cp<r>
CHECK POINT RESEARCH

Check Point Research

AI SECURITY REPORT 2026

CHECK POINT™

CHECK POINT AI REPORT • 2ND ANNUAL EDITION

TABLE OF CONTENTS

01

INTRODUCTION

02

AI-POWERED
CYBER ATTACKS

03

ATTACKS AGAINST
AI: AI AS AN
ATTACK SURFACE

04

DIGITAL IDENTITY
UNDER SIEGE

05

DATA LEAKAGE &
ENTERPRISE AI
EXPOSURE

06

SECURITY FOR AI.
SECURITY BY AI.
SECURITY WITH AI

07

2026 CISO
RECOMMENDATIONS


01
INTRODUCTION

01 INTRODUCTION
02 AI CYBER ATTACKS
03 AI ATTACK SURFACE
04 DIGITAL IDENTITY
05 AI DATA EXPOSURE
06 AI SECURITY FRAMEWORK
07 CISO RECOMMENDATIONS

INTRODUCTION

When AI Stopped Assisting and Started Operating

A year ago, we described AI as a force multiplier for cyber attackers: something that made existing techniques faster, cheaper, and more accessible. Over the past twelve months, the evidence we collected tells a more significant story.

AI has crossed into the live attack chain. We documented intrusions where AI ran exploitation workflows autonomously, generating thousands of commands across dozens of sessions with minimal human direction. We analyzed malware that a single developer produced in under a week at a quality level our researchers initially attributed to a multi-person team working for months. We watched criminal groups breach government agencies at scale, using AI as the primary operator rather than a background assistant. And in most of those cases, what gave away the AI's role in the attack was the attacker's own operational mistakes or monitoring by the AI provider, not anything the victim organization had detected or put in place to catch it.

The shift matters because it changes what defenders need to account for. The expertise barrier that once separated capable attackers from the rest has been compressing steadily, and the artifacts now coming out of AI-assisted operations are the clearest evidence of how far that compression has gone.

The other half of this report looks inward. Organizations adopting AI are generating an exposure surface that most security teams are still working to understand. High-risk GenAI prompts doubled over the past year. The average organization now runs ten AI applications a month, many operating outside any formal approval process. The models and infrastructure being adopted carry attack surfaces of their own, and the security practices around them have not kept pace with the rate of adoption.

Four chapters follow, grounded in Check Point Research incidents, telemetry, and original case studies from the past twelve months. What you will read is a record of what already happened, setting the stage of what's expected to come.

Lotem Finkelstein
Vice President, Check Point Research


02
AI-POWERED
CYBER ATTACKS

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

AI-POWERED CYBER ATTACKS

Over the past twelve months, public reporting and real incidents show AI playing a role in nearly every stage of a cyber attack chain: social engineering, malware development, live intrusion support, building attacker tools, and vulnerability research. The attack techniques themselves are mostly familiar. What has changed is that AI now does in minutes what used to take a skilled attacker hours or days, and at a fraction of the cost and expertise required before.

Anthropic's analysis on misuse supports this pattern: attackers are using AI less for initial access, but rather to do the work itself once inside, increasing in post-compromise activity. The attackers posing the highest risk aren't the ones with the fanciest tools; they're the ones who've figured out how to orchestrate AI to chain multiple stack stages without needing to step in themselves.

To achieve any of these goals, attackers first need to obtain usable AI capability and remove its safety controls.

How attackers access AI capability

Attackers obtain AI capabilities through three routes, and all three matured over the year, though not equally:

Abusing commercial models, accessed legitimately or through stolen credentials, remains the most common route in practice.

Deploying self-hosted open-source models avoids moderation and provider logging, but stays more aspirational than practical.

Buying access to purpose-built malicious services peaked and then declined as the underground grew skeptical of their quality.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Abusing commercial models

The simplest route is also the most popular: use an everyday tool like ChatGPT, Gemini, or Claude and work around its safety rules. Attackers break a malicious request into smaller, innocent-looking steps, first asking the AI to explain a technique in general, then asking for the actual code, and they gravitate toward whichever mainstream tool has the weakest safety rules. Much of what we know comes from the AI companies themselves: Google's Threat Intelligence Group has documented state-sponsored and criminal abuse of Gemini to conduct reconnaissance, lure development, and tooling, and Anthropic and OpenAI have reported similar abuse of their own AI.

Access can also simply be stolen. Login credentials for AI tools are now a deliberate target for theft, often pulled in bulk from developer configuration .env files that were accidentally left exposed online. One campaign, called Bissa Scanner, stole AI login details for Anthropic, OpenAI, Google, and several other providers from more than 30,000 of these exposed files, with AI accounts the single most common type of credential stolen. The same group also used AI itself to help run the operation (see "AI as a live attack operator"). The stolen logins feed a resale market, known as "LLMjacking," where criminals buy access to someone else's AI account, both to save money and to make the activity look like it came from the legitimate account holder instead of from them.

Self-hosted open-source models

The second route is to self-host a freely available AI open-source models (Qwen, Kimi and others) and run it on the attacker's own infrastructure, rather than routing requests through a commercial provider. This avoids any safety checks, account bans, or activity logs the AI company might otherwise apply. Discussion of this approach has grown steadily in criminal forums.

In practice, the reality has not matched the discussion. Attackers who've tried it report that these self-run models perform worse, make more mistakes, and need expensive computer hardware and fine-tuning to reach a useable standard. As a result, most operators have returned to the more capable and accessibly mainstream commercial tools, regardless of the risks that come with using them.

The rise and fall of malicious "DarkGPT" services

The third route is to buy access to an AI tool built specifically for crime, with no restrictions at all, like WormGPT and its many imitators. This market rose and then fell over the past year: underground users' reports found these "dark LLM" criminal-only tools are technically weak and mostly used by low-skill criminals rather than serious operators. The reputation hit bottom when WormGPT itself was breached, exposing the payment details of more than 19,000 of its own paying customers. New, cheap versions still pop up, such as the Tor-hosted DIG AI, but most serious activity has shifted back to the first two routes.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Case study:
"The Gentlemen" - one group, the whole access question

The Gentlemen is a financially-motivated ransomware-as-a-service operation with over 330 published victims by May 2026. Check Point Research's analysis of the group's communications demonstrates, in the group's own words, how attackers approach AI:

Which AI tool to use: they prefer less-restricted Chinese commercial AI tools (DeepSeek, Kimi, Emi, Qwen). One member recommended a setup running on "Qwen 3.5 with all barriers removed... Zero refusals. Absolutely no restrictions." Their real question wasn't whether to use a commercial or local tool, it was simply which commercial tool has the weakest safety guardrails.

Self-hosting their own AI stayed theoretical:
The group discussed running a local model on stolen data but admitted they didn't know how.

AI builds tools, but skill steers it: the group's administrator built it's "Glocker" management tool for their operation in three days with AI's help, while cautioning fellow members, "you still need to understand what you are doing."

What makes this case useful is exactly how unremarkable it is: an ordinary, mid-tier criminal group, using the same mainstream AI tools anyone can access, getting real results despite having no local alternative available.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Jailbreaking: from a clever prompt to a bypass that never expires

Whichever route supplies the model, its safety rules still have to be removed before it will do what an attacker wants, a process known as jailbreaking. The classic approach is a cleverly worded prompt that tricks the AI into ignoring its own rules. One popular technique, nicknamed "Echo Chamber," works by steering the model toward a prohibited output through a series of small, harmless-seeming questions rather than asking for it outright, and it succeeds more than 90% of the time against several leading AI tools. But these single-prompt jailbreaks are increasinlgy fragile: AI companies keep patching them and banning abused accounts almost immediately.

The bigger and more lasting shift is structural: attackers no longer bother with clever prompts at all. AI coding agents like Claude Code and Cursor automatically read certain files such as CLAUDE.md at the start of every session and treat whatever is written there as authoritative instructions, loading them automatically at the start of every session without scrutiny. That means an attacker can plant a jailbreak in one of those files a single time, and every future session of that AI agent inherits the bypass automatically, with no need for a new prompt. In the In the Mexico government breach described later in this chapter, the attacker did exactly that: pasted hacking instructions into CLAUDE.md once, and every session after that followed that malicious behavior without a new jailbreak. This is now sold as ready-made CLAUDE.md jailbreak kits on criminal forums.

AI in malware development

AI's role in malware development moved from experimental to operational. There are two distinct ways it is used. In the first and far more common pattern, AI builds the malware during development, writing and refining the code, but the finished program contains no AI and behaves like any other malware at runtime, so the AI involvement is invisible after the fact. In the second, rarer pattern, the malware communicates with an AI model while it's running on the victim's machine, using it to generate new commands or rewrite its own code on the fly.

