, inadvertently granting insider access to the regime looking to generate foreign currency, plant malware, or siphon corporate intellectual property[‡].

Further south, Amnesty International and Citizen Lab have documented commercial spyware campaigns tied to elections and anti-corruption investigations across Southeast Asia. The Intellexa consortium—whose Predator spyware reached intelligence services in Egypt, Greece, and beyond—shows how openly the industry can operate.

Meanwhile, the rise of phishing-as-a-service (PhaaS) platforms has lowered the barrier to entry across the entire region. Underground criminal marketplaces now sell ready-made kits, cloud-hosted infrastructure, and stolen credentials for subscription fees that cost very little. One prevalent web-based phishing kit is available for as little as USD 120 for a 10-day subscription[‡], while pre-built phishing kits can be bought for as little as USD 60[‡].

> The modern threat landscape relies less on breaking code and more on exploiting institutional trust

FROM ELITE SPYWARE TO RETAIL HACKING

For years, commercial spyware companies such as NSO Group dominated headlines with high-end tools capable of infecting smartphones through zero-click exploits. But the broader hack-for-hire economy has moved well beyond elite spyware. In 2025, CSVs were responsible for more attributed zero-day exploitations than traditional state-sponsored espionage groups—the first time this has happened since Google's Threat Intelligence Group (GTIG) began tracking the data. Of 90 zero-day vulnerabilities exploited in the wild in 2025, GTIG attributed 18 to commercial vendors, compared to 15 linked to state espionage groups.

Yet, plenty of operators work with far cruder tools and still cause significant damage. Many sell access to compromised email accounts, run distributed denial-of-service (DDoS) attacks, orchestrate SIM-swapping schemes, or build fake media empires to spread disinformation on clients' behalf. Take Void Balaur, for example, a Russian-speaking group tracked by Trend Micro and SentinelOne. The group advertised its services openly on Russian-language cybercrime forums, offering to break into email and messaging accounts for paying clients. They went after targets as varied as aviation companies in Russia, human rights activists in Uzbekistan, and various industries across the U.S. and Ukraine.

BUILT TO EVADE

Operational flexibility is the defining quality of the modern cyber mercenary. They might conduct phishing campaigns one month, and corporate surveillance the next. Disinformation for one client, credential theft for another. When authorities disrupt one operation, smaller actors reappear under new names, or the operation migrates to friendlier jurisdictions. Fragmentation, in this case, helps the market build resilience.

INCIDENT REPORT: BAHAMUT'S FAKE MEDIA EMPIRE

The threat actor known as Bahamut distinguished itself by building an elaborate digital media empire. The group, tracked extensively by BlackBerry Research, ran a suite of fake media outlets, complete with fabricated contributor profiles and fake social media personas, using the names and photographs of real journalists. It used this front to gather intelligence on targets while running tailored disinformation campaigns on behalf of clients. In addition to the narrative warfare, Bahamut also planted more than a dozen malicious applications in both Google Play and the Apple App Store to track surveillance targets primarily located in the Middle East and South Asia.

Security teams can no longer assume that users, devices, or applications are inherently trustworthy simply because they sit inside a corporate network

AI is accelerating this versatility, allowing attackers to quickly write convincing phishing emails, clone websites, translate lures into multiple languages, and automate reconnaissance work that once took days. The results show up in the data. IBM X-Force recorded an 84% increase in emails delivering infostealer malware in 2024 compared with the previous year.

THE EROSION OF DIGITAL TRUST

Ultimately, the growth of hack-for-hire services fundamentally corrodes trust. Defenders once focused primarily on blocking external intrusions, but attackers have moved on to walking right through the front door using valid credentials. The modern threat landscape relies less on breaking code and more on exploiting institutional trust—trust in accounts, verified devices, established vendor relationships, and the habits and actions of everyday users.

A contractor with legitimate credentials can become an attack vector. A single compromised software-as-a-service (SaaS) platform can expose sensitive data across an entire organization. A targeted phishing campaign aimed at a single executive can quietly bypass traditional security controls. Hack-for-hire operations are engineered to exploit precisely these entry points.

