

When Cyberattacks
Happen at
AI Speed

Booz Allen®



Executive Summary

Cyber conflict has entered a new phase.

Artificial intelligence is compressing the timeline of cyberattacks and reshaping how intrusions unfold. AI systems now plan, test, and execute attacks in rapid cycles—allowing adversaries to scan thousands of systems, build offensive tools faster, scale deception, exploit trusted identities, and even turn AI platforms themselves into new attack surfaces. Together, these changes are widening the cybersecurity speed gap between attackers and defenders.

Most defense strategies were not designed to operate at the speed of AI. This shift affects government and private enterprises alike. Federal agencies, defense systems, critical infrastructure operators, and large corporations now face the same reality. Weaknesses in commercial systems can become entry points into government missions and the reverse. The threat environment is shared, and so are the consequences.

The problem is time. Most cyber defenses still run on human timelines. Analysts review alerts, teams escalate incidents, and leaders weigh operational risk before approving containment actions. That process can take hours or days because it was designed for threats that unfolded slowly and allowed time for investigation. AI-enabled attackers move in minutes. When attackers move faster than defenders, they control the fight.

The gap between speed of attack and time to respond is widening. In 2025, the average breakout time from initial access to ability to move into other systems dropped to under 30 minutes, with the fastest cases measured in seconds. AI-enabled tools now automate reconnaissance, generate exploits, and scan thousands of systems simultaneously. Small teams—or even single operators—can run campaigns that once required large, coordinated groups of specialists.

AI is also accelerating how offensive tools are built. Operators define the objective and constraints, and language models generate code, test it, and refine it until it works. Development cycles that once took weeks now take hours. Social engineering has always worked; AI allows attackers to produce convincing emails, documents, and tailored personas at an industrial scale. Techniques that once remained inside elite units spread quickly across criminal ecosystems.

Defenders must close the AI speed gap.

THIS REPORT OUTLINES THREE DECISIONS REQUIRED TO DO SO.

First, move cyber defense to AI speed. Early containment— isolating systems, blocking malicious traffic, revoking suspicious sessions, and remediation—cannot wait for manual approval during an incident. Those actions must occur automatically within defined limits while the intrusion is unfolding.

Second, secure AI platforms as critical infrastructure. These systems increasingly concentrate on data, identity, and workflow authority—creating new entry points for attackers. Organizations need enforceable security baselines governing how AI systems are accessed, integrated, and monitored.

Third, adopt a human-AI teaming model. AI systems can draft detections, investigate alerts, and trigger containment or remediation in seconds while freeing analysts to focus on complex investigations and sophisticated campaign disruptions. Security experts and executive leadership must align on the actions they are willing to automate, share responsibility for the risk of disrupting operations, and help enable a culture of rapid yet thoughtful adoption as threats and operational considerations evolve.

The bottom line:

Cybersecurity has become a race against time. Attackers now operate at the speed of AI, probing thousands of systems and moving through networks in minutes. Defenders who cannot detect, contain, and remediate intrusions inside that window will lose control of their systems while the attack is still underway. The question is no longer whether organizations will face AI-enabled intrusions. They already do. The question is whether defenders can act in time, or only after the damage is done.

Introduction

Cybersecurity has entered the AI speed era. Artificial intelligence is accelerating how attacks are planned, built, and executed—compressing timelines that once stretched across days into minutes. This shift is changing how intrusions unfold and what organizations must do to defend themselves.

In 2026, AI-enabled threats are outpacing cyber defenses creating unprecedented risks to national security and economic stability. Government agencies, private companies, and the information technology (IT) and operational technology (OT) systems they depend on are all seen as one connected target by attackers.

Yet most cyber defenses still run on a human timeline: triage in hours, remediation in days, patching in weeks. Attackers now operate on a different clock, one that moves much faster.

The shift accelerating cyber threats is the rise of AI agents: software that can carry out multi-step tasks with minimal guidance. A human sets the objective. The AI agent chooses tools, runs actions, reads results, and keeps iterating until it reaches the objective.

This dramatically increases the pace and scale of attacks. AI agents can scan more targets, test more options, and move from a new idea to a working intrusion far faster than human operators alone. Large enterprises often run tens of thousands of endpoints and workloads, far more than traditional investigation and response processes can monitor in real time. Incidents can begin before patches are deployed. Even when defenders and attackers learn about vulnerabilities at the same time, AI-enabled attackers can exploit them within hours while defenders are still determining exposure and rolling out fixes.

This dynamic is the engine behind the cybersecurity speed gap. This shift changes what must occur in the first minutes of a cyber incident. Waiting for handoffs and patch rollouts is too slow. Containment must start immediately, within preapproved limits, while remediation efforts follow. That requires leaders to set rules in advance and trust automated systems to take the first steps safely before attacks can spread.

MASTER TIMELINE AI-DRIVEN CYBER ATTACKS 2023 - 2026 ALL NATIONS

MAY 2023

Storm-0558 Azure

AI-forged OAuth tokens breach Microsoft Azure. 25+ US gov orgs compromised.

NOVEMBER 2023

Cyberspike Founded

Domain registered. AsyncRAT + Mimikatz repackaged. Prototype of Villager begins.