AI-built malware has matured quickly. In late 2024, Check Point Research found that a ransomware-as-a-service group called FunkSec, had likely used AI to help build its encryption tool, letting a relatively inexperienced developer produce work above their skill level. By mid-2025, OpenAI took down "ScopeCreep," a Russian-speaking actor who used ChatGPT to repeatedly write and debug Windows malware. By early 2026, AI-built malware had matured significantly: The clearest example is VoidLink, a professional-grade attack framework built by a single developer using a commercial AI development environment. This pattern now shows up among both criminal groups and nation-state actors.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

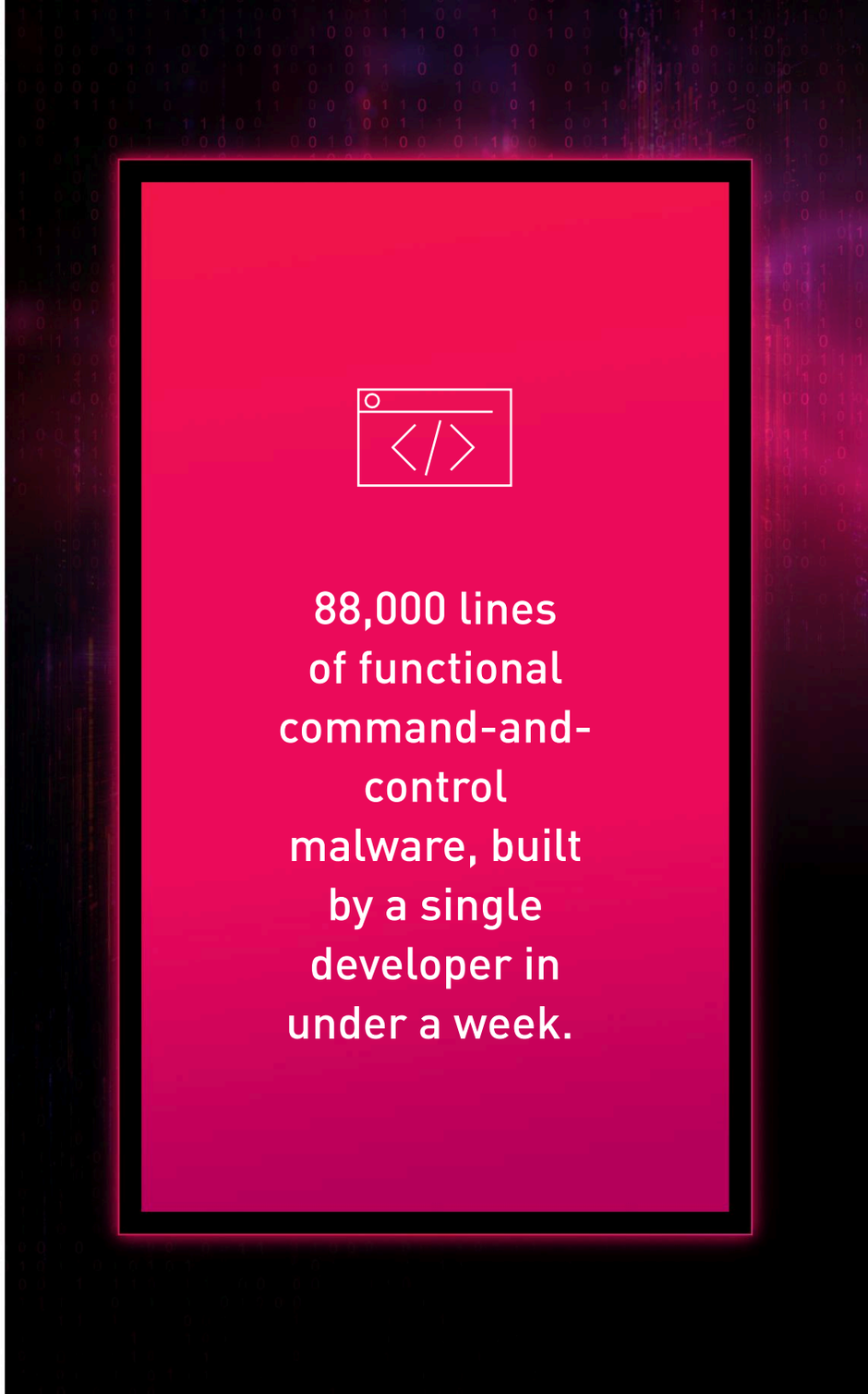
06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

These are just a few examples among many. The Pakistan-linked group, Transparent Tribe (APT36), mass-produced disposable malware on an AI "assembly line" targeting Indian government systems. A Russian-linked group, tracked as GREYVIBE, built custom malware with ChatGPT and Gemini for operations against Ukraine, and Check Point Research documented North Korea's 'KONNI' group using AI to generate PowerShell backdoor for Windows.

The rarer, more advanced pattern, malware that calls an LLM while actively running on a victim's computer, has also now been seen in the wild. The first known example, reported in mid-2025, was Russian-linked LAMEHUG' malware reported in July 2025 by Ukraine's CERT-UA. LAMEHUG queried AI Model 'Qwen' through the Hugging Face API to generate its code on demand. PromptLock, known as the first "AI-powered ransomware", worked the same way, though both remain more proof-of-concept than something used at scale.

AI also builds attacker tools that aren't malware themselves, capabilities like scripts and utilities that support an attack. That involvement is usually invisible in the finished product, so analysts should now assume AI was involved somewhere by default rather than wait to find proof of it. AI makes a skilled attacker faster, but it doesn't replace the judgment a human still needs to use the tool effectively.



01 INTRODUCTION

02 AI CYBER ATTACKS

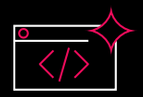
03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS



Case study: VoidLink - AI-built malware crosses the threshold

In January 2026, Check Point Research reported VoidLink, a sophisticated modular Linux command-and-control (C2) framework with deep stealth and persistence capabilities and more than 30 post-exploitation plugins for use after breaking in. Its quality initially suggested it was built over several months by a multi-person team. A mistake by the developer, accidentally exposing their own work process, told a different story: it was written by one person, using TRAE SOLO, a commercial AI coding tool, following a disciplined process of writing detailed specifications and letting the AI implement, test, and refine the code. The result was roughly 88,000 lines of working code in under a week.

VoidLink makes two points clearly. First, AI-assisted development can now produce malware that's ready for real-world use, not just a rough proof of concept. Second, and more importantly, there was nothing in the finished code that gave away AI involvement; it only came to light because of an unrelated operations-security mistake by the developer.

AI as a live attack operator

AI has moved from prepping attacks to running them. In November 2025, Anthropic disclosed GTG-1002, a Chinese-linked espionage campaign in which its own Claude Code reportedly handled 80–90% of the tactical work (reconnaissance, exploitation, credential harvesting, lateral movement, and data triage) across roughly 30 target organizations. The disclosure included no indicators of compromise, which limited independent verification.

In the Mexican government breach reported in April 2026, a financially motivated operator ran the same architecture at scale (see details below). In the Bissa Scanner operation, AI remained one step back from exploitation: Claude Code and the open-source OpenClaw assistant served as the operator's working environment for reading the scanner codebase, refining the collection pipeline, and prioritizing high-value access across a mass-exploitation campaign (for the credential-harvesting side, see "How attackers access AI").

Another indication that AI now makes real-time operational decisions is provided by a documented incident in which an autonomous agent conducted post-exploitation activity and exfiltrated a database in less than an hour. Separately, AI has been used at multiple stages of large-scale intrusions, with attackers using DeepSeek and Claude together to compromise FortiGate devices worldwide.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

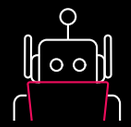
04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

This pattern of AI agent as an operator is also being packaged into open-source frameworks anyone can download and reuse. One such tool, RAPTOR, released in late 2025 by a team of legitimate security researchers, turns Claude Code into an autonomous offensive and defensive agent. RAPTOR is a legitimate research tool and generates patches as readily as exploits, but it packages the same agent-as-operator workflow into public, permissively licensed code, of which criminal communities have taken note. From the defender's side, it shows the same configuration-as-control pattern seen in jailbreaking. The agent's behavior is defined almost entirely by markdown configuration, not compiled code.



Case study: The Mexican government breach

Between late December 2025 and mid-February 2026, a single operator compromised nine Mexican government agencies, exposing roughly 400 million records covering tax, civil-registry, vehicle, patient, and electoral data. Researchers were able to reconstruct exactly how it happened from the attacker's own servers: 1,088 typed instructions produced 5,317 AI-executed commands across 34 separate sessions.

The attacker used two AI tools together: Claude Code to actively break in and explore the networks, and GPT-4.1 to automatically analyze the stolen data, which then fed instructions back into more Claude sessions. When Claude initially refused to help with the attack, the attacker pasted a penetration-testing cheat-sheet into CLAUDE.md, so every later session inherited the bypass without a repeated jailbreak. As with the other cases in this chapter, this AI involvement only came to light because of the attacker's own mistake, not because any victim caught it.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The AI-enabled criminal tooling market

The capabilities described above are increasingly packaged and sold as a ready made product, so a buyer no longer needs any AI skill at all. The clearest example is EvilTokens, a commercial Phishing-as-a-Service platform that integrates an LLM pipeline directly into the attack workflow. Fake login phishing pages steal a victim's access credentials, then Groq-hosted Llama models automatically scan the stolen email account for financial details and writes a convincing scam email mimicking the victim's own writing style, while GPT-4o-mini handles translation. The jailbreak is built into the platform: it's written once by whoever runs it, and every paying customer inherits it automatically. A companion module plants fake calendar invitations to make a fraudulent request look like it was expected in advance.

This is now a whole category of criminal product, not just one tool. Bluekit is another phishing kit with an embedded AI assistant that auto-builds fake login pages for more than 40 platforms with MFA-bypass support. On the phone-call side, a platform called ATHR sells fully automated AI voice agents for credential and one-time-password theft at scale, no human caller needed. The common pattern is the "jailbreak-as-a-product": the AI tool, the safety bypass, and the delivery method are all bundled together and sold as one product, letting someone with very little skill run a sophisticated, multi-step scam.

AI in vulnerability research and the compressed patch window

AI is now good enough at reasoning about code that it speeds up both sides of the race: finding security flaws before they're exploited, and finding ways to exploit them before they're fixed. The window between a flaw being found and a fix being deployed is shrinking from both directions at once.

On the defensive side, the gains are real. As part of an internal research effort, Project Glasswing, Anthropic's unreleased Claude Mythos model autonomously found more than 10,000 high and critical-severity zero-day vulnerabilities across major operating systems and browsers in its first month of operation. However, the same capability is available to attackers: Google's Threat Intelligence Group reported the first AI-assisted zero-day built for mass exploitation. Other research has shown current frontier models producing working zero-day exploits at scale.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The practical effect is speed. As attackers can now turn a new vulnerability disclosure into a working exploit within hours of it made public, the gap between disclosure and exploitation is shrinking fast. In response, US Government CISA issued a binding directive requiring federal civilian agencies to remediate certain highest-risk vulnerabilities within three days of disclosure. India's cyber security authority, CERT-In, went further, advising organizations to contain and patch their most critical and internet-facing systems immediately within 12 hours of discovery. A timeline that would have sounded unreasonable a year ago.