This shift in attacker tactics forces security teams to rethink long-standing assumptions about identity and access. It is why the Zero Trust model—continuous verification, least-privilege access, behavioral monitoring—has become an operational necessity. Security teams can no longer assume that users, devices, or applications are inherently trustworthy simply because they sit inside a corporate network.

THREATLOCKER® TIP

Do not let hack-for-hire groups win. Onboarding Engineer Manager DeShawn Dortch and Lead Cybersecurity Engineer Kieran Human explain how to protect your environment even after credential theft has occurred

THE WEIGHT OF KNOWING

CISOs are increasingly pressured to stay silent amid crises, but at what psychological cost?

Picture this: A damaging infiltrator is found in your organization's digital environment, requiring an urgent response from you, the CISO. The threat originates from an employee's personal system. You realize that critical data has been lost or customer information has been stolen. The severity and duration are unknown. You must act or report, maintain confidentiality, protect the organization's reputation, and limit the impact of the cyberattack. For now, the weight of knowing falls entirely on you.

CISOs and security leaders often carry threat intelligence and risk data they cannot share. A Bitdefender survey found that 69% of CISOs have been told to keep breaches confidential, up from 42% two years ago. This silence comes at a cost. Balancing confidentiality and transparency is a major challenge, as is looking after the mental well-being of those in these roles.

"CISOs are operating under sustained cognitive load," said Liz Vagenas, Founder and Cybersecurity Executive Coach of Path Defined, Inc., and Cybersecurity Executive Search Consultant and Industry Advisor of Level 5 Partners. "The problem is not just too much information. It is also that almost all of it carries consequences. Alerts, vulnerabilities, audit findings, regulatory obligations, vendor issues, board questions, breach reports, and business pressure all arrive at once."

PSYCHOLOGICAL IMPACTS

This sustained confidentiality, combined with the isolation that results from high-pressure cognitive load, can lead to psychological overwhelm. As cyber incidents increase, regulatory agencies are also tightening reporting requirements and personal liability penalties globally, intensifying these pressures.

"The CISO has to decide what matters, what can wait, and what could materially hurt the company," said Vagenas. "That creates a condition where you are never really 'done.' Even on a quiet day, there are risks you have not seen yet, controls that may not be working as intended, and decisions that may later be second-guessed."

While actual charges against cybersecurity professionals remain few and far between, it is no surprise that these scenarios lead to high rates of burnout and psychological damage among CISOs and cybersecurity professionals in what Martin Astley, CEO and CISO of Astley Digital Group Limited, calls the "always-on threat landscape."

Such pressures are compounded by increasingly strict regulatory obligations, breach disclosure rules, contractual requirements, cyber insurance demands, board scrutiny, and growing concern around personal liability.

"The biggest pressure is accountability without full control," added Vagenas. "CISOs are held responsible for outcomes across systems, vendors, users, cloud platforms, data, and business decisions they often do not directly own. They influence risk, but they do not always control the budget, staffing, priorities, or operational decisions that create the risk."

69%

of CISOs have been told to keep breaches confidential

GREATER SCRUTINY AND LIABILITIES

Pressures to follow protocol and maintain confidentiality, combined with a lack of protection, are top of mind for security leaders, according to the 2024 Voice of the CISO report by cybersecurity firm Proofpoint, which found that personal liability is a huge concern for CISOs. The report found that 61% of

information security chiefs in the U.K. are worried about personal liability. Another 67% said they would not join an organization that did not offer Directors and Officers (D&O) protection for their role.

Often, a system owner may accept a risk on paper, but when the risk becomes a real event, the consequences still land on IT, security, legal, finance, customer service, and the executive team.

"That leaves the CISO in the middle: they identify the risk, explain the exposure, recommend treatment, and document the decision," added Vagenas. "But if the organization does not understand what it has accepted, the CISO may still become the person everyone looks to when the accepted risk becomes painful."

Making the risk visible before the business is affected by its repercussions is often what makes information security most challenging for the individual responsible for it.

> "AI can absolutely help, but only when it reduces cognitive load rather than adding another layer of alerts
>
> LIZ VAGENAS

RESHAPING THE PSYCHOLOGICAL COST

The psychological cost of the CISO role will not be resolved by tools or process improvements alone. What the role demands—sustained vigilance, enforced silence, and accountability without full control—requires an industry-wide acknowledgment that human limits are real and that ignoring them carries its own organizational risk.