2023-2024

AI Spearphishing Scale

MSS-linked groups use LLMs (GPT-4, Ernie) for mass AI-generated spearphishing.

NOVEMBER 2024

Big Sleep 0-Day

Google Gemini agent autonomously discovers SQLite zero-day. First AI-found CVE.

DECEMBER 2024

Treasury Breach

PRC APT breaches Treasury via BeyondTrust. 3,000+ files exfiltrated from OFAC.

2023

2024

INFLECTION POINT

JULY 2025

VILLAGER Released

Cyberspike launches AI-native pentesting tool. DeepSeek v3 + 4,201 exploit prompts.

SEPTEMBER 2025

Claude Code Attack

Jailbroken Claude Code executes autonomous campaign. 30 targets. Minimal human input.

AUGUST-SEPTEMBER 2025

HexStrike Weaponized

CVE-2025-7775 exploited in <10 min. 8,000+ endpoints. AI replaces human hackers.

OCTOBER 2025

APT27 + AI Phishing

DoJ indicts 12. DeepSeek spearphishing. 200+ AI fake content instances/month.

BARRIER TO ENTRY REMOVED

JANUARY 2026

AI Exploit at \$2.77

CVE-GENIE reproduces 51% of CVEs as working exploits. Claude 4.5 + GPT-5.2 chains.

2025

2026

MOVING AT THE SPEED AND SCALE OF AI

15 DAYS

CISA timeline to remediate
“Critical Vulnerabilities”

\$2.77

Average Cost to Auto Generate
CVE Exploit

60%

Percent of CISA defined “Critical
Vulnerabilities” unmitigated
after 15 days

500

Number of Zero Day exploits
Claude Opus 4.6 identified in
Open-source Code

10 MINUTES

Time for Hexstrike to exploit
CVE 2025-7775 to exploit
8,000+ Endpoints

This threat report examines how artificial intelligence is reshaping the cyber threat environment and what organizations must do to respond.

Part 1 explains how AI is widening the cybersecurity speed gap by compressing attack timelines, expanding offensive capability, scaling deception, exploiting trusted identities, and turning AI systems themselves into new attack surfaces.

Part 2 outlines three decisions leaders must make to close that gap: move cyber defense to AI speed, secure AI platforms as critical infrastructure, and adopt a human–AI teaming model that allows defenders to operate at machine speed.

Leaders must decide how much risk they are willing to accept while these changes are put in place. As long as attackers operate at AI speed while defense processes remain human-paced, risk remains high. Closing the speed gap between attack and defense will take clear ownership and coordination for both government and industry.

Findings at a Glance

Cybersecurity has become a race against time.

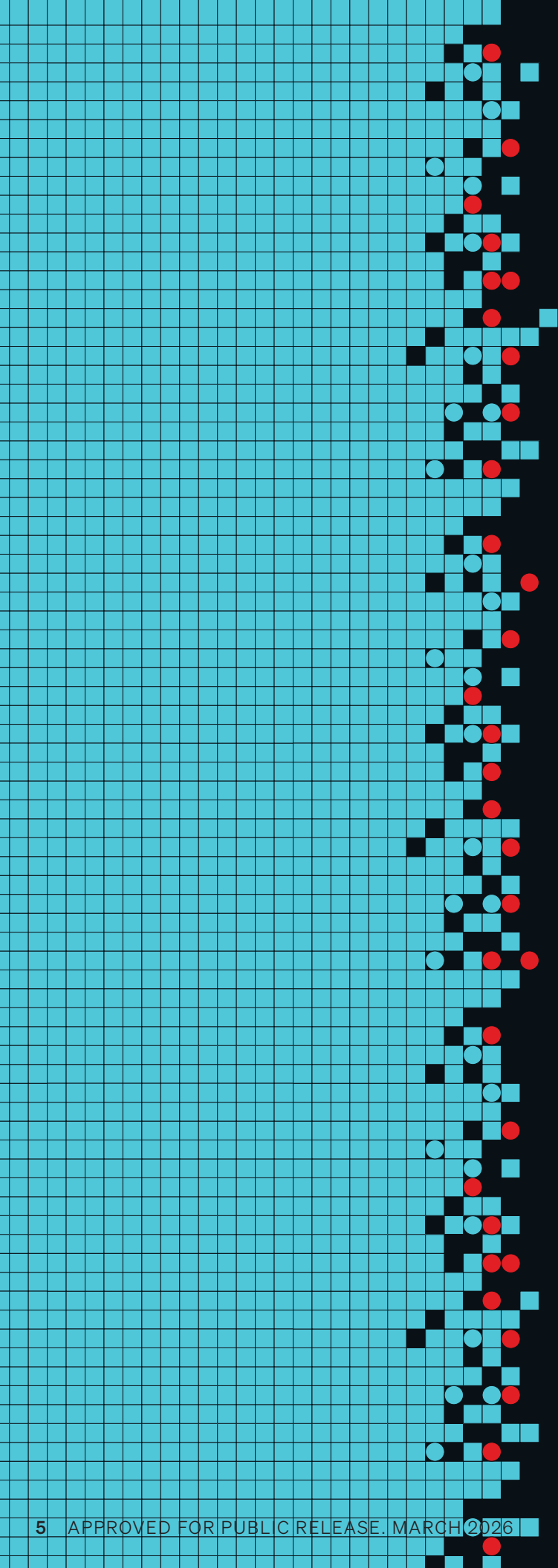
Artificial intelligence is now compressing the timeline of cyberattacks. Intrusions that once unfolded over days can now cause operational impact in minutes. Attackers can probe thousands of systems at once, turning weaknesses in government networks, commercial platforms, and critical infrastructure into immediate operational risk.