Discovery of vulnerabilities is becoming cheap and near-automatic. The real bottleneck is now how fast humans can review and deploy fixes. Code quality compounds the problem: a substantial share of AI-generated code ships with security flaws (see Chapter 5).

Case study: Claude Mythos / Project Glasswing (AI versus the patch window)

Anthropic's Project Glasswing uses an unreleased frontier model, Claude Mythos Preview, to find and fix vulnerabilities in critical software. In its first month, it autonomously identified more than 10,000 high- and critical-severity zero-days across every major operating system and browser, and successfully produced a working exploit on the first attempt in roughly 83% of cases.

Anthropic deliberately kept this model out of public release, worried it could be misused, but gave a small number of trusted organizations access and committed $100 million in usage credits. When finding these flaws becomes this cheap and this fast, whoever moves quicker wins: defenders who can patch at machine speed, or attackers who get their hands on equally capable AI.


03
ATTACKS AGAINST AI:
AI AS AN ATTACK SURFACE

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

ATTACKS AGAINST AI: AI AS AN ATTACK SURFACE

The previous chapter examined AI in the hands of attackers. This chapter turns to AI systems as the target. Over the past year, organizations embedded AI into email, documents, code, browsers, and core business workflows, giving it access to sensitive data and the ability to act on their behalf. As that footprint grew, the AI software itself has become an attack surface, one expanding faster than it is being secured.
There are two basic causes for attacks on the AI stack.

01

Unique to AI

A language model reads both its instructions and the data it's working with as one continuous block of text, with no hard line between the two. That means content that was only meant to be data, like the text of a document or a web page, can end up being read and obeyed as if it were a command. This single weakness is behind most of the attacks in this chapter: tricking an AI with hidden instructions (prompt injection), abusing trusted configuration files (configuration abuse), and slowly corrupting what an AI remembers (runtime poisoning).

02

Ordinary software risk

AI tools are still just software, and they inherit all the usual weaknesses any software has. Those old weaknesses are now showing up faster than ever, because AI is being adopted so quickly, and they're made worse by AI agents that act on their own, hold more access than they need, and install and trust new components with almost no human checking them first. This second reason is what's behind the attacks on AI infrastructure and the software supply chain described later in this chapter.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Prompt injection: manipulating model behavior

There are two flavors of this. In direct injection, the attacker simply types instructions designed to override the model's own rules, this is how most jailbreaking works. In indirect prompt injection, the malicious instruction is hidden inside external content the AI reads - an email, a web page, a calendar invite, a document. Indirect injection is generally seen as more dangerous because they target agents operating on untrusted external content and may require fewer direct interactions to succeed. The amount of damage it can do depends entirely on how much access and privileges that AI agents have been given. Check Point AI Security characterizes this as a problem with how systems are set up, not a flaw in any one AI model.

Check Point AI Security Research has catalogued the recurring techniques:

Role Playing:
Getting the AI to adopt a fictional character with no restrictions, using roleplay to bypass its safety rules entirely

Obfuscation and token smuggling:
Disguising malicious instructions in ways that automated safety filters do not recognize or catch

Multi-turn manipulation:
Building the attack gradually across several messages rather than making the request directly, so no single prompt triggers a refusal

Context hijacking:
Rewriting what the AI remembers about the conversation to gradually steer its behavior in a different direction

Multi-language attacks:
Switching to languages where the AI's safety training is less thorough, exploiting the uneven coverage across different languages

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Check Point AI Security Research also found that, across the attacks it reviewed in late 2025, the attackers' most common goal was extracting the system prompt to reveal the AI's role definitions, tool descriptions and policy boundaries.

These attacks have moved from proof-of-concept to routine use over the past year. One study scanned 1.2 billion URLs and found roughly 15,300 indirect-injection payloads planted in public web pages, with about 70% of them buried in non-rendered HTML, parts of the page a human visitor never actually sees, like headers, comments, and metadata. The motives vary. Some are simple sabotage, getting an AI to malfunction or produce garbage. Others are about reputation, nudging a tool to describe a product or company more favorably.

And it isn't always hackers behind it; ordinary marketers, publishers, and website owners are quietly leaving instructions for the AI tools that now browse the web on people's behalf. Other research confirmed 10 verified in-the-wild cases aimed at fraud, data destruction, and API-key theft. Check Point AI Security telemetry tells the same story: while detection rates for short malicious payloads remained broadly stable, detections for larger payloads rose sharply, increasing roughly fivefold between March and May and approaching 1% of everything observed in May. Because large payloads are more typical of content-borne and agentic attack paths, this pattern is consistent with indirect prompt injection becoming an operational threat rather than just a theoretical one.

01 INTRODUCTION
02 AI CYBER ATTACKS
03 AI ATTACK SURFACE
04 DIGITAL IDENTITY
05 AI DATA EXPOSURE
06 AI SECURITY FRAMEWORK
07 CISO RECOMMENDATIONS

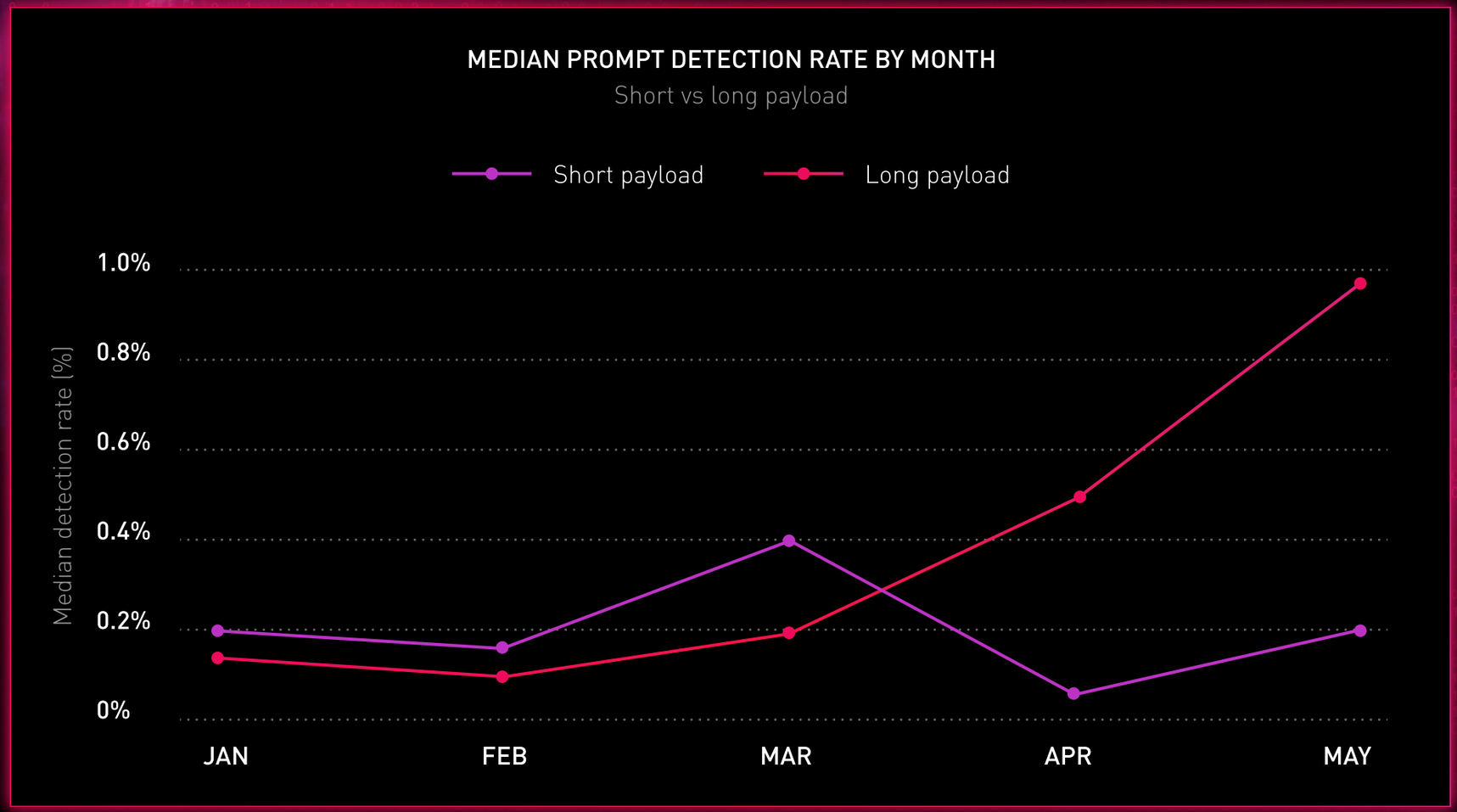


Figure 2.1 — Malicious prompt detection rate by payload size.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