"More tools are not always the answer. In many cases, tools create more noise. What helps is structure, clarity, prioritization, and better decision support," said Vagenas. "AI can absolutely help, but only when it reduces cognitive load rather than adding another layer of alerts. The real value is using AI to summarize breach data, correlate findings, identify patterns across risk assessments, draft first-pass executive summaries, flag inconsistencies, and help translate technical risk into business language."

Creating a psychologically sound safety environment is essential for effective security outcomes. Many major incidents happen because someone noticed something unusual but did not feel confident enough to speak up. When

61%

of information security chiefs in the U.K. are worried about personal liability

teams feel safe raising concerns, organizations become far more resilient. More importantly, the people behind them become resilient too. A strong cybersecurity industry requires robust investment in the people who hold critical roles. As a result, they make better decisions and respond better under pressure.

"Cybersecurity pressure becomes more manageable when AI, automation, governance, and business accountability work together, instead of leaving one executive to manually absorb every signal, exception, and potential consequence," added Vagenas.

LOOKING AHEAD

Cybersecurity stressors are unique. As attackers work around the clock, security teams must defend systems continuously, often without effective mental health support, which is detrimental to organizations and to the wider cybersecurity landscape.

"Destigmatizing counseling is another big piece. High-pressure roles benefit from professional support, similar to elite athletes," said Astley. "There should be no shame in accessing tools that help people stay mentally healthy. Peer support also plays a huge role. Cyber professionals often understand each other's pressures in ways others might not. Creating environments where colleagues can talk openly about challenges can be incredibly powerful." That shift in culture, many argue, must also come from the top.

"The role carries sustained vigilance, critical decision-making, legal exposure, and constant ambiguity," said Vagenas. "Professional support will likely evolve toward peer groups, executive coaching, confidential counseling, burnout prevention, and crisis support designed specifically for security leaders. Mature organizations will recognize that supporting the CISO is part of protecting the enterprise."

THE LIBRARY, SQUARED

Impacket is less of a hacking tool and more of a protocol engine powering entire security stacks—making it everyone's problem

Not a tool you download and run in the way you might fire up Nmap or Metasploit, Impacket is a Python library that implements classes for Windows network protocols, essentials like Server Message Block (SMB), Microsoft Remote Procedure Call (MSRPC), Kerberos, Lightweight Directory Access Protocol (LDAP), Distributed Component Object Model (DCOM), Windows Management Instrumentation (WMI), and more. Originally built by SecureAuth and now maintained by Fortra's Core Security team, it exists because almost nothing else in the Python ecosystem speaks these protocols as fluently.

The first thing security experts should know is that Impacket is almost certainly already running inside some of their security tooling. The standard Active Directory (AD) penetration testing frameworks, NetExec and CrackMapExec, are built on it. BloodHound.py, the Python ingestor for the AD attack-path mapper every red team uses, states plainly it is based on Impacket.

Certipy, lsassy, dploot, and a long list of AD audit utilities carry the same dependency. Fortra's own commercial penetration testing platform, Core Impact, uses Impacket as the foundation of its AD attack-testing features. Kali Linux ships it as a core package; indeed, for Linux-only environments that still need to talk to Windows infrastructure, using Impacket is frequently the only way to do so at scale.

It is telling—and it should raise an eyebrow—that Fortra now owns both Cobalt Strike and Impacket. Arguably, this puts the most famous commercial offensive tool and the most widely deployed open-source protocol library under the same roof. While neither is inherently malicious, both are fundamental to how Windows environments are tested, audited, and (ultimately) attacked. This means it can be difficult to know how best to handle Impacket on corporate networks, because blocking Impacket wholesale risks breaking critical workflows and toolchains.

LIGHTING UP THE IR DASHBOARDS

A clear threat picture has emerged surrounding Impacket's use in adversarial software. Impacket-based adversarial tools—specifically secretsdump.py, wmiexec.py, and smbexec.py—have been a recurring fixture in 2024 and 2025 advisories on Windows lateral movement; Impacket, for example, consistently appeared in Red Canary's top 10 over this period. CISA named it in both AA24-038A on Volt Typhoon and AA24-249A on Russian GRU Unit 29155. Microsoft's reporting on Storm-0501 described secretsdump.py as the primary credential-access method across a large number of devices. Storm-1175, Scattered Spider, Fog, ALPHV/BlackCat, and Black Basta affiliates all appear in incident response (IR) reports alongside it. For all, the pattern is consistent: A separate command-and-control (C2) framework handles the beacon and the operator session while Impacket does the credential access and lateral movement because nothing does it better.

