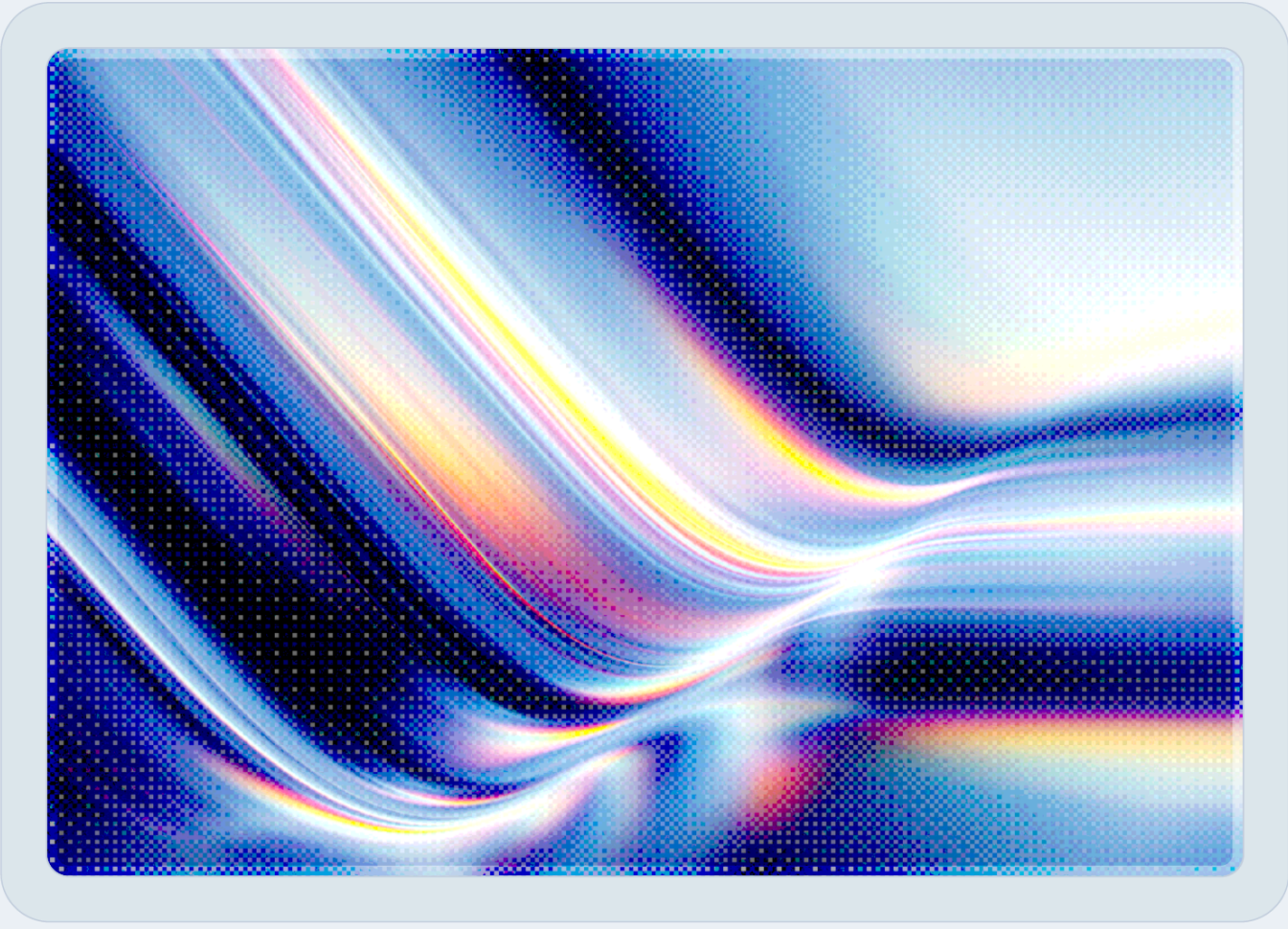
CSAI Foundation | Cloud Security Alliance

LLM-Accelerated Attack Pipelines: AI Agents as Offensive Force Multipliers

How Autonomous AI Is Reshaping the Offensive Threat Landscape and What Defenders Must Do Now

2026-05-28

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Table of Contents

Executive Summary

Artificial intelligence has arrived on the offensive side of the security boundary faster than most enterprise security programs anticipated. Large language models and autonomous AI agents — systems originally designed to assist researchers, developers, and knowledge workers — are now integral components of sophisticated attack pipelines. This shift is not hypothetical: documented incidents from 2024 through the first quarter of 2026 confirm that threat actors spanning nation-states, organized ransomware groups, and low-skill opportunists have each adapted AI to dramatically reduce the cost and increase the scale of their operations.

The core dynamic is asymmetric amplification. An AI agent that autonomously conducts open-source intelligence gathering, drafts targeted spear-phishing content, identifies exploitable misconfigurations, and adapts its behavior based on environmental feedback allows a small team of attackers to operate at the scale of a much larger adversary. University of Illinois Urbana-Champaign researchers demonstrated in 2024 that GPT-4 could autonomously exploit 87 percent of a curated set of 15 tested one-day vulnerabilities at an estimated cost of under nine dollars per successful exploit [1]. By early 2026, this research-level capability had migrated into criminal operations: a financially motivated threat actor used AI-assisted tooling to compromise more than 600 FortiGate devices across 55 countries in a five-week campaign without exploiting a single previously undisclosed vulnerability — relying instead on AI-assisted reconnaissance and credential harvesting to accomplish at scale what would previously have required a coordinated team [2].