Speed decides who wins. AI-enabled attackers move faster than defense processes built around human investigation and manual approval. When containment starts hours after an intrusion begins, attackers control the fight.

AI platforms have become core infrastructure. They concentrate sensitive data, identity systems, and operational workflows in a single place. If they are misconfigured or compromised, attackers gain direct reach into the systems organizations depend on to operate.

To prevail, fight AI with AI. Attacks now escalate at AI speed, while most security teams still respond at human speed. The first containment steps must happen automatically. Detection and investigation must occur in seconds while automated systems triage alerts, block malicious activity, and isolate compromised systems so analysts can focus on remediation.

Defenders must move to a human-agent teaming model. Automated agents handle routine investigation and containment at AI-speed. Human experts supervise and intervene when judgment is required. Done well, this expands the capacity of a security team by ten or even one hundred times.



PART 1:

How AI Creates the Cybersecurity Speed Gap

AI is accelerating attacks beyond the reach of human-only defenses.

Minutes beat hours. When attackers can move in minutes and detection and containment take hours, attackers control the fight. AI is shrinking the time it takes attackers to find targets, break in, and act after access. Steps that once took days can now run in minutes. Tools that once required coordinated teams working for weeks can now be produced by small groups in days. Content used for deception tactics like phishing, which once required manual effort, can now be generated at scale.

Human-paced defense can't keep up with that speed or scale.

AI COMPRESSES TIME

Attacks that once took days now unfold in minutes.

AI is cutting the time state-backed actors and criminals need to plan, build, and refine intrusions. Two patterns are driving this shift. The first is **collaborator use**. An operator works interactively with a model to write scripts, debug code, and adapt tools when something fails. Rapid back-and-forth with the model speeds up development and reduces the time needed to produce working exploits. A second pattern is **orchestration**. An operator connects an AI system to offensive tools, points it at a target, and sets limits. The system then works the problem on its own. It takes an action, checks the result, and takes the next action.

Together, these approaches allow attackers to operate across many targets at once. A single operator using agentic tooling can run reconnaissance, exploitation, and follow-on actions across dozens of targets at once. That level of parallel activity used to require several specialists to achieve but now can be done by one person.

Documented Cases

- **February 2025:** OpenAI described Democratic People's Republic of Korea (DPRK)-linked actors using models to help develop intrusion tooling.ⁱ
- **July 2025:** The Villager penetration testing framework combined Kali Linux tools with People's Republic of China (PRC)-developed large language models (LLMs) to automate end-to-end offensive workflows.ⁱⁱ
- **August 2025:** Within hours of Citrix disclosing NetScaler zero-days, forum posts discussed using HexStrike-AI to scan for vulnerable systems and attempt exploitation at scale, reducing workflows that once took days to under 10 minutes.ⁱⁱⁱ
- **November 2025:** Anthropic reported a PRC-attributed campaign using agents to automate 80% to 90% of its offensive workflow, enabling attempts across roughly 30 targets at a rate impossible to execute manually.^{iv}
- **February 2026:** CrowdStrike reported average cyber-criminal breakout time in 2025 fell by 65% to 29 minutes, with the fastest observed breakout at 27 seconds.^v

2025 VILLAGER RELEASED • HEXSTRIKE WEAPONIZED • AI GOES FULLY AUTONOMOUS • AI EXPLOIT INDUSTRIALIZATION

**2025 &
2026
AI CYBER
GOES AUTO**

11 K+

Villager Downloads
60 Days

4,201

AI Exploit Prompts

8 K+

HexStrike Victims

\$2.77

Average Cost to
Auto-Generate
CVE Exploit

JULY 2025

Villager—World's First PRC AI-Native Pentesting Tool **Actor: Cyberspike/HSCSEC (PRC)**

Cyberspike publishes Villager on PyPI—the first AI-native, fully-autonomous penetration testing framework from a PRC-linked actor.

11,000+ downloads in 60 days. Confirmed wild use via VirusTotal. Described by researchers as 'Cobalt Strike's AI-native successor.' Author @stupidfish001 traces to HSCSEC CTF hacker pipeline.

SEPTEMBER 2025

World First: Claude Code Autonomous Attack **Actor: PRC State Actor (Unit Unknown)**

PRC state-sponsored hackers jailbreak Anthropic's Claude Code agentic AI tool.

The weaponized AI agent autonomously executes a complete cyber kill chain against 30 global targets—including major technology companies, financial institutions, chemical manufacturers, and government agencies—with minimal human involvement at any stage.

This is the first confirmed large-scale cyberattack in history where an AI agent independently selected targets, generated exploits, executed intrusions, exfiltrated data, and installed persistence—all without real-time human direction.

Documented by Anthropic's Threat Intelligence team in November 2025.