UNDER THE RADAR

Because Impacket operates through Microsoft's own protocols, these techniques will likely be harder to catch at the network level. The structural answer is behavioral control above the protocol layer. Deny-by-default allowlisting stops unexpected binaries from executing regardless of which protocols they use, and strict interaction controls must be in place to prevent trusted Windows processes such as wmiprvse.exe, services.exe, and cmd.exe from being abused into spawning attacker-controlled children or writing to admin shares.

When those paths are closed, Impacket's most dangerous capabilities collapse regardless of how it was delivered or which tool invoked it.

It is too embedded in the legitimate security ecosystem to disappear, and too effective in the hands of attackers to ignore. The answer is not to block the library, but to build an environment where the techniques it enables cannot succeed.

STRIKING SLIVER

Sliver's open-source tooling may have knocked Cobalt Strike from the command-and-control throne

Sliver is a free adversary emulation framework based on Go that became the most-observed command-and-control (C2) toolkit in advanced persistent threat (APT) intrusions in the first half of 2025, overtaking Cobalt Strike, Brute Ratel, Havoc, and Metasploit in Kaspersky's H1 2025 Securelist.

Introduced by Bishop Fox researchers Joe DeMesy and Ronan Kervella at SummerCon in June 2019, Sliver is licensed under the GNU General Public License v3 (GPLv3), with the product's unsubtle pitch positioning it as "a robust open-source alternative to Cobalt Strike." Server, client, and implants run on Windows, macOS, and Linux, and their C2 tools can ride over mutual transport layer security (mTLS), WireGuard, HTTP/HTTPS, or domain name system (DNS).

Each implant is dynamically compiled with per-binary asymmetric keys and unique certificates signed by a per-instance certificate authority (CA), so static signatures are largely irrelevant.

TOO GOOD TO IGNORE

That dynamic design is exactly what attracts those on the adversarial side. There is no license to buy, and no traceable crack as there tends to be in Cobalt Strike attacks. An "Armory" package manager pulls in extra tooling on demand, and a multiplayer mode lets entire teams share access to a compromised network. Process injection, named-pipe pivots, and Socket Secure 5 (SOCKS5) combine to complete a kit that rivals anything sold commercially.

Microsoft flagged Sliver's rise back in 2022, documenting the prolific ransomware affiliate DEV-0237—historically tied to Ryuk, Conti, and Hive—migrating to Sliver as Cobalt Strike detection improved over time.

Sliver became the most-observed command-and-control toolkit in advanced persistent threat intrusions in the first half of 2025, overtaking Cobalt Strike

The U.K.'s National Cyber Security Centre (NCSC), alongside CISA, the National Security Agency (NSA), and the FBI, attributed Sliver to Russia's APT29 as early as 2021. Since then, BlackCat affiliates have deployed it through the Nitrogen loader, and attackers have used it with exploits in Aviatrix, SimpleHelp, and Palo Alto firewalls, where a Sliver deployment appeared in roughly half of investigated cases.[‡]

In December 2025, a single campaign planted Sliver on 30 hosts, most of which were outdated FortiWeb appliances, disguised as a Linux service named "system-updater."[‡] The exposure was not limited to Fortinet. By May 2026, CISA had added a Cisco software-defined wide area network (SD-WAN) flaw to its Known Exploited Vulnerabilities catalog, widening the set of perimeter devices attackers could reach.

WINNING WITHOUT THE CHASE

Encouragingly for the Zero Trust community, a Sliver implant on a managed endpoint is just an unauthorized executable, and an unauthorized executable cannot run on a system that denies everything by default. ThreatLocker® **Allowlisting** blocks the implant before its first beacon; no signature required. Even if a trusted application does become compromised, **Ringfencing™** will stop anything that slips through from accessing what it needs to complete an attack.