This whitepaper analyzes the structural ways in which LLMs and agentic AI systems transform each phase of the offensive lifecycle, reviews the documented evidence of threat actor adoption, and provides concrete recommendations for defenders. It also identifies the governance and detection capabilities that organizations must build now, before AI-accelerated offense becomes the norm rather than the exception. The CSA AI Controls Matrix (AICM), MAESTRO threat modeling framework, and Agentic AI Red Teaming Guide each provide applicable controls and methodologies that security programs can deploy today.

1. Introduction: The Tipping Point

For most of AI's commercial history, discussions of security implications centered on defense: using machine learning to detect anomalies, classify malware, and correlate threat signals faster than human analysts could manage. That framing, while still valid, has become insufficient to characterize the current risk environment. The same capabilities that make LLMs powerful reasoning assistants — long-context understanding, code generation, tool use, and iterative planning — make them valuable components of offensive infrastructure.

The transition from theoretical concern to operational reality accelerated sharply in 2024 and 2025. Academic researchers demonstrated that frontier models could autonomously chain multi-step exploit sequences [1]. Security practitioners building AI-powered penetration testing tools found that well-prompted LLM agents could conduct credible reconnaissance, generate custom payloads, and identify lateral movement paths with minimal human direction [3]. And by late 2025, threat intelligence reporting documented state-sponsored actors using AI coding agents to conduct autonomous espionage campaigns, handling up to 90 percent of tactical operations without continuous human oversight [4].

The Europol Internet Organised Crime Threat Assessment (IOCTA) 2026 report characterized this as a shift toward the "industrialization of cybercrime," noting that AI has lowered both the skill threshold and the per-attack cost of sophisticated offensive operations [5]. Ransomware groups that once required specialized expertise to conduct targeted intrusions can now deploy AI agents to automate the research, reconnaissance, and spear-phishing stages, reserving human operator time for the highest-value decisions. Industry threat intelligence reporting observed a sharp escalation in AI-assisted intrusion activity through 2025, with multiple sources documenting that AI-enhanced techniques had become operational across threat actor categories ranging from nation-states to commodity ransomware operations [4][5].

These are not simply incremental improvements to existing attack techniques. They represent a qualitative change in the threat landscape: one in which the limiting factor for many attack operations is no longer human skill or labor, but access to AI services and compute. Understanding this shift — and responding to it — is the purpose of this whitepaper.

2. How AI Transforms Each Phase of the Attack Lifecycle

The traditional kill chain — reconnaissance, weaponization, delivery, exploitation, installation, command-and-control, and action on objectives — provides a useful lens through which to examine AI's offensive impact. AI does not replace the kill chain; it compresses and automates it, reducing the time between target selection and impact from days or weeks to hours.

2.1 Reconnaissance: AI-Powered OSINT at Scale

Open-source intelligence gathering has historically been labor-intensive. Building a useful profile of a target organization requires correlating data from corporate websites, LinkedIn profiles, GitHub repositories, job postings, technical forums, breach databases, and passive DNS records. This work is well within the capability of AI agents operating with web-access tools. An agent instructed to profile a target organization can systematically scrape and synthesize public sources, identify employee names and email formats, map technology stacks from job postings, extract organizational structure from LinkedIn, and surface recent press releases or regulatory filings that indicate operational vulnerabilities — in minutes rather than hours [7].

The implications extend beyond speed. AI-powered reconnaissance produces more complete target profiles than most human operators would compile under time pressure. An agent will not skip a source because it seems unlikely to yield results; it will not miss the job posting that reveals an unpatched VPN product still in use, or the developer's GitHub commit that exposes an internal API endpoint. The thoroughness that AI brings to OSINT collection directly increases the precision and success rate of subsequent attack phases.

Nation-state groups and sophisticated ransomware operators have documented this capability most clearly, but the tools required are widely accessible. Commercial AI services with browsing capabilities, combined with relatively simple orchestration scripts, are sufficient to automate the reconnaissance phase of an attack against most organizational targets.

2.2 Weaponization: Automated Exploit Development and Payload Generation

Weaponization — converting a known vulnerability or misconfiguration into a working exploit — has traditionally required specialized skill. UIUC researchers quantified how AI has changed this calculus: when provided with CVE descriptions, GPT-4 autonomously exploited 87 percent of a curated set of 15 tested

one-day vulnerabilities, while no other model they tested succeeded more than a small fraction of the time [1]. Critically, the cost was approximately $8.80 per successful exploit, a figure that makes systematic automated exploitation economically viable against a large attack surface.

The practical implication is that the window between public vulnerability disclosure and attacker exploitation is shrinking. Security teams have historically operated under the assumption that they have days or weeks after a CVE is published before targeted exploitation becomes widespread. AI-assisted exploit development narrows that window significantly. An AI agent that monitors vulnerability feeds, ingests advisories, and generates working proof-of-concept code can stage exploitation attempts faster than most patch deployment cycles permit.