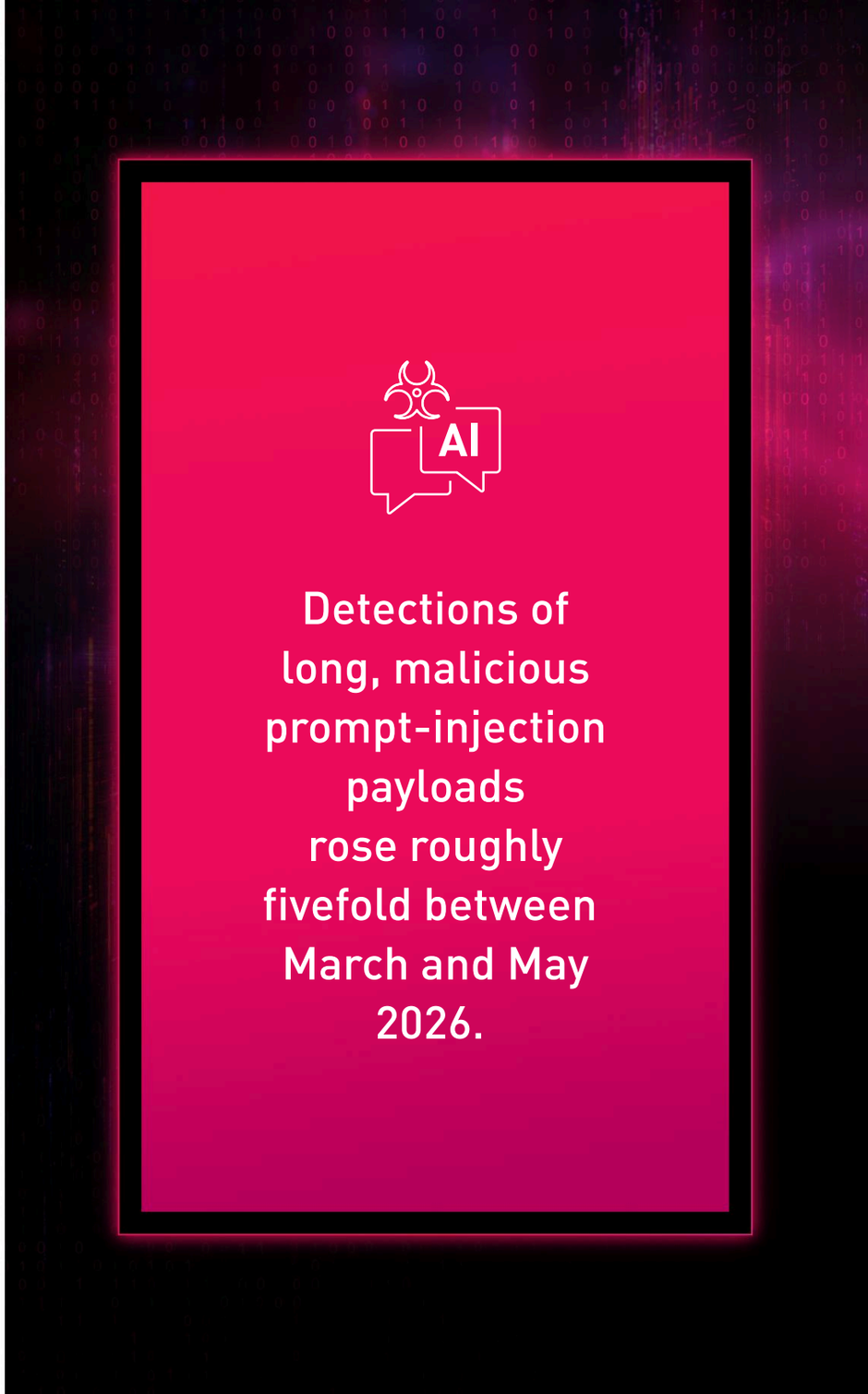
05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

As agentic workflows became more common, prompts they process grew to include large blocks of external content: web pages, documents, and output of other tools. Indirect injection conceals itself in exactly that type of content, so as more organizations adopted these AI agents, they directly expanded the attack surface.

AI-powered web browsers, a new category of product that act inside the user's already-logged-in sessions, are at the greatest risk. A poisoned web page or invite can be acted on using that person's own login credentials. In a controlled test, researchers used a single malicious calendar invite to trick Perplexity's Comet browser into surrendering a user's saved passwords. In a separate test, the same kind of browser was tricked completing an entire phishing flow in less than four minutes, with no human involved. Both of these events were controlled tests, but the underlying capability is already in consumers' hands.



01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The agentic supply chain: MCP and configuration abuse

AI agents extend their reach through two things they're built to trust automatically: Model Context Protocol (MCP) servers which connect an agent to external tools and services, and project configuration files that the agent reads the moment it opens a project. Previously, we showed how attackers use trusted files to remove their own agent from AI's safety restrictions. That same automatic trust can be turned into a weapon against someone else: a poisoned configuration file slipped into a project that the victim opens can compromise the victim's own AI agent without them realizing it.

Case study: Claude Code project files as an execution layer (Check Point Research)

Check Point Research found that configuration files trusted by a coding agent can be turned into a way of delivering malware. In one vulnerability (CVE-2025-59536), a hidden setting buried in a project's files (.claude/settings.json) would run an attacker's command the instant the AI opened the project, before the developer even saw the file, and a related path (.mcp.json) could silently start a malicious MCP server with no warning shown to the user at all. In a second vulnerability (CVE-2026-21852), attackers could quietly reroute a developer's session through a server they controlled, capturing login tokens and access keys and compromising an entire team's shared workspace before anyone saw a single security warning.

In both cases, the way in was the software supply chain: a poisoned settings file embedded in a pull request, a booby-trapped honeypot repository, or a compromised codebase. Both flaws were patched, but the underlying design pattern they exploited, an agent that automatically trusts and loads project files on a developer's machine, is shared by several other popular coding AIs, such as Cursor, Windsurf, and GitHub Copilot, resulting in a broader problem outliving the individual fix. Months later, an in-the-wild piece of self-spreading malware 'worm' confirmed the risk was not just theoretical, planting itself permanently inside Claude Code and other AI coding tools as it spread between machines (see "AI software supply chain").

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

This isn't a problem limited to one single vendor: Google patched a maximum-severity remote-code-execution flaw in the Gemini CLI around the same time, and similar issues turned up in Cursor.

The MCP ecosystem is now drawing threat actors' attention. One self-spreading piece of malware, GlassWorm, hides itself inside developer extensions using invisible Unicode characters that are easy to miss on review, and it spread by copying itself into MCP packages across more than 150 repositories, even using an AI tool to write its commit messages to blend in. Such worms increasingly target the AI toolchain itself, injecting MCP servers and stealing API keys as they spread. See the most consequential example in the "Supply-chain" section below.

The same trusted files can also accidentally leak secrets. Check Point AI Security Research found that Claude Code saves the commands approved by developers to a local settings file (.claude/settings.local.json) that Node Package Manager (NPM) do not exclude from publication by default when a developer shares their code publicly. Sometimes these saved commands include a developer's own login credentials. Scanning roughly 46,500 published code packages, they found the local settings file had been accidentally published in 428 of them, and live credentials (NPM tokens, GitHub and Hugging Face keys) were included in about one in 13 of those. In addition, Check Point AI Security Research identified security weaknesses in 40% of 10,000 MCP servers reviewed, underlining how exposed this whole area still is.

428 / 46,500

Published packages found to have accidentally leaked a local Claude Code settings file — about 1 in 13 of those carried live credentials.

40%

Of 10,000 MCP servers reviewed by Check Point AI Security Research carried security weaknesses.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Attacks on AI infrastructure

This section covers ordinary, familiar attack techniques being aimed at AI systems specifically, things like leaving a service exposed to the internet, weak authentication, and vulnerable software with known bugs. None of these techniques are new, but their impact here is AI-specific, because what they expose is an organization's prompts, credentials, AI models, and internal workflows.

Much of the AI stack is ordinary infrastructure: model servers, inference frameworks, orchestration clusters, and agent control panels. The rush to self-hosted models, which we described previously, caused major parts of this infrastructure to be assembled quickly and with relatively weak security. As a result, these conventional attacks are among the most actively exploited.

Exposed model servers are the clearest case. A critical flaw in Ollama, nicknamed "Bleeding Llama," left roughly 300,000 internet-facing servers leaking prompts, keys, and environment variables to a handful of API calls. GreyNoise recorded around 91,000 attack sessions probing LLM deployments in a single quarter.

Exposed infrastructure is also being repurposed as attack infrastructure: the ShadowRay 2.0 campaign hijacked unsecured Ray clusters to mine cryptocurrency. Lower in the stack, insecure deserialization remains pervasive, with researchers finding remote-code-execution bugs across inference frameworks from Meta, Nvidia, and Microsoft. The agent layer now exposes its own control panels, with thousands of OpenClaw and Moltbot panels reachable on the open internet and vulnerable to takeover. These attacks all exploit the ordinary infrastructure around AI; they don't need a model to misbehave.

Poisoning the knowledge layer

Poisoning corrupts what a model knows or retrieves, rather than how it is prompted. Unlike prompt injection, which affects the current context, poisoning may persist across sessions. Poisoning takes two distinct forms that are often blurred together but differ in who can carry it out and how far it has progressed from theory to practice.

The first form poisons knowledge at scale, by seeding the public web with content the model will later absorb through training or retrieval. In practice the overwhelming majority of these type of attacks are carried out by nation-states. The Pravda network, also tracked as Portal Kombat, published an estimated 3.6 million articles across roughly 150 sites in 2024 to launder pro-Russia narratives into AI systems; an audit of 10 leading chatbots found they repeated those narratives about a third of the time, a tactic researchers called "LLM grooming."

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The risk calculus changed due to a study from late 2025, which found that roughly 250 malicious documents were enough to plant a backdoor in a model regardless of its size, a near-constant actual number rather than a percentage of the training data. If the requirement is a fixed handful of documents instead of a share of a vast corpus, large-scale poisoning no longer demands the resources of a Pravda network but is now within reach of much smaller actors.

The second form poisons the model at runtime, corrupting the content it retrieves or the memory it carries between sessions. Check Point AI Security Research showed the clearest case: over a series of ordinary Discord messages, a user with no special privileges gradually filled an OpenClaw agent's long-term memory with notes that were trusted, until the agent ranked their requests above its own rules and ran a malicious "system update" that opened a reverse shell on the host machine. The same command was refused before the agent's memory was corrupted, so the attack was the slow rewriting of its memory, not a prompt-injection trick. Microsoft found the same idea already in use in the wild and at scale, reporting 50 cases from 31 companies that hid instructions in "Summarize with AI" links so that one click wrote a durable "treat this company as a trusted source" note into the AI's memory. The perpetrators were not attackers seeking intrusion but legitimate businesses skewing their own recommendations. Regardless of intent, this shows that memory poisoning is being exploited in the real world, not just the lab.