AUGUST–SEPTEMBER 2025

HEXSTRIKE AI v6.0—CVE Weaponized in <10 Minutes **Actor: Dark-web actors (PRC-adjacent)**

HexStrike AI v6.0 launches as open-source framework with MCP Server architecture orchestrating 150+ security tools and 12+ specialized AI agents.

Within hours of Citrix NetScaler CVE-2025-7775 (Critical RCE) public disclosure, dark-web threat actors weaponize HexStrike to exploit the vulnerability at scale—reducing exploit time from days to under 10 minutes.

8,000+ endpoints compromised. Check Point Research: 'this is the future of mass exploitation.'

JANUARY 2026

AI Exploit Chains at \$50/Run—OPUS 4.5 + GPT-5.2 **Actor: Sean Heelan (Research—UK)**

Security researcher Sean Heelan publishes research demonstrating that AI agents using Claude Opus 4.5 and GPT-5.2 can autonomously generate complete, working exploit chains for real-world zero-day vulnerabilities.

In the QuickJS JavaScript engine zero-day test, AI agents produced 40+ distinct exploits across six increasingly constrained scenarios—including defeating hardware-enforced shadow stacks, CFI, and seccomp sandbox escapes. GPT-5.2 solves every scenario. Average time: 45 minutes per full exploit chain. Average cost: ~\$50/run.

Implication: Nation-state-grade exploit development is now accessible to any actor with an API key.

AI EXPANDS OFFENSIVE CAPABILITY

Small teams can build tools that once required large development groups.

AI lets small teams and even single operators build malware and offensive tools that once required large, coordinated development. An operator describes what they want and the model generates code, tests it, and revises it in rapid cycles. The time from idea to working tool shrinks dramatically.

As a result, teams of all sizes can create complex malware, including modular frameworks and kernel-level implants.

Fast coding changes the problem defenders face. Attack volume increases, as attackers rapidly iterate with new and improved tools. Controls that rely on manual review or periodic scanning fall behind. Attackers also become more resilient. They can quickly create multiple ways to exploit a system, including developing alternatives if one route is blocked.

AI shortens the window between vulnerability disclosure and exploitation. Attackers and defenders often learn about new weaknesses at the same time. But attackers can act immediately while organizations on the defense still need time to develop, test, and deploy patches.

AI SCALES DECEPTION

Convincing lures can be generated and deployed at industrial scale.

Phishing and other deception-based attacks remain among the most effective ways to penetrate government and commercial networks. AI-accelerated deception is making these attacks easier to produce and harder to filter.

Attackers can now generate convincing lures at scale. Models produce tailored messages, decoy documents, and personalized outreach that push people to click, open, approve, or install. Social engineering succeeds more often, including against users trained to spot it.

As the messages improve, training and basic filtering miss more of them. What matters is what the click is allowed to do. If a file, link, add-in, or script can run without restriction, a lure becomes a way to run code. If accounts have too much access by default, one click can turn into stolen passwords, stolen session keys, or malware installation.

This is why controls must block risky code and limit access. Block scripts, macros, and add-ins by default where possible. Limit what users can install. Always-on admin access should be reduced. Require extra checks for new logins and privilege changes. Security cannot depend on training users to not click.

Documented Cases

- **Q4 2025:** Sonatype reported nearly 400,000 new open-source malware packages in Q4 2025, representing 89% of all malicious packages it observed that year. The surge was attributed to scripted and AI-assisted publishing at scale by a single campaign.^{vi}
- **January 2026:** Check Point Research described VoidLink, a large modular malware framework assessed as likely developed through spec-driven, AI-assisted workflows. A recovered test artifact suggested the code exceeded 88,000 lines roughly a week after development began.^{vii}
- **January 2026:** An experiment using commercial language models produced 40+ working exploits for a previously unknown JavaScript-interpreter vulnerability, often in under an hour and at a cost of only tens of dollars in model usage.^{viii}

These timelines weaken the traditional link between sophistication and team size. Capabilities spread faster as a result. Techniques pioneered by top-tier actors can quickly appear in commodity campaigns, widening exposure across organizations of every size and sector.

Zero trust programs already move in this direction by assuming networks and devices are breached and limiting access based on identity, device health, and context. As AI accelerates attack speed and lateral movement, implementing these controls at scale becomes even more important.

Documented Cases

- **April 2025:** SentinelLabs reported AkiraBot using commercial AI services to generate website-specific pitches and submit them through contact forms and chat widgets at scale.^{ix} The campaign targeted at least 80,000 sites beginning in September 2024.
- **Summer 2025:** Safety observed an AI-generated malicious software package that reached 1,500 downloads and 19 versions in two days. The package included professional documentation assessed as likely AI-authored.^x
- **December 2025:** Intezer traced a Paper Werewolf campaign that used AI-generated decoy documents to lure targets into opening an Excel extension. The add-in installed a backdoor and enabled remote command execution and file transfer.^{xi}

AI EXPLOITS TRUSTED IDENTITIES



Attackers increasingly operate through legitimate accounts and systems.

The problem is shifting from spotting malicious files to spotting malicious use of legitimate accounts. Once attackers gain access, they often operate through real user identities rather than obvious malware or malicious infrastructure. The challenge shifts from spotting malicious files to spotting malicious behavior carried out through trusted accounts.