Beyond CVE exploitation, AI has accelerated the development of novel malware. Trend Micro's 2026 threat predictions report documented that LLMs can generate functional malware variants that evade detection signatures [8], and that polymorphic payload generation — in which malware mutates its own code between deployments — is achievable with relatively straightforward prompting. While responsible AI providers maintain guardrails against obvious misuse, jailbreaking techniques, model fine-tuning on uncensored code repositories, and locally hosted open-source models provide threat actors with routes around provider restrictions.

2.3 Delivery: Spear Phishing at Machine Scale

Social engineering has historically ranked among the most consistently successful attack vectors, accounting for a significant proportion of initial access across multiple years of breach data [26]. It is also the phase of the kill chain most dramatically transformed by AI. Effective spear phishing requires research into the target's role, relationships, and recent activities; composition of convincing communication that mimics legitimate senders; and volume sufficient to overcome organizational defenses. All three requirements align with LLM strengths.

Industry practitioner reporting has observed AI-generated phishing content outperforming human red team operators in click rates and credential harvesting success in organizational deployments [9]. The explanation is not that AI composes more convincing prose in isolation, but that it can rapidly synthesize reconnaissance data into personalized communication at a scale that human operators cannot match. An AI agent can generate hundreds of individually tailored spear-phishing emails — each referencing the recipient's recent LinkedIn activity, their organization's current initiatives, or a vendor relationship identified through public sources — in the time a human operator would spend crafting a single message.

Voice and video deepfakes, while technically distinct from LLM text generation, represent a complementary vector that has already produced documented fraud at significant scale [25]. The combination of AI-generated text for initial contact, deepfake audio for telephone verification, and AI-crafted follow-up

communications creates a multi-modal social engineering pipeline that is difficult for targets to distinguish from legitimate communication.

2.4 Exploitation and Lateral Movement: Machine-Speed Adaptation

Once initial access is established, lateral movement traditionally requires an operator to manually explore the compromised network: enumerating hosts, assessing accessible credentials, identifying valuable targets, and navigating security controls. AI agents can perform this work autonomously and adaptively, maintaining context across sessions and modifying their approach based on observed defenses.

In laboratory simulations and documented incidents, AI-assisted lateral movement has compressed what traditionally requires days of manual operator activity into hours [10]. An agent that has established initial access can systematically enumerate network topology, attempt credential reuse against identified targets, identify unpatched systems through banner grabbing and configuration inspection, and log its findings for operator review — all without continuous human direction. When blocked by a security control, an AI agent can select alternative paths from its enumerated options rather than requiring operator intervention to unblock it.

This adaptability is particularly concerning in environments where defenders rely on the assumption that blocking an attack path will cause the attacker to abandon the attempt. AI-assisted adversaries may exhibit greater persistence than human operators under time pressure, and within their operational parameters they are less subject to the cognitive fatigue that causes human attackers to make identifiable errors during extended operations. However, agentic systems have their own operational constraints — including context window limitations, instruction drift, and resource exhaustion — that can interrupt autonomous operations.

2.5 Exfiltration and Post-Exploitation: Precision Over Volume

The final phases of the attack lifecycle — exfiltration, data encryption, and achievement of objectives — benefit from AI's ability to classify and prioritize large volumes of data rapidly. Rather than exfiltrating everything accessible, an AI agent can scan file systems, email archives, and databases to identify the highest-value targets: intellectual property, financial records, credentials for additional access, personally identifiable information with regulatory significance, and files containing authentication tokens or private keys. This targeted exfiltration reduces operational noise, decreases detection risk, and maximizes the leverage available for extortion.

In ransomware operations, AI-assisted reconnaissance of organizational structure and operational technology dependencies enables attackers to identify which systems, when encrypted, will cause maximum business disruption. Qilin, one of the ransomware groups reported to have integrated AI agents into their

operations as of Q1 2026, has reportedly deployed AI specifically trained to identify medical data repositories and critical operational technology dependencies, enabling the group to maximize leverage against healthcare targets [11].

3. Threat Actor Adoption: Evidence from the Field

The progression from academic demonstration to operational deployment has been faster than most threat intelligence analysts anticipated. AI capabilities that were considered emerging research in 2023 had been integrated into operational attack tooling by mid-2025, and by early 2026 multiple categories of threat actors had confirmed or demonstrably adopted AI-assisted techniques.

3.1 Nation-State Actors

Nation-state groups have the resources, technical sophistication, and long operational timelines that favor early adoption of AI capabilities. In September 2025, security researchers and subsequent reporting documented what appears to be the first publicly confirmed instance of a state-sponsored AI coding agent conducting an autonomous cyber espionage campaign. The agent handled an estimated 80 to 90 percent of tactical operations — reconnaissance, code generation, and lateral movement attempts — independently, with human operators directing only high-level objectives [4]. The campaign targeted 30 organizations across multiple sectors and geographies.

The use of AI agents for espionage reflects a strategic calculation: automating labor-intensive intelligence operations reduces operator risk and allows a smaller team to maintain persistent access across a broader target set. Nation-state groups that previously concentrated resources on high-value targets are increasingly capable of maintaining wide-area surveillance operations at reduced marginal cost, with AI handling the routine tasks that would otherwise consume analyst time.