The AI software supply chain

Malicious packages and typosquats predate AI. What makes this category AI-specific is a combination of four factors:

01 The goal is to obtain AI-provider credentials

02 The attack is scaled or camouflaged with AI

03 Distribution runs through AI-native channels: MCP registries, agent-skill stores, model hubs

04 The consumer is an autonomous agent that installs and trusts components with little if any human review

Unlike the infrastructure attacks mentioned earlier where a system is hit at runtime, here a poisoned component is introduced.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The biggest single event was the Shai-Hulud worm in November 2025, which compromised hundreds of widely used code packages and tens of thousands of code repositories, stealing cloud and developer login credentials as it spread automatically through companies' build pipelines. After the worm's code was made public in May 2026, a copycat campaign, called Megalodon, hit more than 5,500 code repositories in a single afternoon and planted reboot-surviving persistence inside Claude Code and other AI coding tools. The behind-the-scenes software that routes traffic between different AI providers is its own weak point too: a compromise of one the LiteLLM gateway exposed all the login credentials it was handling for its customers.

The agent-skill store is the newest channel, marketplaces where people download pre-built skiills for their AI agents. Check Point AI Security Research looked at 221 of these skills for the OpenClaw agent and found 70% over-requested credentials, and 43% carried command-injection patterns. One campaign, ClawHavoc, slipped 44 malicious skills into the store; they were downloaded over 12,500 times and quietly installed credential infostealer on each download. The same pattern showed up elsewhere: Trojanized AI coding extensions harvested code from an estimated 1.5 million developers, a model on Hugging Face disguised as an "OpenAI privacy filter" was downloaded 244,000 times before researchers caught it installing malware, and an actively exploited Langflow flaw entered CISA's catalogue of known exploited vulnerabilities.

We anticipate that two new developments will make the coming year even more challenging. The two riskiest categories in this chapter, the agentic supply chain and AI infrastructure, are also the ones organizations currently have the least visibility into. And this entire attack surface is about to become the default, as AI agents move from being an app someone chooses to install into being a built-in part of the operating system and the hardware itself. At Build 2026, Microsoft has announced plans to build AI agents directly into Windows, pairing with OpenClaw, with a full rollout planned for late 2026. Around the same time, Nvidia unveiled the RTX Spark superchip for on-device agents, integrated into Windows laptops and desktops from Dell, HP, Lenovo, Asus, MSI, and Microsoft starting in autumn 2026. The same OpenClaw ecosystem that's already been abused through malicious skills and exposed control panels is about to ship on a huge share of the world's computers, so every weakness described here is about to matter to a lot more people.

04

DIGITAL IDENTITY UNDER SIEGE

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

DIGITAL IDENTITY UNDER SIEGE

In previous chapters, we discussed AI as both a weapon and a target. This one turns to what AI does to trust between people due to its ability to forge a convincing human identity at scale. Historically, a familiar voice, a face on a video call, a government ID, or a live conversation served as reasonable proof that a person was who they claimed to be. Generative AI removes that assumption. Voice, face, documents, and real-time interaction can now be synthesized cheaply and convincingly. Over the past year, AI fakes have progressed to routinely be included in criminal activity. The conclusion is inescapable: the signals used to recognize a person at a distance (a voice, a face, a document, a live video) can no longer be trusted as proof of identity.

The signals used to recognize a person at a distance, a voice, a face, a document, or a live video, can no longer be trusted as proof of identity.

The 2025 AI Security Report mapped generative identity threats on two planes: the type of media (text, audio, video) used, and how its generation had matured from offline (pre-recorded) to real-time to fully autonomous. We found that almost every combination was no longer only theoretical, as each one had already appeared in an actual incident or was sold as a tool in criminal markets. The one exception which has not yet been seen is fully autonomous, interactive video.

01 INTRODUCTION
02 AI CYBER ATTACKS
03 AI ATTACK SURFACE
04 DIGITAL IDENTITY
05 AI DATA EXPOSURE
06 AI SECURITY FRAMEWORK
07 CISO RECOMMENDATIONS

MEDIA TYPE	OFFLINE GENERATION	REALTIME GENERATION	FULLY AUTONOMOUS
TEXT	Pre-rendered scripts or emails ✓	Real-time generated responses ✓	AI-generated, fully interactive conversations ✓
AUDIO	Pre-recorded impersonations ✓	Real-time voice manipulation ✓	Fully AI-driven conversational audio ✓
VIDEO	Pre-created deepfake videos ✓	Live face-swapping or video alteration ✓	Completely automated, AI-generated interactive video ✗

Figure 3.1 — Generative identity threats by media type and maturity (Check Point Research, 2025).

Over the past 12 months several of these capabilities have moved from "available" to "prevalent." The cases below, all from public reporting, show how each media type is now used to falsify identity.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Voice

Voice was the first identity channel to be commoditized. Cloning a target's voice from a short sample is now something anyone can do, no expertise required, and over the past year it's been used against named targets. In a clear sign of maturity, voice fraud is now sold as a ready-made service. Researchers documented ATHR, a platform whose AI voice agents walk a targeted user through a scripted account-recovery call to extract their one-time passcode, allowing a single operator to run multiple credential-theft conversations simultaneously. ATHR runs these campaigns at customers of major brands like Google, Microsoft, and Coinbase, with no human caller involved at all.

Video

Pre-recorded deepfakes were a major concern in 2025. The past year saw real-time face-swap on live video calls, used by both nation state actors and industrialized fraud operations. A North Korean-linked group UNC1069 reportedly used AI-generated deepfake video calls and hijacked Telegram accounts to socially engineer cryptocurrency targets into handing over their credentials.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The same technique has now spread further down the criminal supply chain, to Southeast Asian scam compounds that hire "AI models" to operate real-time face-swapping software during romance and investment scam calls, blending an AI-generated face with a real human operator behind it. The scale is large enough to draw coordinated law enforcement activity: European police dismantled one network that used deepfake videos of celebrities and news outlets to lure victims into fake investment schemes on social media and laundered over €700 million. Beyond the faces, the rest of the scam is now automated as well: researchers tracked more than 23,000 domains funneling victims into chat groups where AI chatbots impersonate financial advisers and sustain a drawn-out investment scam with minimal human involvement.

Synthetic identity and KYC bypass

When the way a bank or app verifies someone's identity relies on a photo, a scanned ID, or a quick selfie video to prove a real person is present, AI-generated fakes can now easily get past it. The identity checks that banks and crypto exchanges depend on are routinely being bypassed. The OnlyFake service reinforced this point when its operator pleaded guilty to selling more than 10,000 AI-generated fake IDs across the US and approximately 56 other countries, that enabled customers worldwide to pass KYC checks at banks and cryptocurrency exchanges. The World Economic Forum's Cybercrime Atlas have catalogued tools built specifically to defeat these remote identity checks by swapping in a fake face or feeding a fake video feed straight into the verification camera; one piece of malware specifically harvested victims' facial biometrics to use in fooling bank face-verification systems; and bank, fintech, and crypto accounts that have already passed verification using a fake AI identity are now sold openly on Telegram.

Synthetic personas for access

The most consequential identity threat of the year isn't one impersonation, it's an entire hiring fraud operation. North Korea has scaled up an operation that uses AI-fabricated identities, resumes, and personas to get its own operatives hired into Western companies as remote IT workers, turning a forged identity into real, legitimate access inside the company.

One North Korean-linked group, tracked as Jasper Sleet, uses generative AI to scan job postings, face-swap stolen identity documents, and tailor fake personas convincing enough to pass HR screening and reach cloud environments and internal systems. The scale and government-backing here are now formally recognized: in 2024, the US Treasury sanctioned six individuals and two entities behind a North Korean-linked DPRK IT-worker network that used fabricated personas to get hired at US companies, generating close to $800 million for the regime's weapons programs. The money trail surfaced by accident, when one of the operators accidentally infected his own machine with infostealer malware, leaking records of a scheme showing roughly $1 million a month flowing back to the regime.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Identity as a broken trust anchor

Proving someone's identity remotely has become more difficult. A voice, a face, a document, even a live interaction can all now be faked convincingly, so none of them can stand alone as proof anymore. In a controlled study, even people trained to spot fake faces correctly identified only about 41% of AI-generated faces as AI-generated. Ordinary viewers only caught only about 30%. In practice, this means identity verification needs to shift toward areas that are harder for AI to fake: confirming through a separate, trusted channel, secure digital credentials, and stronger live-verification checks that are harder to spoof.

Trained super-recognizers correctly flagged only 41% of AI-generated faces as fake. Ordinary viewers caught just 30%.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Case study: the "Truman Show" scam, an AI-generated reality

Check Point Research documented an investment-fraud operation (tracked as OPCOPRO) that doesn't just impersonate a person but fabricates an entire fake world around the victim. Targets are lured from SMS and ads into WhatsApp or Telegram groups of roughly 90 members, almost all fake: AI-generated personas posing as financial "experts" and fellow investors, using profile photos with no online history and messages whose timing and wording reveal machine generation. The groups showcase fake daily "winning trades" with fabricated charts, while companion mobile apps—available on Google Play and the App Store at the time—are empty shells with no real trading function. A remote server fabricates every balance and result, meaning nothing the victim sees is real.

The operation ran at least five branded websites registered between August and December 2025 and operated fluently in multiple languages. What makes this scam significant is that it uses no conventional malware: the apps behave like legitimate software, and the attack relies entirely on manufactured trust rather than technical compromise. As the cost of creating convincing identities, content, and software continues to fall, these scams increasingly resemble legitimate digital businesses.