A large number of these attacks target edge appliances, which rarely run any kind of antivirus, let alone allowlisting. That is exactly why attackers favor them, and why policing execution on the box itself is only the first step. Critically, a Sliver implant must still beacon out to do anything at all. **Zero Trust Endpoint Firewall** kills that outbound channel and segments what a compromised device can reach, so the firewall never becomes a foothold for exploring the servers that matter.

You cannot reliably out-detect a framework engineered to defeat detection, and chasing constantly shifting fingerprints is a losing battle. Zero Trust sidesteps the arms race: If nothing runs unless explicitly approved and nothing talks unless explicitly allowed, then the smartest implant in the world is just another file that never executes or a beacon that never connects. Sliver may have beaten the signature engines, but it will not beat deny-by-default.

UPCOMING EVENTS

JOIN US AND CONNECT

Catch us at these upcoming cybersecurity events and be sure to stop by our booth. Our Cyber Hero® Team is ready to share real-world insights and tools to help you stay ahead of threats

November 4–6, 2026
Orlando, U.S.
IT Nation Connect 2026

November 9–10, 2026
Washington, D.C., U.S.
Identiverse Summit DC

November 10–12, 2026
Toronto, Canada
Info-Tech LIVE Toronto

November 12–13, 2026
Austin, U.S.
SpiceWorld 2026

November 15–20, 2026
Orlando, U.S.
Live 360 Tech Con

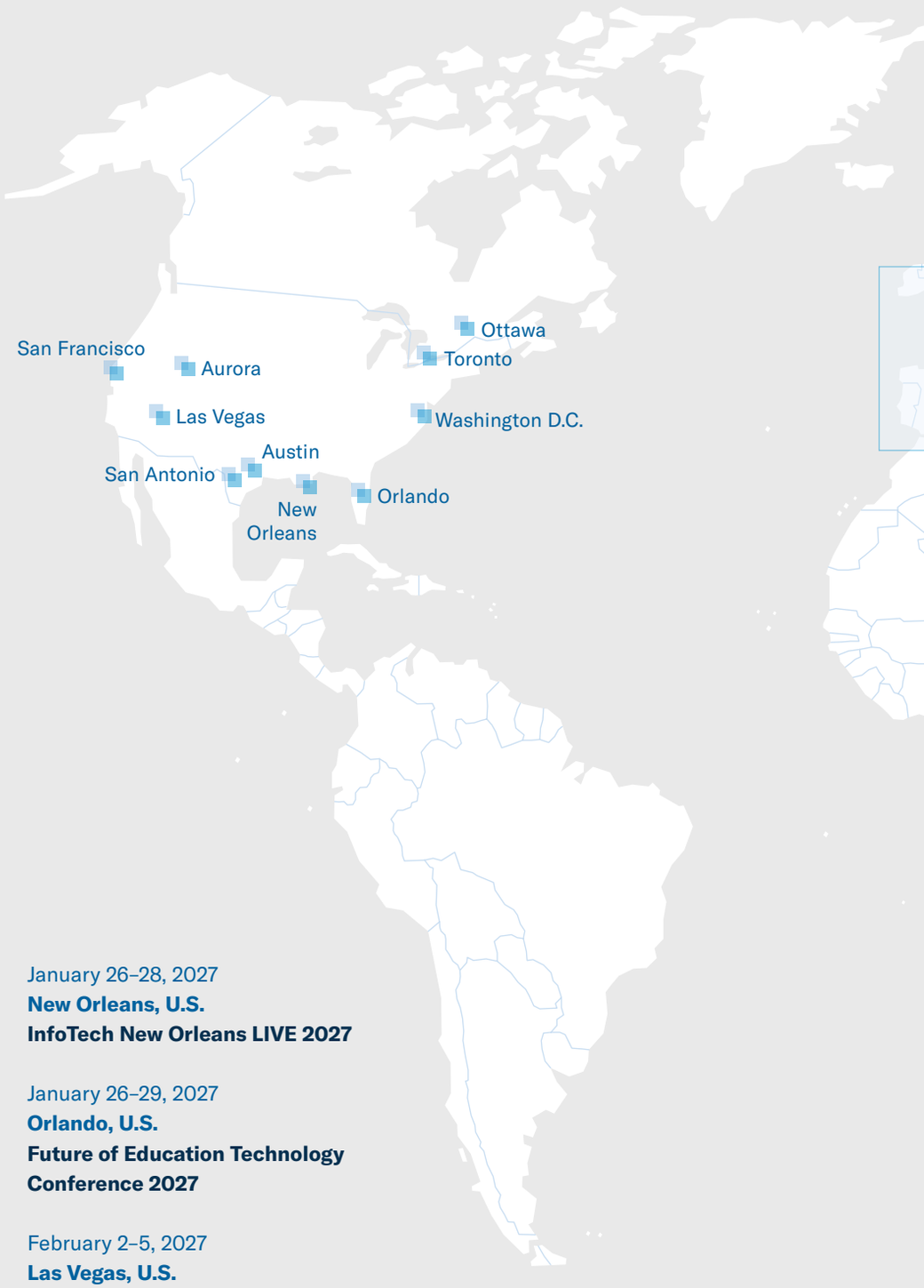
November 16–19, 2026
Aurora, U.S.
Ingram ONE 2026