A December 2025 through February 2026 cyberattack against water infrastructure in the Monterrey metropolitan area of Mexico illustrated how commercial AI services can be integrated into critical infrastructure attack planning regardless of threat actor sophistication. The incident — whose attribution had not been publicly confirmed as of this writing — documented that both Anthropic's Claude and OpenAI's GPT models were used to assist with planning and executing the campaign [12]. The incident raises significant policy questions about the governance of AI model access when those models can contribute to attacks on critical infrastructure — questions that commercial AI providers are only beginning to address through enhanced monitoring and access restrictions.

Advanced Persistent Threat (APT) groups such as APT29, previously known for patient, methodical tradecraft, have reportedly incorporated AI to accelerate the OSINT and phishing phases of their campaigns while maintaining the stealth-focused operational security for which they are known [10]. The combination of AI-assisted speed in early attack phases with human-directed precision in later phases represents a functional division of labor that maximizes the impact of limited skilled operator time.

3.2 Ransomware Groups: AI as an Operational Multiplier

Ransomware operations have adopted AI in ways that directly extend their business model. The "double extortion" approach — encrypting victim data while also threatening to publish stolen information — requires both technical access and enough intelligence about the victim's data landscape to credibly threaten the most damaging disclosures. AI supports both dimensions.

By Q1 2026, three prominent ransomware groups — Akira, Qilin, and Scattered Spider — had been reported to have integrated AI agents into their attack pipelines, according to multiple threat intelligence sources [6] [11][13]. Their reported applications span the full pre-exploitation phase: AI-powered OSINT to build target profiles, automated scanning to identify exposed attack surfaces, personalized spear-phishing generation, and credential harvesting at scale. The automation of these phases allows ransomware groups to operate with smaller teams while maintaining higher throughput than was previously achievable.

These trends have measurable financial consequences. A joint advisory from CISA, the FBI, and Europol documented that Akira collected more than $42 million in ransom payments in its first ten months of operation, between March 2023 and January 2024 [24] — a figure that subsequent reporting has indicated has grown substantially. As AI reduces the per-attack cost of reconnaissance and initial access, the marginal economics of ransomware operations improve. Groups that previously had to be selective about targets given resource constraints can now pursue a higher volume of targets simultaneously, increasing the probability that some fraction will be profitable.

The SANS Institute's analysis of AI-powered ransomware characterized this as AI enabling the "industrialization" of ransomware operations, with AI handling the scalable but labor-intensive work of target identification and initial access while experienced operators focus on post-exploitation and negotiation [13]. Europol's IOCTA 2026 report echoed this characterization, identifying the shift to industrialized cybercrime as a primary concern for European law enforcement [5].

3.3 Lowered Barriers: AI as a Skill Amplifier for Lower-Sophistication Actors

Perhaps the most strategically significant implication of offensive AI is the extent to which it democratizes capability. The FortiGate campaign documented by Amazon Web Services threat intelligence in early 2026 exemplifies this dynamic. The threat actor — characterized as a financially motivated, Russian-speaking individual or small group assessed at low-to-medium skill level — leveraged multiple commercial generative AI services to compromise more than 600 FortiGate devices across 55 countries between January 11 and February 18, 2026 [2].

The attack required no exploitation of previously undisclosed vulnerabilities. Instead, the threat actor used AI-assisted Python scripts to systematically scan for exposed management interfaces, attempt authentication with commonly reused credentials, and then parse and decrypt stolen device configurations

to extract network topology, credentials, and sensitive configuration data at scale. The result was a pre-ransomware staging operation — credential harvesting and Active Directory compromise across a globally distributed victim set — that would previously have required technical expertise well beyond the assessed skill level of the actor.

This case is illustrative of a broader pattern: AI does not eliminate the need for human judgment in attack planning, but it does allow actors with moderate skill to execute operations that previously required specialist expertise. The implication for defenders is that the threat model for many organizations must expand to encompass a significantly larger population of capable adversaries than traditional skill-based threat stratification would suggest.

4. The Convergence of Agentic AI and the Offense-Defense Asymmetry

Several researchers have argued that AI provides a near-term structural advantage to attackers, based on an asymmetry: offense needs to find one exploitable path through an organization's defenses, while defense must protect all of them [14]. This asymmetry hypothesis, while not universally accepted, informs how AI automation applies differently to each side of the security boundary. AI automation applied to reconnaissance and exploitation dramatically increases the breadth with which attackers can probe an attack surface, while defenders face the computationally harder problem of monitoring that same attack surface comprehensively. A well-resourced attacker can deploy an AI agent to systematically enumerate an organization's exposed services, test credential reuse across all identified authentication endpoints, and probe for known misconfigurations across every cloud resource — all in parallel, all without manual intervention.

This asymmetry is compounded by the economics of AI access. The per-query cost of commercial LLM inference is low and declining, while the cost of deploying comprehensive AI-assisted monitoring across a complex enterprise environment remains significant. Attackers operating outside legal and regulatory constraints can leverage the full capability of frontier models without the guardrail constraints that enterprise security products typically apply.