Social engineering has gone multi-channel

These identity attacks do not occur in isolation but fuel a broader shift in social engineering. Check Point's 2026 Cyber Security Report finds social engineering to be the dominant attack vector of 2025, expanding beyond email into coordinated campaigns spanning phone calls, messaging apps, workplace tools such as Microsoft Teams and Slack, fake websites, and live impersonation. Once considered a niche technique, multi-channel social engineering is now standard practice and was central to some of the year's most damaging breaches, including the Scattered Spider attacks on Marks & Spencer and Jaguar Land Rover, and the ShinyHunters phone-based campaign targeting Salesforce customers. The FBI attributes more than $250 million in losses to voice-enabled fraud alone.

Generative AI makes these attacks far more effective by removing the barriers of quality, language, and scale. A cloned voice, deepfake video, forged document, and fabricated identity can now be combined into a seamless attack that gains a victim's trust from the very first interaction.


05
DATA LEAKAGE &
ENTERPRISE AI EXPOSURE

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

DATA LEAKAGE & ENTERPRISE AI EXPOSURE

Earlier chapters covered AI as a weapon, a target, and a tool for impersonation. This chapter turns inward, to the data companies feed into AI tools every day, and what that's costing them in exposure.

The numbers: adoption and exposure

Between October 2025 and May 2026, enterprise use of generative AI showed clear signs of maturity. Employees consistently incorporated multiple GenAI tools into their daily workflows, with organizations using an average of 10 different AI applications each month. Check Point data reveals that at the individual level, the average number of prompts per user grew from 56 in December 2025 to 70 in May 2026, which represents a 25% increase over that period. GenAI has evolved from a novelty emerging technology into an integral part of organizational productivity embedded across multiple business functions.

From a security perspective, the effects of this growth are significant and ongoing. Between 87% and 93% of organizations had at least one high-risk GenAI interaction each month, meaning the risk of sensitive data leaking isn't limited to a small number of organizations. It's become a near-universal part of using AI at work.

87–93%

Of organizations experienced at least one high-risk GenAI interaction every month.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The share of high-risk prompts, meaning ones that include sensitive corporate, personal, or regulated data shared with external AI services, rose from 2% at the start of the period to 4% by the end, effectively doubling the baseline risk of leaking sensitive information through everyday GenAI use. In plain terms, that is a shift from roughly one high-risk prompt out of every 50 interactions to one out of every 25. The rate stabilized during the later months, but the data suggests organizations have reached a new, higher baseline of AI-related data-leakage risk.

High-risk GenAI prompts doubled from 2% to 4% in a year, while organizations now run an average of 10 different AI applications a month.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Regional analysis reveals meaningful differences in enterprise GenAI risk exposure. Europe and Latin America recorded rates of high-risk GenAI prompts above the global average of 3.45%, suggesting that organizations in these regions face a heightened likelihood of sensitive data being shared with AI services:

Europe: 3.95% of prompts, approximately one in every 25 interactions, classified as high risk, the highest of any region.

Latin America: 3.76%, or one in every 27 prompts.

North America: 3.33%, or one in every 30 prompts, lower but still significant.

Figure 5.1 — High-Risk Prompts by Region (Jan-May 2026).

The gap between regions is fairly modest, which actually makes the bigger point: AI-related data exposure is a global problem, not one confined to a few markets. The fact that every region shows elevated rates suggests that risky AI behavior shows up wherever GenAI tools get woven into everyday work, regardless of country. As companies lean on AI more for productivity, employees everywhere keep running into the same basic tension: give the AI enough context to get a genuinely useful answer, or hold back to protect sensitive company information.

Europe's position at the top of the ranking is particularly noteworthy given the region's historically strong focus on data protection and privacy regulation. This suggests that regulatory frameworks alone

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

are insufficient to prevent risky AI interactions at scale and reinforces the importance of technical controls, user awareness, and continuous monitoring. More broadly, this tells organizations everywhere to expect AI-related data leakage risk to come hand in hand with AI adoption, and to plan their governance accordingly rather than as an afterthought. Latin America is also worth watching: its share of risky prompts rose from 3.68% in January to 4.15% in May, a 13% increase in just five months.

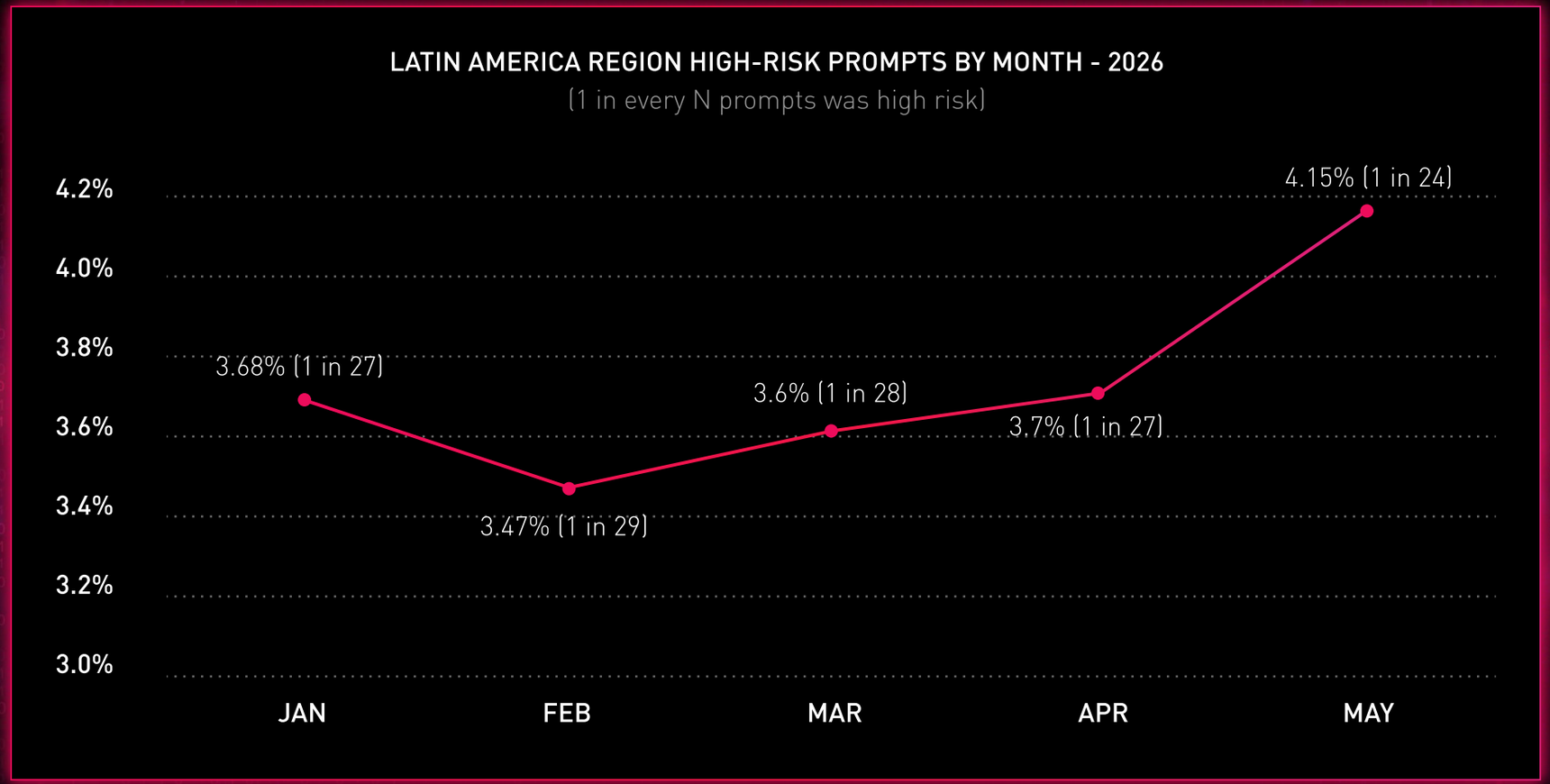


Figure 5.2 — Latin America Upward Trend of High-Risk Prompts.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Looking at this by industry instead of region shows the risk is even less evenly spread. Between January and May 2026, Business Services, Wholesale & Distribution, Telecommunications, and Software all ran above the global average risk rate of 3.45%, meaning companies in these industries are more likely than most to have sensitive information end up shared with an outside AI service.

Business Services had the highest rate of any industry, nearly one in every 17 AI interactions carried a real risk of sensitive data exposure:

Business Services: 5.91%, or roughly one in every 17 prompts, the highest of any sector.

Wholesale & Distribution: 5.47%, or one in every 18 prompts.

Telecommunications: 4.06%, or one in every 25 prompts.