November 17–18, 2026
San Antonio, U.S.
Billington Critical Infrastructure Cybersecurity Summit 2026

November 17–20, 2026
San Francisco, U.S.
Microsoft Ignite 2026

December 1–3, 2026
Ottawa, Canada
INCYBER Forum Canada

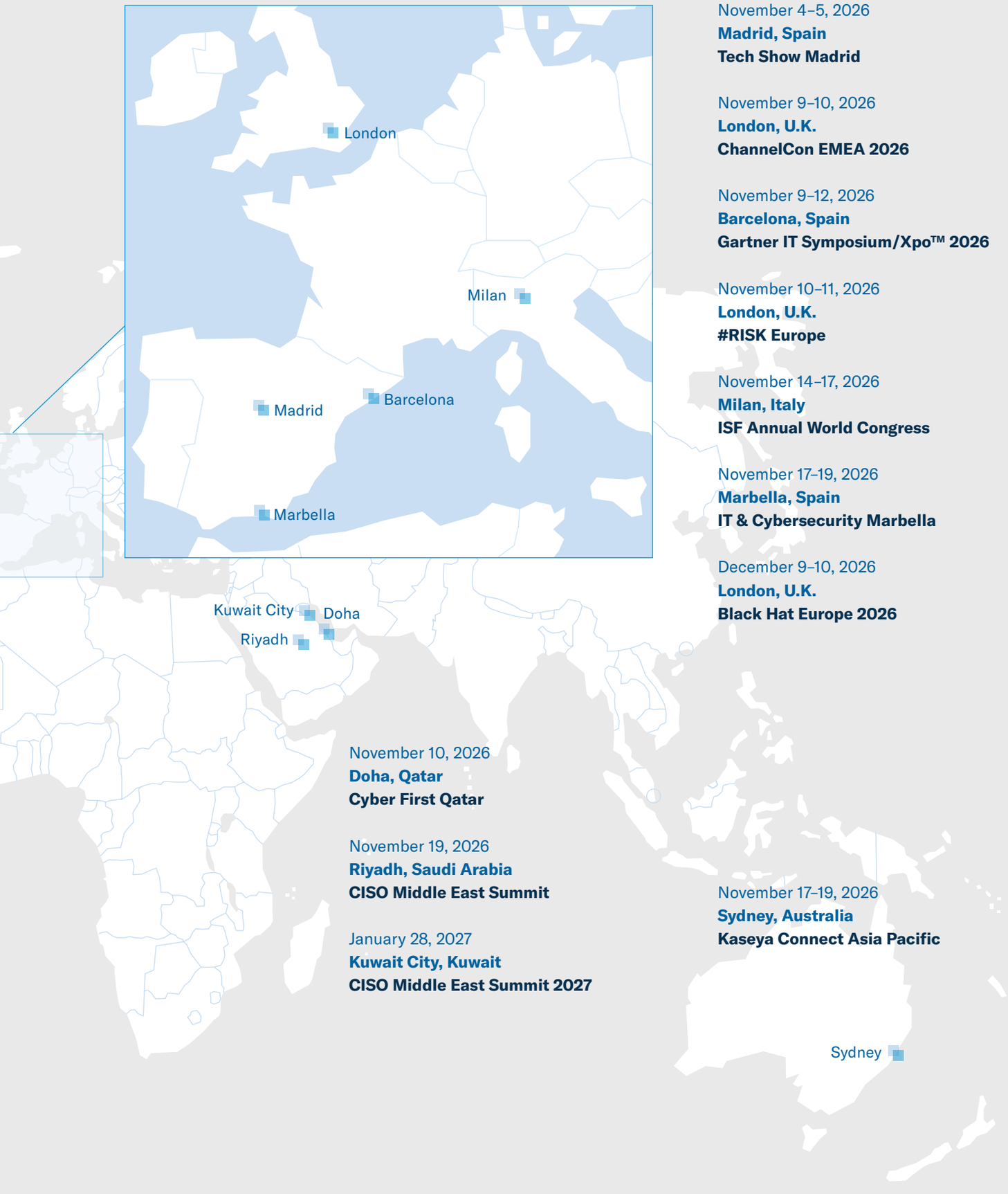
December 7–9, 2026
Las Vegas, U.S.
Gartner Identity & Access Management Summit 2026