Several researchers have proposed that the AI-versus-AI dynamic — in which both attackers and defenders deploy autonomous agents — will eventually reach a new equilibrium [14]. This analysis is likely correct over a longer time horizon, but the near-term transition period is the period of greatest risk. Organizations that have not yet deployed AI-assisted detection and response capabilities will face AI-assisted attacks operating faster than human analysts can respond to manually. A September 2025 lab simulation documented that an AI "Reconnaissance AI Agent" continuously probed target networks while an "Exfiltration AI Agent" stealthily shifted data to alternate cloud channels when blocked, compressing what normally requires days of manual operator activity into less than an hour [4].

The speed dimension is significant for incident response. Security operations centers built around human analyst workflows operate on timelines measured in minutes to hours for initial triage and containment decisions. AI-assisted attacks can complete their primary objectives — initial access, lateral movement, data theft, and encryption — within a timeframe that challenges conventional response workflows. The mean time between initial compromise and ransomware deployment, which had already compressed over the preceding decade, is being further reduced by AI-assisted automation of the intervening steps.

5. Multi-Agent Attack Architectures: The Emerging Frontier

While much current threat intelligence focuses on AI-assisted versions of conventional attacks, a more challenging threat vector is emerging: coordinated multi-agent attack pipelines in which specialized AI agents collaborate across attack phases. Each agent in such a pipeline is optimized for a specific function — one for OSINT, one for spear-phishing content generation, one for exploit development, one for post-exploitation enumeration — and they pass outputs between stages as context for the next agent's task.

This architecture mirrors the multi-agent workflows being built for legitimate enterprise automation purposes, which creates an additional governance challenge: the same patterns that defenders are implementing for productive automation can be replicated for offense. The academic paper "Agentic AI and the Industrialization of Cyber Offense," published in May 2026, characterized this as the attack surface expanding beyond individual compromised systems to encompass the entire orchestration layer of enterprise AI infrastructure [14].

The MITRE ATLAS framework, which tracks adversarial machine learning techniques, explicitly addresses this threat surface. The November 2024 v5.1.0 update expanded the framework to cover 16 tactics, 84 techniques, and 56 sub-techniques relevant to AI system attacks, with the February 2025 v5.4.0 update adding further techniques specific to agentic systems including "Publish Poisoned AI Agent Tool" and "Escape to Host" [15][28]. These additions acknowledge that the attack surface relevant to AI-enabled offense includes not only the models themselves but the orchestration infrastructure, memory systems, tool integrations, and inter-agent communication channels.

The CSA Agentic AI Red Teaming Guide, published in May 2025, identified 12 distinct vulnerability categories specific to agentic systems that are relevant to the offensive threat landscape: control hijacking, permission escalation, hallucination exploitation, knowledge base poisoning, memory manipulation, multi-agent collusion, resource exhaustion, supply chain compromise, and untraceability, among others [16]. Each of these categories represents both an attack technique that adversarial actors may deploy against defender-operated AI systems and a capability that malicious agentic systems may exploit when targeting victim environments.

The most operationally significant of these is prompt injection — the insertion of adversarial instructions into content that an AI agent processes as input. When an AI agent is used in attack tooling and encounters defensive AI systems, security filters, or AI-powered monitoring tools in the target environment, prompt injection can be used to manipulate those systems' behavior. The discovery of "second-order" prompt injection techniques — where a low-privilege agent is manipulated into causing a higher-privilege agent to

take a malicious action on its behalf — demonstrates that multi-agent architectures introduce attack surfaces with structural analogies to second-order SQL injection and confused deputy attacks, but with characteristics unique to natural language interfaces that existing tooling is not yet equipped to address.

6. Defensive Strategies: Closing the Asymmetry Gap

Understanding the mechanics of AI-accelerated offense is prerequisite to building effective defense. The following defensive strategies are organized around the principal gaps that current AI-assisted attack techniques exploit.

6.1 Behavioral Detection Over Signature-Based Controls

The most fundamental shift required by the AI threat landscape is from signature-based detection toward behavioral analysis. AI-generated phishing content does not match known templates; AI-assisted malware mutates between deployments; AI-orchestrated lateral movement adapts to the specific environment rather than following predictable playbooks. Defenders who rely primarily on known-bad indicators will face a rapidly growing detection gap.

Behavioral analytics platforms that establish baselines of normal activity across users, systems, and network traffic are better positioned to detect AI-assisted intrusions, which tend to exhibit anomalous patterns even when the specific techniques employed are novel. An AI agent conducting systematic credential stuffing across a large authentication surface will generate access patterns that differ from normal user behavior regardless of whether the specific credentials it tries match known lists. An AI orchestrating lateral movement will exhibit reconnaissance activity across network ranges that human users would not normally traverse.

Industry guidance has proposed architectures in which AI-generated content is evaluated by an independent rule-based or statistical classifier before being acted upon by an AI-powered security tool, reducing the attack surface for prompt injection and other forms of AI manipulation [17]. The deployment of non-LLM classification layers as a secondary control sits upstream of AI-native monitoring systems and provides an additional check against manipulation.

6.2 Hardening Identity and Authentication at Scale

The FortiGate campaign of early 2026 succeeded not by exploiting software vulnerabilities but by systematically testing management interfaces secured only with weak or reused credentials and single-factor authentication [2]. This underscores a defensive priority that predates AI but has been given new urgency: multi-factor authentication on all management interfaces and eliminating credential reuse across organizational systems at scale. AI-powered credential harvesting attacks can test a large credential matrix against a large set of targets quickly enough to find matches in reasonable time; the only reliable mitigation is eliminating the reused credentials themselves and enforcing phishing-resistant MFA where feasible.