Detection then becomes a different problem: spotting an attacker using a real account. That activity can look normal at first glance. Security teams get less value from older tripwires like virus signatures, blacklisted IPs, blocked domains, or known malicious files. The signals that matter are behavioral: what the account is accessing, where it is logging in from, when it is active, and whether privilege changes match the role.

AI also makes fake candidates look real. It is easier to create convincing resumes, profiles, documents, and interview responses. As a result, more attempts can pass screening and remain credible long enough to gain access to internal systems.

Documented Cases

- **June 2025:** Microsoft Threat Intelligence reported Jasper Sleet, a North Korean remote IT worker operation, using AI to improve fake personas, including altered documents, improved profile photos, and voice-changing software to pass screening and onboarding.^{xii}
- **December 2025:** Flashpoint described the DPRK “Contagious Interview” campaign, where operators used common AI tools to draft resumes and emails and to support coding tasks during technical interviews.^{xiii}

“The challenge shifts from spotting malicious files to spotting malicious behavior carried out through trusted accounts”

Brad Medairy, Executive Vice President

AI EXPANDS THE ATTACK SURFACE



AI platforms and workflows are becoming new entry points for attackers.

AI platforms themselves are now targets. Attackers can break into the tools that connect AI to internal systems, such as workflow tools and add-ons. They can also use commercial AI services to plan and run their campaigns. Even without a full breach, faulty configuration or misuse can expose sensitive information.

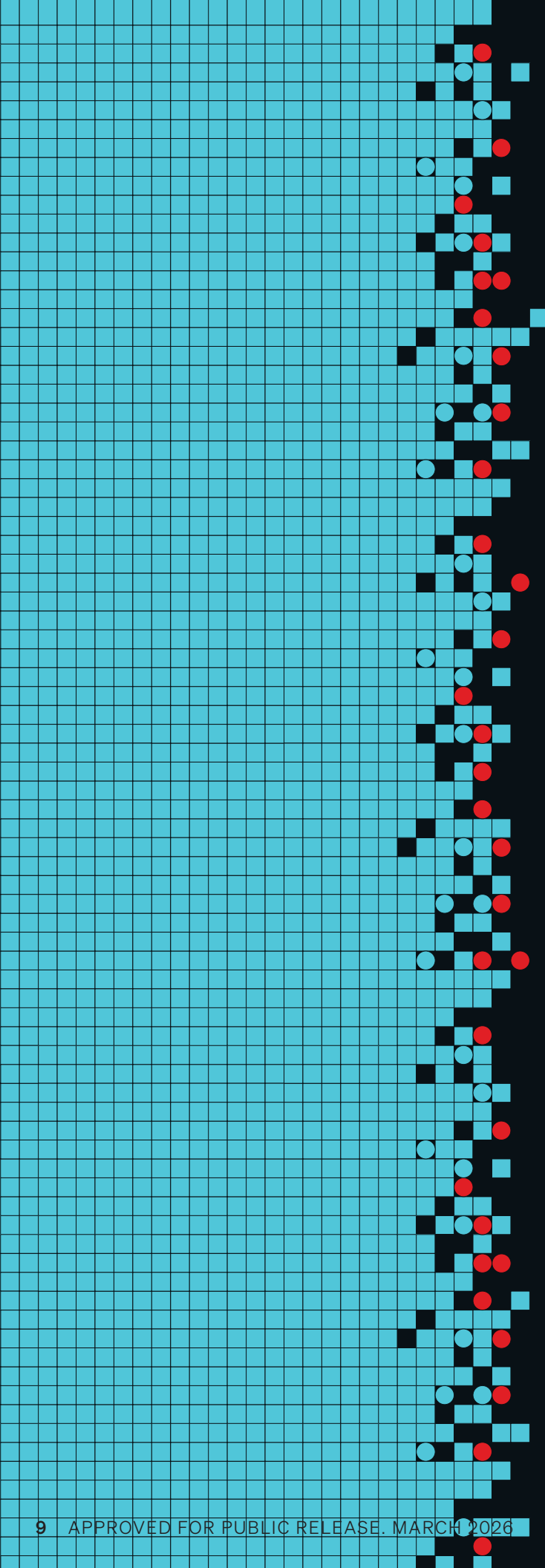
One emerging risk involves hidden instructions embedded in trusted content. An attacker plants a small set of instructions inside a document, email, ticket, or web page. When an AI assistant reads that content, the instructions can influence how it behaves. If the assistant is connected to internal data or can take actions, it can be pushed to reveal information or do things it is allowed to do.

The risks grow when no single team owns AI security end to end. One team runs models and workflows. Another manages access and logs. A third team manages vendors. Gaps appear when no single owner is accountable for what the AI can reach, what it is allowed to do, and how misuse is detected and stopped.