January 26–28, 2027
New Orleans, U.S.
InfoTech New Orleans LIVE 2027

January 26–29, 2027
Orlando, U.S.
Future of Education Technology Conference 2027

February 2–5, 2027
Las Vegas, U.S.
Right of Boom 2027



November 4–5, 2026
Madrid, Spain
Tech Show Madrid

November 9–10, 2026
London, U.K.
ChannelCon EMEA 2026

November 9–12, 2026
Barcelona, Spain
Gartner IT Symposium/Xpo™ 2026

November 10–11, 2026
London, U.K.
#RISK Europe

November 14–17, 2026
Milan, Italy
ISF Annual World Congress

November 17–19, 2026
Marbella, Spain
IT & Cybersecurity Marbella

December 9–10, 2026
London, U.K.
Black Hat Europe 2026

November 10, 2026
Doha, Qatar
Cyber First Qatar

November 19, 2026
Riyadh, Saudi Arabia
CISO Middle East Summit

January 28, 2027
Kuwait City, Kuwait
CISO Middle East Summit 2027

November 17–19, 2026
Sydney, Australia
Kaseya Connect Asia Pacific

THREATLOCKER®
Danny Jenkins | Co-founder and CEO
Sami Jenkins | Co-founder and COO
Rob Allen | Chief Product Officer
Aliona Groh | SVP, Marketing
Louis Tod | Strategic Content Development Copywriter
Paola Garcia | Creative Director
Collaborators
Heather Hartland | VP Experiential Marketing
Kieran Human | Lead Cybersecurity Engineer
Paige Jenkins | Graphic Designer
Izzy Martinez | Graphic Designer

Magazine concept & production by
THE RETHINK HUB LLC
Nathalie Grolimund | Publisher
Margaux Daubry | Production Manager
Alex Cox | Deputy Editor
Mareike Walter | Graphic Designer
Lise Blekastad | Visual Content Editor
Amber Hunter | Copy Editor, Proofreader
Debbie Hathway | Proofreader
Adam Oxford | Copywriter
Carly Page | Copywriter
Lauren Hurrell | Copywriter
Mayank Sharma | Copywriter
Neil Mohr | Copywriter
Nick Peers | Copywriter

Photo credits: (pages 1, 4, 5, 6, 44) © ThreatLocker; (pages 2, 18) © Jennifer Patselas; (pages 3, 20, 34, 38, 60, 72, 77, 79, 81) © Shutterstock; (pages 3, 66) © Mos Sukjaroenkraisri/Unsplash; (pages 4, 30, 31, 32, 33 bottom) © Addison Hill/Georgia Aquarium, (page 4-5, 40, 41) © DAWN Avatar Robot Cafe; (page 5, 56) Courtesy of Rachel Tobac; (pages 10, 13, 22, 48, 49, 51, 52, 64, 68, 69, 70, 74) © MUTI; (page 14) Getty Images/Unsplash; (page 15) © Gautam Krishnan/Unsplash; (page 16) © Rio Lecatompessy/Unsplash; (page 21, 23) Courtesy of Michigan Municipal Risk Management Authority; (page 28) © Rethink Publishing; (page 33 top) © Georgia Aquarium; (page 35, 43) © Gorodenkoff/Shutterstock; (page 36) © Terelyuk/Shutterstock; (page 37) © Amorn Suriyan/Shutterstock; (page 42) © Josh Soto/Unsplash; (page 53) © DC Studio/Shutterstock; (page 54) © ByDroneVideos/Shutterstock; (page 57) © Amrut Roul/Unsplash; (page 59) © Austin Distel/Unsplash; (page 78) © Lise Blekastad; (page 82) Courtesy of Liz Vagenas; (page 83) © Helena Lopes/Unsplash; (page 88) © Orlando downtown by remainphotographyLLC