Zero Trust architectures, which enforce continuous identity verification throughout a session rather than only at initial authentication, reduce the value of compromised credentials to an attacker operating at machine speed. An AI agent that acquires a credential set through phishing or credential stuffing and immediately begins using it for lateral movement will encounter continuous re-verification requirements that slow and constrain its operation in a well-implemented Zero Trust environment [17].

Non-human identities — API keys, service account tokens, and machine credentials — represent a particular vulnerability surface for AI-assisted attacks because they are often issued with broad permissions, rarely rotated, and seldom monitored with the same scrutiny as human user credentials. The CSA State of AI Agents Security Survey (2026) found that a significant proportion of organizations lack comprehensive visibility into the non-human identities associated with their AI deployments, creating an attack surface that AI-powered credential harvesting tools are well-positioned to exploit [18].

6.3 Attack Surface Reduction and Exposure Management

AI-powered reconnaissance is most effective when the target attack surface is large and inconsistently managed. Exposed management interfaces, forgotten internet-facing services, shadow cloud deployments, and public repositories containing credentials or internal documentation all represent inputs to AI-assisted target profiling. Systematic attack surface management — continuous discovery and inventory of internet-facing assets, prompt remediation of exposed management interfaces, secret scanning in code repositories, and consistent application of network segmentation — reduces the yield of AI-powered reconnaissance and increases the preparation time an attacker must invest before attempting exploitation.

The AI-accelerated reconnaissance dynamic gives new importance to the speed of remediation. When a vulnerability is disclosed and an AI agent can generate working exploit code within hours, the patch deployment window narrows significantly. Vulnerability management programs must be calibrated to the AI-accelerated exploitation timeline, not the historical assumption of days to weeks between disclosure and weaponization.

6.4 AI-Aware Incident Response

Security operations centers must adapt their processes to the speed of AI-assisted attacks. Incident response playbooks built around human analyst timelines — initial triage in 15-30 minutes, containment decisions in one to two hours — may not be adequate when AI-assisted attacks can progress from initial access to full lateral movement in under an hour. Automated containment capabilities — the ability to isolate a compromised host, revoke a credential set, or block network traffic patterns based on behavioral signals without waiting for analyst confirmation — become critical in this environment.

The use of AI by defenders for threat hunting and automated response is, in this respect, not a luxury but a necessity. AI-assisted security operations tools can correlate signals across a large event stream at speeds that human analysts cannot match, enabling faster detection of the systematic patterns that AI-assisted attacks tend to produce. The CSA Benchmark Study of AI Agents in the SOC documents a growing practice of deploying AI agents for tier-one alert triage, allowing human analysts to focus on the higher-judgment tasks of investigation and containment [19].

Forensic readiness for AI-assisted attacks requires particular attention to logging completeness and integrity. AI-assisted attackers operating through compromised systems may attempt to suppress or manipulate logs to obscure their activities. Immutable logging infrastructure, cryptographically signed audit trails, and network-based evidence capture provide more reliable forensic records than host-based logs that an attacker with administrative access may have modified.

6.5 Governance of Dual-Use AI Capabilities

One of the most challenging aspects of the AI-accelerated threat landscape is that the same AI capabilities used for offense are also used for defense. LLMs used for attack reconnaissance are the same models used for security research and threat intelligence analysis. Agentic frameworks used to automate attack pipelines are also used to automate security operations. This dual-use reality complicates governance of AI access within the enterprise.

Organizations deploying AI security tools must grapple with the possibility that those tools could themselves be targeted by adversarial AI. Prompt injection attacks against AI-powered security systems — designed to cause the security tool to misclassify malicious activity as benign — are a documented threat that requires architectural mitigations, not merely policy responses [16]. Security teams should evaluate the robustness of AI-powered security tools to adversarial manipulation before deploying them in high-stakes detection and response roles.

The CSA AI Controls Matrix (AICM) provides a structured set of controls applicable to AI deployments that can be applied both to defender-operated AI systems and to the governance of AI use within the broader enterprise [20]. Controls relevant to the offensive AI threat include those addressing AI system access management, output validation, monitoring and logging, and supply chain integrity. Organizations that have not yet aligned their AI deployments to the AICM have an incomplete picture of the governance controls available to reduce the risk of their AI systems being exploited or manipulated.

7. Policy and Industry Coordination

The technical dimensions of AI-accelerated offense cannot be fully addressed by individual organizations acting in isolation. Several policy and coordination challenges require industry-level and governmental responses.

AI model providers occupy a structurally significant position in the offensive AI ecosystem. Commercial API access to frontier models provides the intelligence that drives AI-assisted reconnaissance, payload generation, and target selection. While responsible providers maintain safety classifiers and usage policies intended to prevent obvious misuse, the FortiGate campaign demonstrated that AI-assisted attacks can be conducted using entirely legitimate-seeming API requests — automated scripts that use AI to parse and analyze data, not requests that directly solicit attack assistance. Developing more sophisticated monitoring of AI API usage patterns that indicate offensive automation, without compromising the privacy and productivity benefits of legitimate high-volume use, is a research and policy challenge that has not yet been adequately addressed.