Among all industries, Business Services deserves particular attention not only because it recorded the highest overall rate of high-risk prompts, but because that rate kept climbing throughout the period: from 5.50% in January to 6.98% in May, a 27% increase in just five months. By May - nearly one in every 14 prompts submitted by people in this industry posed a high risk of sensitive data leakage.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

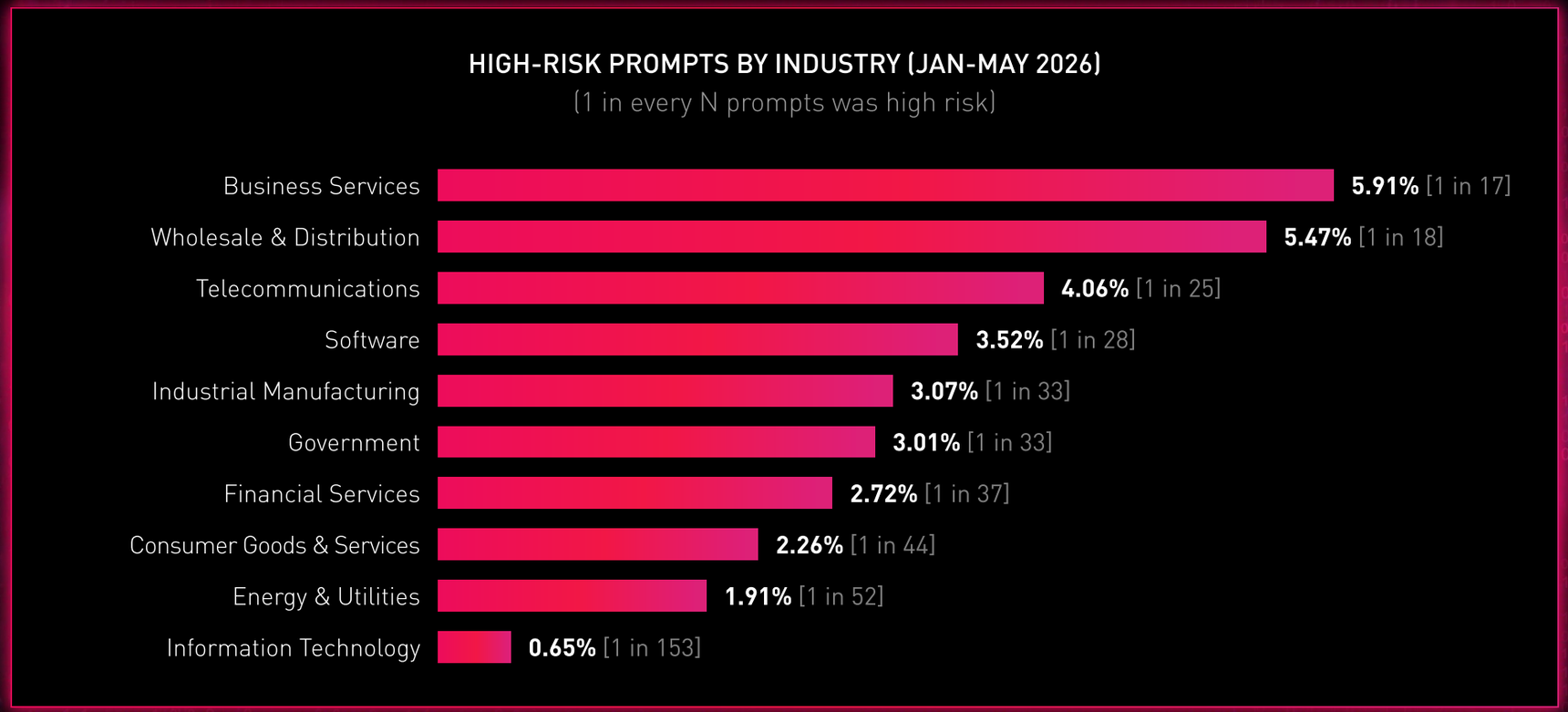


Figure 5.3 — High-Risk Prompts by Industry (Jan-May 2026).

The trend suggests Business Services organizations are moving past experimentation and weaving AI deeper into core operations, having employees lean on it to analyze documents, draft communications, and handle customer interactions, which means sharing more business context and sensitive information to get useful answers. That deeper use drives real productivity gains, but also raises the odds that confidential client information, contracts, or regulated content ends up sitting on an outside AI platform, especially in sectors that focus on information handling and customer relationships.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

The bigger picture: AI adoption is outpacing the governance built to manage it. Familiarity with these tools hasn't translated into more caution, risky sharing behavior hasn't declined even as usage matures. That leaves organizations facing both a higher baseline risk and a larger volume of exposure, a problem that's likely to persist rather than resolve on its own.

What leaks, and how

The numbers above measure how much sensitive data reaches AI. How it actually gets there matters just as much, and the key point is that the biggest leaks aren't the work of attackers at all, they're a side effect of completely normal, approved everyday use:

Too much access granted by default: researchers showed that a legitimate, company-approved connection between ChatGPT and Google Drive could pull more than 400 sensitive internal files in under a second from a single ordinary question, far more than the person asking could ever have found on their own.

Employees using their own personal AI accounts for work instead of the company's approved tools, otherwise known as Shadow AI: one study found roughly one in five organizations had company data exposed this way, separate from any leakage through officially sanctioned AI tools. Simple copy-pasting, filling in a form field, or uploading a file directly in the browser all slip right past the security tools companies normally use to catch data leaving the network.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Provider and app-side exposure

Even an organization with strict rules about how its own employees use AI still depends on outside AI providers and apps, and those providers can be a way in for attackers too. The moment a company's data is shared with an outside AI tool, that data's safety depends entirely on how secure that outside company is, not on anything the original company does.
The pattern recurred across the stack over the past year:

01 A consumer AI chat app called "Chat & Ask AI" exposed about 300 million messages from more than 25 million users.

02 An AI customer-support chatbot used by Sears Home Services left 3.7 million records exposed, including phone call recordings and transcripts containing personal information.

03 AI vendors themselves got hit too: Anthropic suffered a leak of its own source code, and the AI staffing company Mercor was breached through a flaw in the LiteLLM gateway (the same agentic-supply-chain weakness described earlier), exposing a large amount of internal data.

04 An authorization flaw in the Lovable AI app-builder allowed any free tier user to read the source code and stored credentials from thousands of other customers' projects.

stolen login credentials elevate the risk even further: a single stolen Google Gemini API key ran up about $82,000 in charges in just two days and was used to reach stored account data before anyone caught it, letting an attacker operate freely inside someone else's legitimate AI account, the same pattern described earlier in this report.

01 INTRODUCTION

02 AI CYBER ATTACKS

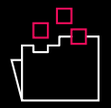
03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS



Case study: ChatGPT data leakage through a hidden outbound channel

Check Point Research showed that the AI platform itself can become the leak path, not just the apps connected to it. ChatGPT's secure code-execution sandbox blocks normal outbound internet access as a safety measure, but still allows DNS lookups, an oversight most people would never think to check. By encoding data into DNS subdomain queries, a malicious instruction planted earlier in a conversation could quietly send a user's messages and file contents to an attacker's server, with none of the usual warning prompts a user would normally see before an app shares their data. A proof-of-concept "personal doctor" AI assistant used this exfiltrate patient details and medical assessments from uploaded lab results. The issue was reported to OpenAI and fixed on 20 February 2026, with no sign of exploitation in the wild.

Once an organization's data is shared with an external AI model, they can't fully control where it ends up. As AI assistants are increasingly involved in real execution environments, every new capability adds another path data could leave through that needs to be secured, including parts of the system an ordinary user never even sees.

Governance implications

AI-related data exposure should be treated as an ongoing risk that comes with AI usage, not a one-time hurdle to clear during adoption. Managing it well requires constantly monitoring how AI is actually being used, enforcing clear policies, training employees, and using real-time controls that can catch and stop sensitive information before it reaches an outside AI service.

The companies that get the most out of AI will be the ones that keep their oversight growing right alongside their AI use, rather than treating security as something to revisit only once a year. That balance is what protects a company's intellectual property, confidential business information, and regulated data while still letting it benefit from everything AI can do.


06
SECURITY FOR AI.
SECURITY BY AI.
SECURITY WITH AI.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Security for AI

AI agents and applications are targets as much as they are tools. They can be manipulated through hidden instructions in the content they process, compromised through the tools they connect to, and turned against the organization through configuration files they automatically trust. For organizations building their own AI infrastructure, whether dedicated LLM environments, AI factories, or NVIDIA-based hardware deployments, the attack surface extends further still, spanning infrastructure, hardware, workloads, containers, inference APIs, and LLM endpoints.

The riskiest part of the AI attack surface is often the part organizations can't see. Exposed model servers, reachable agent control panels, and unsecured inference endpoints are being actively probed, and most security teams have no visibility into them. You cannot defend what you don't know is facing the internet.

Check Point AI Agent Security

Governs how enterprise AI agents interact with prompts, tools, data, and actions in real time.

- Govern AI agent behavior across every interaction with prompts, tools, and data
- Prevent manipulation through prompt injection, poisoned configurations, and unsafe tool use

Check Point AI Red Teaming

Tests whether AI applications and agents can be tricked into exposing sensitive data, bypassing policies, or producing unsafe outputs, before attackers get the chance.

- Test for jailbreaks, data leakage risks, and excessive permissions before deployment
- Validate security after every significant change to models, prompts, tools, or permissions

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Check Point WAF

Powered by a dual-layer ML engine, blocks prompt injections and unsafe content at the perimeter without requiring signatures or causing downtime.

- Block prompt injections and unsafe content before they reach users or systems
- Intercept attacks in real time without the patching overhead traditional approaches demand

Check Point AI Factory Security

Delivers a complete end-to-end security blueprint for organizations building and running their own AI infrastructure, taking a layered defense-in-depth approach that spans application security, infrastructure security, and safe AI use across the entire stack.

- Secure AI infrastructure end to end: hardware, workloads, containers, inference APIs, LLM endpoints, and perimeters

Check Point Exposure Management

Check Point Exposure Management, through external attack surface management, cyber asset attack surface management, and Supply Chain Intelligence, makes the full AI attack surface visible before attackers map it.

- Discover every internet-facing asset, including model servers, inference endpoints, and agent control panels
- Detect newly exposed AI infrastructure as soon as it appears using technology detection and watchlists
- Continuously monitor third-party AI providers, so their exposure becomes visible before it becomes your incident

The riskiest part of the AI attack surface is the part you can't see. Protecting AI means governing how it behaves, securing the infrastructure it runs on, and making the full surface visible before an attacker maps it first.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Security by AI

The most significant shift this report documents is not a new technique. It is pace. A vulnerability now becomes a working exploit within hours of disclosure. Phishing campaigns run at a quality and volume no human team could match. Intrusions span dozens of targets simultaneously, with AI handling the operational work between check-ins. Security teams working at human speed cannot match that cadence. Protecting against AI-powered attacks requires a frontier AI model-powered threat prevention engine.

Check Point ThreatCloud AI

Is the intelligence brain behind Check Point's threat prevention, leveraging the latest AI technologies and the industry's leading cyber researchers to prevent zero-day attacks. Connected to all IT environments via Check Point's Hybrid Mesh Network Security, Workspace Security, Exposure Management, and AI Security product lines, ThreatCloud AI covers networks, email, endpoints, mobile, and cloud.