‡ Data Sources: (page 8) AV-TEST Institute: “Malware Statistics & Trends Report”; (page 9) Verizon: “2026 Data Breach Investigations Report”; (pages 15, 16) Bloomberg Law: “CNA Financial Paid $40 Million in Ransom After March Cyberattack”; (page 16) Australian Cyber Security Centre: “Cyber sanction imposed on Russian cybercriminal for 2022 Medibank Private compromise”; (page 36) Matthew Rodda and Vasilios Mavroudis: “Analysis of Publicly Accessible Operational Technology and Associated Risks” (arXiv, 2025); (page 39) International Federation of Robotics: “Japan’s Car Industry has Highest Robot Installations in Five Years”; (page 50) FBI Internet Crime Complaint Center: “2024 Internet Crime Report”; (page 50) CNN: “British engineering giant Arup revealed as $25 million deepfake scam victim”; (page 53) Parliamentary Monitoring Group: “Ekurhuleni Metropolitan Municipality hearing on latest audit outcomes”; (page 53) Daily Maverick: “Gauteng was lucky with latest 3.8TB data breach, but the luck will run out”; (page 53) IOL: “DDoS attacks hit South Africa: How cybercrime disrupted everyday internet users”; (page 53) DefenceWeb: “Department of Defence suffered nearly 300 cybercrime incidents over five years”; (page 54) Waterways News: “NCS Reinforces B’Odogwu Platform Security Following Port Operations Disruption”; (page 54) African Security Analysis: “Kenyan Government Targeted by White-Supremacist Defacement Attack: Coordinated Intrusion Disrupts State Digital Infrastructure”; (page 54) Resecurity: “Cybercriminals Attacked National Social Security Fund of Morocco – Millions of Digital Identities at Risk of Data Breach”; (page 54) Live Science: “Hackers used AI to steal hundreds of millions of Mexican government and private citizen records in one of the largest cybersecurity breaches ever”; (page 55) Grand View Research: “Online Food Delivery Market Size, Share Report, 2025–2030”; (page 55) African Development Bank: “Over 66 million Africans Gain Digital Access Owing to African Development Bank Infrastructure Push”; (page 55) Global Health Action: “A decade of designing and implementing electronic health records in Sub-Saharan Africa: a scoping review”; (page 69) IBM: “Cost of a Data Breach Report 2025”; (page 73) S&P Global: “S&P Global Report Charts Enterprise Race to Build AI Agent-Ready Infrastructure”; (page 75) IBM: “X-Force Threat Intelligence Index 2025”; (page 76) Google Threat Intelligence Group: “Look what you made us patch: 2025 zero-days in review”; (page 78) Cybersecurity Dive: “Major companies keep hiring North Korean IT workers”; (page 78) Microsoft: “Inside Tycoon2FA: How a leading AiTM phishing kit operated at scale”; (page 78) Trend Micro: “Revisiting 16shop Phishing Kit, Trend-Interpol Partnership”; (page 85) Darktrace: “Darktrace’s View on Operation Lunar Peek: Exploitation of Palo Alto Firewall Devices (CVE-2024-0012 and CVE-2024-9474)”; (page 85) Ctrlaltintel: “FortiWeb to Sliver: Tracking a Long-Term Access Campaign.”

In our customer's words

ThreatLocker was different in that it didn't just respond, it prevented.

That was a game changer. It prevents unauthorized applications from being installed, including malware and ransomware, and protects against living-off-the-land attacks until an admin allows it.

Steve Koenig | IT Manager
Walsh Gallegos