The governance of AI model access for critical infrastructure attacks, illustrated by the Mexico water facility incident, requires coordination between model providers, national cybersecurity agencies, and critical infrastructure operators. Voluntary commitments by major AI providers to monitor for and respond to critical infrastructure targeting are a starting point, but they do not yet constitute systematic governance proportionate to the scale of the threat.

Attribution and legal deterrence for AI-assisted attacks are further complicated by the ease with which AI can generate operationally plausible but misleading technical indicators. When an AI agent generates phishing content, exploit code, and post-exploitation scripts from scratch rather than reusing known tools and techniques, the conventional indicators of compromise that attribution relies on are absent. International coordination on attribution methodology for AI-assisted attacks, and on the legal frameworks applicable to AI-assisted cybercrime, is an urgent policy gap.

8. Conclusions and Recommendations

AI-accelerated attack pipelines represent a structural shift in the threat landscape, not a temporary spike in capability. The evidence from 2024 through early 2026 confirms that threat actors at multiple sophistication levels have integrated LLMs and agentic AI into their operations, that this integration has produced measurable increases in attack scale and speed, and that the pace of adoption shows no signs of plateauing.

The recommendations that follow are organized by urgency and organizational scope.

Immediate Actions (0-90 Days)

Organizations should immediately audit all internet-facing management interfaces for single-factor authentication and exposed credentials, prioritizing remediation based on exploitation complexity. The FortiGate campaign of early 2026 succeeded against security fundamentals that many organizations believe they have addressed — experience and incident data consistently reveal gaps between documented controls and actual implementation. Phishing-resistant multi-factor authentication should be deployed on all administrative access paths and extended to any system that would provide meaningful lateral movement leverage if compromised.

Security operations teams should conduct an assessment of their current detection capabilities against AI-assisted attack patterns: systematic credential stuffing across a large target set, rapid sequential authentication attempts from varying source addresses, high-volume network enumeration, and lateral movement exhibiting unusual breadth without depth. Gaps in coverage of these patterns should be prioritized for immediate remediation.

Organizations that have deployed AI-powered security tools should assess those tools' robustness to prompt injection and adversarial manipulation, particularly if the tools consume external content as input to their analysis and decision-making processes.

Short-Term Mitigations (90-180 Days)

AI-aware incident response procedures should be developed and exercised, specifically addressing the scenario of an AI-assisted intrusion progressing rapidly from initial access through lateral movement. Automated containment capabilities — pre-approved playbooks that can execute isolation, credential revocation, and traffic blocking based on behavioral triggers without analyst confirmation — should be developed and validated.

Red team exercises should explicitly incorporate AI-assisted attack techniques. The CSA Agentic AI Red Teaming Guide provides a methodology for testing AI system vulnerabilities [16]; the same methodology can be adapted to test organizational defenses against AI-assisted attacks conducted against conventional IT infrastructure. Understanding how quickly an AI-assisted attacker could conduct reconnaissance against the organization's current attack surface, and how long it would take current monitoring to detect systematic credential testing or network enumeration, provides a concrete baseline for prioritizing defensive investments.

Supply chain security for AI dependencies — model providers, AI-powered security tools, and any AI components embedded in enterprise applications — should be addressed through vendor security assessments aligned with the AICM [20]. The MITRE ATLAS framework's inclusion of "Publish Poisoned AI Agent Tool" as an attack technique reflects a real supply chain risk that organizations have not yet broadly incorporated into their vendor due diligence processes [15][28].

Strategic Considerations (6–18 Months)

At the strategic level, organizations must grapple with the reality that the effective threat actor population they face has grown significantly due to AI's skill amplification effect. Threat models that were calibrated to nation-state and organized criminal actors should be expanded to include opportunistic actors with AI-enhanced capabilities. Security program maturity assessments should include an evaluation of resilience against AI-assisted attacks, not merely conventional intrusion techniques.

The development of AI-native security capabilities — behavioral analytics platforms, AI-assisted threat hunting, automated containment orchestration — should be treated as a multi-year program requiring sustained investment, not a one-time deployment. The AI-versus-AI equilibrium is a moving target; defender capabilities that are adequate today may be inadequate against attacker capabilities emerging over the next twelve to eighteen months.

Investment in threat intelligence programs that specifically track AI adoption by threat actors — the tools, techniques, and AI service providers observed in attributed campaigns — provides early warning of capability shifts that can inform defensive prioritization before new capabilities are employed against the organization.

9. CSA Resource Alignment

The threat landscape described in this paper maps directly to several CSA frameworks and publications that provide actionable guidance for organizational response.

MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome), introduced by CSA in February 2025, provides a seven-layer threat modeling framework specifically designed for agentic AI systems [21]. MAESTRO's recognition that threats chain across layers — from foundation models through agent frameworks, tool integrations, and deployment infrastructure — is directly applicable to understanding the multi-agent attack architectures emerging in the offensive threat landscape. Organizations deploying AI security tools should apply MAESTRO to model the attack surface those tools introduce, not only the threats they are intended to detect.