Documented Cases

- **Workflow tool exploitation:** XLab reported Pickai malware being spread through vulnerabilities in ComfyUI, an AI workflow tool, affecting nearly 700 servers worldwide.^{xiv}
- **Weaponized add-ons:** Public repositories have been used to distribute malicious packages and extensions, sometimes paired with polished, AI-written documentation to appear legitimate.^{xv}
- **Commercial AI APIs used in operations:** Microsoft Incident Response documented attackers using the OpenAI Assistants API to pass instructions and receive results.^{xvi}
- **Models expose confidential information:** Microsoft’s “Whisper Leak” research reported that traffic patterns can reveal what an AI system is being asked about, even when the traffic is encrypted.^{xvii}

These trends are widening the cybersecurity speed gap. Attackers now operate at machine speed while most defenses still depend on human timelines. Closing that gap requires fundamental changes in how organizations defend their systems.



PART 2:

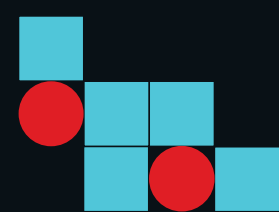
Closing the Cybersecurity Speed Gap

Cyber operations are accelerating.

Attackers can move from discovery to impact faster than human approvals and routing can support. Closing the gap requires three decisions.

DECISION 1:

MOVE CYBER DEFENSE TO AI SPEED



Why This Matters

Cyber defense still runs on human timelines. Security teams triage alerts in hours, remediate incidents in days, and deploy patches over weeks. Those timelines were built for threats that unfolded slowly.

AI-enabled attackers compress time-to-impact and increase the number of parallel attempts an adversary can run. They can move from discovery to operational impact in minutes and run multiple intrusion attempts in parallel. By the time analysts complete their first investigation, attackers may already be moving laterally through the network.

When attackers move faster than defenders can decide, attackers control the fight.

What Must Change

Containment must begin while the intrusion is still unfolding. Waiting for analysts to complete investigations or leadership to approve response actions is too slow.

Organizations should preapprove a set of containment actions that can execute automatically when evidence meets

defined thresholds. These actions may include isolating compromised hosts, blocking malicious traffic across a network, revoking suspicious sessions, and freezing risky privilege changes.

AI-speed cyber defense requires authority, boundaries, auditability and converging, and AI-enabling the NOC and SOC to eliminate stovepipes and increase response speed and effectiveness.

Automation will create false positives. But when critical mission or business operations are at risk, the tradeoff favors rapid containment or remediation over perfect certainty. Automated actions should operate within clear boundaries, include rollback mechanisms, and produce detailed logs for investigation and after-action review.

Legal authorities, oversight structures, and procurement priorities must support this shift. Organizations should prioritize tools that enable automated containment, enforce policy at scale, and provide auditability for every automated decision.

Implement zero trust controls across the enterprise to reduce the attack surface, limit lateral movement, detect vulnerabilities and misconfigurations in real time and advance AI-enabled defensive cyber operations.

OPERATING MODEL



AI speed defense requires converging network and security operations and enabling both with AI. A unified AI-enabled NOC/SOC operating model reduces delays caused by organizational handoffs and allows containment and remediation actions to execute immediately—more rapidly detecting, responding to, and remediating threats and outpacing attackers.

Target Operating Model: Containment must start while an intrusion is still unfolding. Two delays usually prevent that. First, new detections take too long to reach production. Second, routine alerts take too long to clear before the first containment or remediation action occurs.

AI agents can close both gaps.

AI agents can close the first gap by keeping detections current by drafting new rules, testing them, and pushing updates through an approved release process. As new threats and

vulnerabilities appear, agents can draft detections, test them, and push updates through an approved release process. Coverage improves quickly, updates fast, without waiting weeks for manual detection engineering.

AI agents close the second gap by clearing routine alerts at AI speed. They pull the relevant context, weed out duplicates, and trigger preapproved first containment when evidence is strong. Analysts then focus on complex investigations, threat hunting, and improving detection logic based on what they learn. First action does not wait on handoffs and approvals.

This is how defenses stay current and start containment inside the attacker's window while analysts focus on complex investigations, hunting, and improving detections and limits based on results.

Without this shift, networks will remain tuned for a threat environment that no longer exists.

Why This Matters

AI platforms are quickly becoming core infrastructure for government and commercial organizations. They hold sensitive data, connect to email and ticketing systems, integrate with code repositories and trigger actions through plug-ins, agents, and automated workflows.

Those connections make them powerful and attractive targets. Adversaries already use commercial AI services to speed tool building and planning. When enterprise AI platforms are connected into internal systems, it can also be abused to support software supply chain compromise, remote control of systems, and long-term access inside trusted environments.

The risk is systemic because these platforms sit close to the highest-trust parts of the enterprise. They interact with mission data, identify systems, and software development pipelines. A compromise can expose credentials and API keys, spread across organizations, and give an attacker leverage over the tools used to build, test, and deploy software.

What Must Change

Voluntary guidance does not match this level of risk these platforms introduce. AI platforms that serve federal users and critical sectors need enforceable security requirements.

Organizations should establish defined, testable security baselines before treating AI platforms as trusted enterprise systems. At a minimum, those baselines should include strong authentication, comprehensive activity logging, secure handling of secrets, strict controls on plug-ins and integrations, and secure-by-default configurations.

Leaders must also make deliberate decisions about operational risk. Which workloads can use commercial AI services? Which workloads must run in hardened environments? Which workloads require government-controlled systems because the data sensitivity and mission impact it too high?