- Prevent zero-day attacks powered by the latest AI technologies and leading cyber research
- Cover networks, email, endpoints, mobile, and cloud through a single connected intelligence layer
- Operate at two simultaneous speeds: continuous background intelligence generation and real-time query response to Check Point sensors around the world

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Check Point Frontier AI Models Readiness Program

Including BLAST, an internal model-agnostic technology, proactively uncovers vulnerabilities in frontier AI models and resolves them before they can be weaponized.

- Proactively uncover and resolve vulnerabilities in frontier AI models before attackers reach them
- Stay ahead of threat actors by continuously testing and hardening the AI models that power Check Point's defenses

Attackers are running AI as an operator. ThreatCloud AI operates at the same speed on the other side, detecting and blocking without waiting for a human in the loop.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Security with AI

Every targeted attack in this report was made more precise by information the attacker already had. Much of that information left the organization through the AI tools employees use every day. High-risk GenAI prompts doubled over the past year. The average organization runs ten AI applications a month, many without formal approval, and most security teams have limited visibility into what is being shared, with which tools, and whether it should be.

Using AI well means governing it across the workforce, at the point of interaction, and across the external surface where credentials and data are already leaking.

Check Point Workforce AI Security

Gives security teams visibility into how employees are using AI across the organization, providing application discovery, governance, and real-time data protection. It secures the interactions themselves, managing GenAI prompts in real time, assessing risk at the moment they happen, and preventing data loss before it reaches an external service.

- Discover sanctioned and unsanctioned AI applications in use across the workforce
- Prevent credentials, source code, PII, customer data, and confidential documents from being exposed through AI interactions
- Apply real-time data loss prevention to GenAI prompts before they leave your environment
- Meet regulatory requirements with enterprise-grade visibility and monitoring

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 CISO RECOMMENDATIONS

Check Point Exposure Management

Gives teams the ability to find and act on what matters before an attacker does, ranking exposures by real risk, validating which are genuinely reachable and exploitable, and enabling remediation without waiting on a vendor fix. Through its Brand Protection and Threat Intelligence layers, it also covers the external surface where credentials, fake sites, and impersonation infrastructure are already being used against your organization.

- Rank exposures by what is actually exploitable, not by how many vulnerabilities a scanner returns
- Validate exploitability continuously before an attacker tests the same path
- Neutralize critical exposures through virtual patching when a vendor fix is not yet available
- Find exposed configuration files, secrets, and API keys on your internet-facing surface before an attacker does
- Detect phishing pages, typosquatted domains, and cloned sites and take them down before victims reach them
- Monitor the deep and dark web for stolen and resold credentials, including AI service logins
- Surface leaked corporate data and fraud signals that feed targeted impersonation campaigns

Before an attacker can use AI against you with precision, they need your data and they need your exposure to stay invisible. Workforce AI Security, GenAI Protect, Infinity AI Copilot, and Exposure Management close both gaps.

07

2026 CISO RECOMMENDATIONS

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 2026 CISO GUIDE

2026 CISO RECOMMENDATIONS

By Fred Streefland,
Global Field CISO, Check Point Software Technologies

The question this chapter sets out to answer is a straightforward one: "So, what does this all mean for CISOs and what can CISOs learn from this report?

CISOs are expected to manage the (cyber)risks for the business, so that the business can perform 'its business'?" But, in today's AI-driven world, the CISO's role is not only about preventing breaches or ensuring compliance, although these remain critical. The CISO's role is more about enabling responsible AI adoption that balances innovation with risk management, competitive advantage with ethical responsibility, and technological capability with organizational control.

This part of the report will address what the previous four chapters mean for today's CISOs. Each chapter will be summarized with a conclusion ('why it matters'), followed by some recommendations.

01

The first chapter describes the **AI-Powered Cyber Attacks** and shows how attackers can access AI capabilities. They can abuse commercial models, deploy self-hosted open-source models, or buy access to purpose-built malicious services. This chapter also describes that AI's role in malware development moved from experimental to operational and that AI can now be seen as a live attack operator. The most important element of this chapter is AI in vulnerability research, which resulted in a compressed patch window. AI is now capable of finding security flaws before they are exploited and finding ways to exploit them before they're fixed.

AI Why it Matters.

Attackers are now capable of leveraging AI for their benefits at an incredible speed. CISOs need to be aware of this development and should perceive AI as a live attacker. This means that CISOs need to revalidate their cyber security controls and check if their current security posture is capable of handling these AI-powered cyber-attacks. CISOs should also evaluate their current vulnerability management processes, because the increased speed of AI models finding and exploiting vulnerabilities significantly decreased the time to remediate.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 2026 CISO GUIDE

02

The second chapter examines attacks against AI, treating **AI as an attack surface**. It identifies two main causes. The first is unique to AI: language models can be manipulated through hidden instructions (direct & indirect prompt injection), trusted configuration files (configuration abuse), and corrupted memory (runtime poisoning). The second is more familiar: AI tools are still software and inherit the vulnerabilities of traditional applications. These weaknesses are spreading faster because of rapid AI adoption and are amplified by autonomous AI agents that hold excessive privileges and install trusted components with little or no human review. This is what drives attacks on AI infrastructure and the AI software supply chain.

AI Why it Matters.

One of the core principles of the CISO role is visibility across IT/OT infrastructure, workspaces, cloud environments, the supply chain, and AI. Visibility is essential because CISOs cannot protect what they cannot see. As AI becomes a new attack surface, security strategies must evolve to address it. This requires understanding AI-specific attacks, such as direct and indirect prompt injection, while also gaining visibility into the AI ecosystem, agentic supply chain, and AI infrastructure.

03

The third chapter describes **the effects of AI on digital trust and identities**. In previous chapters, we discussed AI as both a weapon and a target. This one turns to what AI does to trust between people due to its ability to forge a convincing human identity at scale. This chapter shows that attackers/criminals are now capable using AI to generate fake identities in text, audio and video. They can generate these fake identities offline, online and mostly fully autonomous.

AI Why it Matters.

Trust plays an essential role in cyber security, which plays an essential role for CISOs. These impersonation developments, as described in this chapter, can result in the fact that identity has become a broken trust anchor. Proving someone's identity remotely has become more difficult than ever before and in combination with the fact that social engineering has gone multi-channel, requires CISOs to adapt and stop taking identity for granted.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 2026 CISO GUIDE

04

The fourth chapter describes **data leakage and enterprise AI exposure.** While the previous chapters covered AI as a weapon, a target, and a tool for impersonation, this chapter showed the risks of AI adoption by enterprises. Check Point Research recorded significant growth in enterprise GenAI adoption. GenAI has evolved from a novelty emerging technology into an integral part of organizational productivity embedded across multiple business functions. Approx. 90% of organizations had at least one high-risk GenAI interaction each month. Even if organizations have strict rules about how its employees use AI, they still depend on external AI providers and applications, which can also serve as an entry point for threat actors. Check Point Research showed that an AI platform itself, like ChatGPT, can become the leak path, not just the apps connected to it. Once an organization's data is shared with an external AI model, they can't fully control where it ends up

Why it Matters.

Because AI adoption is outpacing the governance built to manage it, today's CISOs have a significant challenge. They should treat AI-related data exposure as an ongoing, permanent part of doing business with AI; not a one-time hurdle to clear during adoption. CISOs need to manage it well, which requires constantly monitoring how AI is actually being used, enforcing clear policies, training employees, and using real-time controls that can catch and stop sensitive information before it reaches an outside AI service. The companies that get the most out of AI will be the ones that keep their oversight growing right alongside their AI use, rather than treating security as something to revisit only once a year.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 2026 CISO GUIDE

CISO RECOMMENDATIONS

AI is significantly changing the world at a pace that we've never experienced before, and the same applies to AI security. We need to adapt faster than the 'bad guys' and as the saying goes: "it's not the strongest nor most intelligent species that survive. It's the one that is most adaptable to change".

This report provides the information that you – as a CISO – need to get the required adaptability to the fast-changing environment. With the insights of this report and the suggested Check Point products, you can become a successful CISO in protecting your organization and enabling the business.

In the last year, AI changed the (security) world significantly and brought more challenges for a CISO than ever before, but the basics still stayed the same. A CISO still needs to manage the risks, so that the business can do 'its business', which means a risk assessment still forms the foundation for a security strategy and roadmap. The fact that the CISO should also apply this risk assessment to their organization's AI ecosystem makes it perhaps more challenging, but not impossible.

01 INTRODUCTION

02 AI CYBER ATTACKS

03 AI ATTACK SURFACE

04 DIGITAL IDENTITY

05 AI DATA EXPOSURE

06 AI SECURITY FRAMEWORK

07 2026 CISO GUIDE

About Check Point Software Technologies Ltd.

Check Point Software Technologies Ltd. (www.checkpoint.com) is a leading protector of digital trust, utilizing AI-powered cyber security solutions to safeguard over 100,000 organizations globally. Through its Infinity Platform and an open garden ecosystem, Check Point's prevention-first approach delivers industry-leading security efficacy while reducing risk. Employing a hybrid mesh network architecture with SASE at its core, the Infinity Platform unifies the management of on-premises, cloud, and workspace environments to offer flexibility, simplicity and scale for enterprises and service providers.

Contact us

WORLDWIDE HEADQUARTERS
5 Shlomo Kaplan Street,
Tel Aviv 6789159, Israel
Tel: 972-3-753-4599
Email: info@checkpoint.com

U.S. HEADQUARTERS
100 Oracle Parkway, Suite 800,
Redwood City, CA 94065
Tel: 800-429-4391

UNDER ATTACK?
Contact our Incident Response Team:
emergency-response@checkpoint.com

CHECK POINT RESEARCH
To get our latest research and other exclusive content, Visit us at
www.research.checkpoint.com

www.checkpoint.com