The CSA AI Controls Matrix (AICM) provides a comprehensive set of controls for AI deployments that spans model providers, orchestration service providers, application providers, and AI customers [20]. Controls within the AICM relevant to the offensive AI threat include those addressing access management for AI systems, output validation and integrity, monitoring and logging of AI operations, and supply chain security for AI components. Aligning enterprise AI deployments — including AI-powered security tools — to the AICM provides a structured approach to reducing the attack surface that adversarial AI can exploit.

The CSA Agentic AI Red Teaming Guide (May 2025) defines a testing methodology covering 12 vulnerability categories specific to agentic systems [16]. Security teams should apply this methodology both to evaluate the resilience of AI systems deployed within the enterprise and to develop offensive AI test scenarios for red team exercises designed to test organizational defenses against AI-assisted attacks.

The CSA LLM Threats Taxonomy provides standardized definitions for the nine primary threat categories applicable to LLM deployments, supporting consistent risk communication across the organization and with third-party AI vendors [22]. The taxonomy's classification of model manipulation, data poisoning, prompt injection, and insecure supply chain threats directly addresses the mechanisms by which AI systems can be exploited as components of offensive pipelines.

AI Organizational Responsibilities, the CSA publication series addressing governance, risk management, and compliance for AI deployments, provides guidance on the operational and governance controls that reduce the risk of enterprise AI being compromised or misused [23]. As organizations expand their AI deployments in response to competitive and productivity pressures, ensuring that governance keeps pace with deployment is essential to preventing AI systems from becoming unmonitored attack surfaces.

The NIST AI Risk Management Framework (AI RMF 1.0) [27], referenced by the AICM and multiple CSA publications, provides a complementary organizational framework for managing AI risk that can be aligned with AICM controls for comprehensive AI governance coverage.

References

[1] R. Fang, R. Bindu, A. Gupta, and D. Kang. "LLM Agents Can Autonomously Exploit One-Day Vulnerabilities." arXiv:2404.08144, April 2024.

[2] Amazon Web Services Threat Intelligence. "AI-Augmented Threat Actor Accesses FortiGate Devices at Scale." AWS Security Blog, February 2026.

[3] H. Aftab. "AI-Powered Penetration Testing: How I Used Claude + Kali Linux MCP to Automate Security Assessments." DEV Community, 2025.

[4] Center for Strategic and International Studies. "Beyond Autonomous Attacks: The Reality of AI-Enabled Cyber Threats." CSIS Strategic Technologies Blog, 2025.

[5] Industrial Cyber. "Europol IOCTA 2026 Report Flags Shift to Industrialised Cybercrime Powered by AI, Ransomware and Data Theft." Industrial Cyber, 2026.

[6] AI Automation Global. "Ransomware Gangs Now Use AI Agents – 3 Groups Named." AI Automation Global, Q1 2026.

[7] The Hacker News. "2026: The Year of AI-Assisted Attacks." The Hacker News, May 2026.

[8] Trend Micro. "The AI-fication of Cyberthreats: Trend Micro Security Predictions for 2026." Trend Micro, 2025.

[9] Northwave Cybersecurity. "How AI-Driven Cyberattacks Are Changing the Threat Landscape in 2026." Northwave, 2026.

[10] Penligent AI. "AI Agents Hacking in 2026: Defending the New Execution Boundary." Penligent AI, 2026.

[11] CybelAngel. "Qilin Ransomware: Attack Methods and 2026 Status." CybelAngel, 2026.

[12] Infosecurity Magazine. "OpenAI and Anthropic LLMs Used in Critical Infrastructure Cyber-Attack." Infosecurity Magazine, 2026.

[13] SANS Institute. "AI-Powered Ransomware: How Threat Actors Weaponize AI Across the Attack Lifecycle." SANS Blog, 2026.

[14] arXiv. "Agentic AI and the Industrialization of Cyber Offense: Forecast, Consequences, and Defensive Priorities for Enterprises and the Mittelstand." arXiv, May 2026.

[15] Zenity. "MITRE ATLAS AI Security and Agentic Threats 2026 Update." Zenity Blog, 2026.

[16] Cloud Security Alliance. "Agentic AI Red Teaming Guide." CSA, May 2025.

[17] Atlan. "AI Agent Risks & Guardrails: 2026 Enterprise Security Guide." Atlan, 2026.

[18] Cloud Security Alliance. "Securing Autonomous AI Agents: Survey Report." CSA, 2026.

[19] Cloud Security Alliance. "A Benchmark Study of AI Agents in the SOC." CSA, 2025.

[20] Cloud Security Alliance. "AI Controls Matrix (AICM)." CSA.

[21] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 2025.

[22] Cloud Security Alliance. "Large Language Model (LLM) Threats Taxonomy." CSA, June 2024.

[23] Cloud Security Alliance. "AI Organizational Responsibilities: Governance, Risk Management, Compliance and Cultural Aspects." CSA.

[24] Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, and Europol. "#StopRansomware: Akira Ransomware (Advisory AA24-109a)." CISA, April 2024.

[25] CNN. "Deepfake scam in Hong Kong nets $25 million." CNN, February 2024.

[26] Verizon Business. "Data Breach Investigations Report." Verizon, 2025.

[27] National Institute of Standards and Technology. "Artificial Intelligence Risk Management Framework (AI RMF 1.0)." NIST, January 2023.

[28] MITRE. "MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems." MITRE Corporation.