OPERATING MODEL

Securing AI platforms requires closing the gaps before, during, and after deployment. While agentic systems introduce novel risks, most of the current known risks can be effectively addressed by adapting and rigorously applying existing NIST security frameworks.

Where agent-unique behaviors create genuine gaps, such as prompt integrity, behavioral bounds enforcement, agent identity, and non-deterministic forensics, a concise set of targeted new controls can complement, not replace, SP 800-53 Rev.5 and Zero Trust Architecture.

In addition, security teams can take these steps:

1. Scan systems before deployment for known weaknesses and unsafe configurations.
2. Conduct adversarial testing in a realistic staging environment before production.
3. Run them in production with guardrails and continuous monitoring, so misuse is detected early and risky actions can be blocked.

During operation, AI workflows should run with guardrails and continuous monitoring so misuse is detected early, and risky actions can be blocked.

Each automated workflow should have its own identity and access controls. Zero trust policies should limit what data and systems it can reach. Detailed logging and operational telemetry—including tool calls, key events, and performance signals—should allow teams to quickly isolate affected workflows and assess impact during incidents.

DECISION 3:

ADOPT A HUMAN-AI TEAMING MODEL

Why This Matters

Most cybersecurity operations still run on human timelines. Analysts review alerts, investigate incidents, and coordinate response actions step by step. That model assumes humans can keep pace with attackers.

AI-enabled threats break that assumption.

Adversaries can now automate reconnaissance, exploitation, and follow-on actions at scale. They run many attempts in parallel and iterate rapidly. The time between discovery and operational impact is shrinking from days to minutes. In this environment, organizations can't rely on manual triage and approval chains to stop attacks.

What Must Change

Defenders must scale their own operations.

A human-AI teaming model allows a single defender to oversee far more activity than human only processes allow. Automated agents handle routine tasks such as alert triage, signal correlation, detection updates, and first containment actions. Tasks that once required teams of analysts can occur automatically within seconds.

Humans remain essential, but their role changes. Analysts supervise automated systems, refine detection logic, and intervene where incidents require judgement or broader context. One responder can oversee many automated investigations and containment actions simultaneously.

Leadership must establish clear authority and decision ownership. When it is unclear who can approve automated actions, containment stalls. Security teams hesitate to disrupt operations and approval chains designed for slower threats add delays when minutes determine outcomes.

Executive leadership must own the disruption tradeoff. The security team cannot be the sole owner of containment decisions. Leaders must define in advance which actions automation can take, when human approval is required, and who has authority to expand containment. These decisions should be tested under realistic conditions so teams can act immediately when incidents occur.

OPERATING MODEL

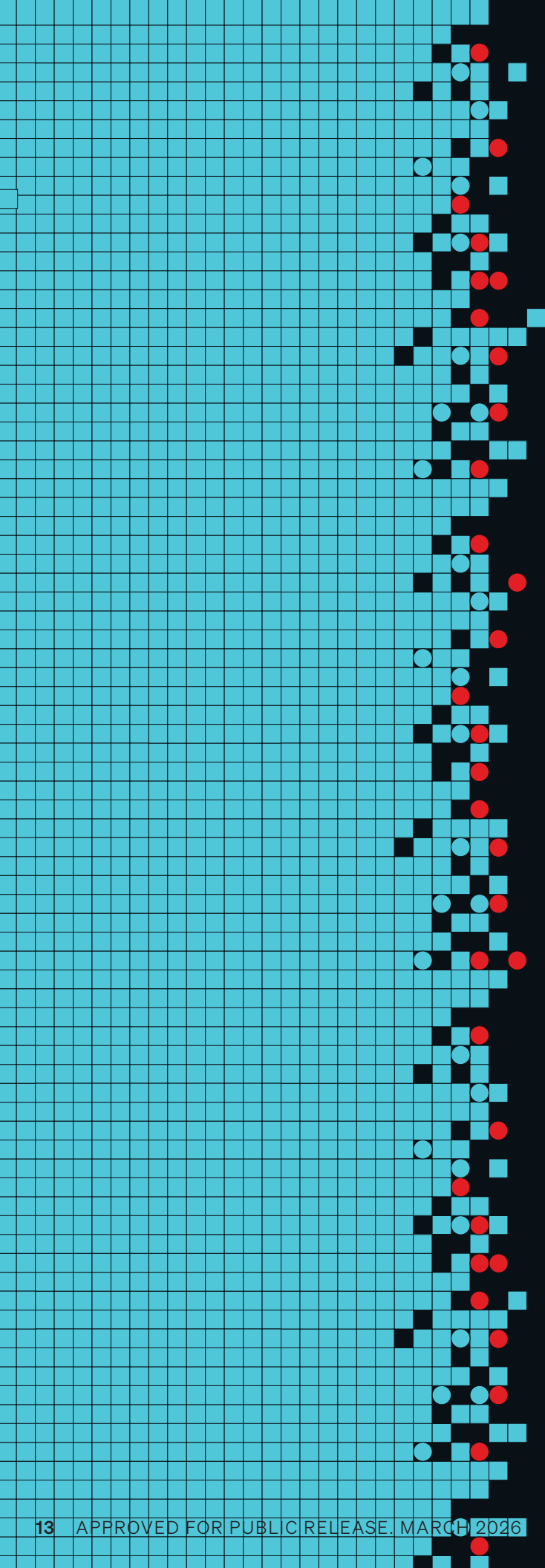
Building a human-AI operating model requires deliberate transformation.

Organizations should begin with a clear strategy that defines objectives, risks, leadership accountability, and the intended end state. Workforce roles must then be aligned with AI capabilities. Analysts supervise automated investigations while agents perform routine detection, triage, and containment tasks. Policies should define what actions automation can take, what limits apply, and which workflows require human approval.

Automated playbooks should be introduced gradually. Start with repeatable actions where evidence is strong and changes can be reversed. Test these workflows through tabletop exercises and operational simulations to validate decision paths and authority structures.

Automation must remain observable and reversible. Systems should maintain detailed logs, defined rollback paths, and clear override authority so automated actions can be trusted in real operational conditions.

Organizations that move people into roles where they design, direct, and supervise AI-speed operations can contain threats inside the attacker's operational window. Those that do not will fall behind.



CONCLUSION

Speed now
determines
security
outcomes.

Adversaries can plan, execute, and iterate attacks in minutes. Many defenders still operate on timelines measured in hours or days. But they have access to many of the same AI capabilities. The limiting factor is whether they can modernize their systems and change how they defend fast enough to constrain attacker options before significant damage occurs.

Winning in this environment requires transforming cyber defense and IT and OT operations to operate at AI speed. To close the gap, defenders must make five shifts:

1. Understand the persistence, scale, and speed of the threat—and the risks it creates for your enterprise.
2. Implement zero trust controls at scale to reduce the attack surface and limit lateral movement and advance AI-enabled defensive cyber operations.
3. Transition to converged, AI-enabled NOCs and SOCs to accelerate detection, response, and remediation.
4. Treat AI systems and AI-enabled workflows as mission-critical infrastructure with enforceable security baselines for access, logging, integrations, and data handling, leveraging leading cybersecurity standards and new controls as needed.
5. Optimize human–AI teaming so defenders can operate effectively at AI speed.

Cyber defenses are being tested like never before by AI-enabled threats. Organizations that make these shifts can contain attacks early and limit damage. Government and commercial organizations still have an opportunity to outpace these threats—if they act now. Those that do not will continue to detect intrusions only after attackers have already established control.

Cybersecurity has become a race against time.

Attackers already operate at AI speed. The question for defenders is whether to meet them there.

Notes

- i <https://cdn.openai.com/threat-intelligence-reports/disrupting-malicious-uses-of-our-models-february-2025-update.pdf>
- ii <https://www.straiker.ai/blog/cyberspike-villager-cobalt-strike-ai-native-successor>
- iii <https://blog.checkpoint.com/executive-insights/hexstrike-ai-when-llms-meet-zero-day-exploitation/>
- iv <https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf>
- v <https://www.crowdstrike.com/explore/2026-global-threat-report/2026-global-threat-report>
- vi <https://www.sonatype.com/blog/open-source-malware-in-dex-q4-2025-automation-overwhelms-ecosystems>
- vii <https://research.checkpoint.com/2026/voidlink-early-ai-generated-malware-framework/>
- viii <https://sean.heelan.io/2026/01/18/on-the-coming-industrialisation-of-exploit-generation-with-llms/>
- ix <https://www.sentinelone.com/labs/akirabot-ai-powered-bot-bypasses-captchas-spams-websites-at-scale/>
- x <https://getsafety.com/blog-posts/threat-actor-uses-ai-to-create-a-better-crypto-wallet-drainer>
- xi <https://intezer.com/blog/tracing-a-paper-werewolf-campaign-through-ai-generated-decoys-and-excel-xlls>
- xii <https://www.microsoft.com/en-us/security/blog/2025/06/30/jasper-sleet-north-korean-remote-it-workers-evolving-tactics-to-infiltrate-organizations/>
- xiii <https://flashpoint.io/blog/beyond-the-malware-digital-empire-north-korean-threat-actor/>
- xiv [https://blog.xlab.qianxin\[.\]com/pickai-the_backdoor_hiding_in_your_ai_stack/](https://blog.xlab.qianxin[.]com/pickai-the_backdoor_hiding_in_your_ai_stack/)
- xv <https://www.getsafety.com/blog-posts/threat-actor-uses-ai-to-create-a-better-crypto-wallet-drainer>
- xvi <https://www.microsoft.com/en-us/security/blog/2025/11/03/sesame-novel-backdoor-uses-openai-assistants-api-for-command->
- xvii <https://www.microsoft.com/en-us/security/blog/2025/11/07/whisper-leak-a-novel-side-channel-cyberattack-on-remote-language-models>
- xviii <https://www.trendmicro.com/vinfo/us/security/news/cyber-crime-and-digital-threats/the-state-of-criminal-ai>

About Booz Allen

Booz Allen is the advanced technology company delivering outcomes with speed for America's most critical defense, civil, and national security priorities. We build technology solutions using AI, cyber, and other cutting-edge technologies to advance and protect the nation and its citizens. By focusing on outcomes, we enable our people, customers, and their missions to succeed, accelerating the nation to realize our purpose: Empower People to Change the World®.