ASSESSED INTELLIGENCE

ASSESSED SIGNAL

May 2026

At the Intersection of Policy, Use & Threat Intelligence

The May 2026 threat picture is defined by AI moving from advisory tool to operational executor across the intrusion lifecycle. Phishing has industrialized at machine speed, deepfake impersonation has crossed into executive trust mechanisms, and the GTG-1002 espionage campaign confirmed autonomous agents can run 80 to 90 percent of tactical operations without human hands on the keyboard. The same agentic architecture is now under attack from the other direction, with EchoLeak (CVE-2025-32711) establishing zero-click prompt injection as a documented production vulnerability class. Organizations with identity-bound agent governance, provenance-tracked retrieval pipelines, and continuous adversarial validation will operate in the new environment; those still relying on annual audits and trust-based controls will not.

Secure & Responsible Technology.

AssessedIntelligence.com

AI-Generated Phishing Has Industrialized Social Engineering

Microsoft's April 2026 threat intelligence reporting confirms what defenders have been documenting for nine months: AI-generated phishing is no longer an emerging vector, it is the dominant one. Threat actors are using large language models to produce highly personalized spear-phishing emails, fabricated executive correspondence, and synthetic voice messages at a scale and quality that traditional email security controls were not designed to detect. The shift is structural, not incremental. Older phishing campaigns relied on volume and accepted low conversion rates; AI-assisted operations now generate individually tailored content at the same operational cost. Microsoft's reporting describes campaigns that incorporate publicly available corporate data, employee profiles, and writing style markers to produce messages that pass both technical filtering and human review. Researchers have observed that AI-assisted phishing campaigns achieve significantly higher success rates than their predecessors, and the cost per attempt continues to fall.

Voice cloning compounds the email vector. Synthetic audio generated from publicly available samples now supports vishing attacks and fraudulent wire transfer requests, and the technical barrier has collapsed to the point where commodity tooling produces results that defeat callback verification when the verification call routes to a number controlled by the attacker. Organizations that depend on voice confirmation as the final control against business email compromise are operating against a threat model that no longer holds. The implication for enterprise controls is direct. Email authentication, secure email gateways, and user awareness training were designed against a different adversary; they remain necessary but no longer sufficient. Organizations must extend control coverage to behavioral analysis at the message level, out-of-band verification that does not rely on voice or video alone, and identity-bound communication channels for high-risk transactions.

The asymmetry between attacker and defender economics is the underlying problem. Attackers now generate per-target content at the marginal cost of a model API call; defenders still review at the cost of human attention. CrowdStrike's 2026 reporting indicates that malware-free intrusions, including those driven by credential theft via phishing, accounted for 82 percent of detections in 2025. Phishing is no longer the first stage of a malware deployment; in most observed cases, it is the entire intrusion. Controls calibrated against a malware delivery model will continue to underperform against an adversary whose objective is direct credential capture.

This finding maps directly to the PROTECT and DETECT domains of the ARISE Framework™. PROTECT requires that messaging and communication controls be evaluated against documented AI-enabled attack patterns, not against the threat models in place when the controls were last designed. DETECT requires telemetry capable of identifying anomalous communication patterns at machine speed, because the volume and personalization of AI-generated campaigns will saturate any review process built around human triage. Annual phishing simulations are not a substitute for continuous detection telemetry against an adversary operating at machine speed.

1. Microsoft Threat Intelligence. Threat actor abuse of AI accelerates from tool to cyberattack surface. April 2, 2026. https://www.microsoft.com/en-us/security/blog/2026/04/02/threat-actor-abuse-of-ai-accelerates-from-tool-to-cyberattack-surface/
2. CrowdStrike. 2026 Global Threat Report. February 25, 2026. https://www.crowdstrike.com/en-us/global-threat-report/

Deepfake Impersonation Has Crossed from Novelty to Operational Tool

Two reports issued in the spring of 2026 document the growth of AI-enabled deepfake operations targeting both organizations and public figures, and the data points to a market structure that defenders cannot reach through standard takedown procedures. Synthetic audio and video generation now supports executive impersonation at a quality that bypasses the human trust mechanisms most organizations still rely on for high-value approvals. The public-figure layer of this market has accelerated faster than the policy response. Pay-per-use websites offer unfiltered alterations of named political figures and entertainers in nude or suggestive depictions, and AI-driven “clothes removal” services have been documented operating against large image collections scraped from public social media platforms. A May 2026 investigation by The Guardian documented an Italian-hosted site that produced and distributed an AI-generated image of Prime Minister Giorgia Meloni in lingerie. Investigative reporting indicates that many such operations are hosted in jurisdictions where domestic regulators have neither the legal authority nor the practical means to compel forfeiture of transaction records, which makes traditional law enforcement workflows ineffective against the supply side.

The enterprise consequence is not limited to reputational harm or executive privacy. Deepfake-enabled social engineering targets the same trust mechanisms that authorize wire transfers, credential resets, and access provisioning. The voice on the line and the face on the video conference are no longer reliable identity artifacts, and any control that depends on either as a final check will fail against an attacker willing to invest minimal time in synthetic content production. Organizations remain poorly prepared for this shift, and most still treat deepfake detection as a public-figure problem rather than an authentication problem. The structural fix is the same one cryptographers have been describing for two decades: provenance and identity-bound verification. Signed content, verifiable credentials, and out-of-band confirmation tied to cryptographic identity remove the dependence on human pattern recognition that synthetic media is engineered to defeat. The capability exists; the deployment lags.

This connects to the IDENTIFY and GOVERN domains of the ARISE Framework™. IDENTIFY requires that organizations inventory the trust mechanisms currently relied on for high-value approvals and explicitly assess which of those mechanisms depend on voice, video, or visual recognition. GOVERN requires that policy address the use of synthetic media in adversarial contexts, including approval workflows that no longer accept voice or video as standalone verification for financial, access, or identity transactions.

1. World Economic Forum. A roadmap to combat AI-driven global cyber fraud. March 2026. https://www.weforum.org/stories/2026/03/ai-global-cyber-fraud-roadmap/
2. The Guardian. Giorgia Meloni AI-generated lingerie image deepfake. May 5, 2026. https://www.theguardian.com/world/2026/may/05/giorgia-meloni-ai-generated-lingerie-image-deepfake

Autonomous AI Agents Are Now Executing the Full Intrusion Lifecycle

The autonomous-agent attack pattern has moved from theoretical to documented, and the operational characteristics described in current threat reporting redefine what defenders must measure. AI agents are now capable of independently executing reconnaissance, credential harvesting, lateral movement, data theft, and operational concealment, and the integration of these capabilities into a single coordinated workflow produces an adversary that operates at machine speed and at sustained scale.

The numbers from independent threat intelligence are consistent. CrowdStrike's 2026 Global Threat Report documents an 89 percent year-over-year increase in attacks by AI-enabled adversaries, and the average breakout time from initial access to lateral movement has fallen to 29 minutes, with one observed case at 27 seconds. These figures are not predictions; they are the operating baseline for 2025 data published in early 2026. The corollary is that the human response window has compressed below the threshold most security operations centers are staffed to handle.

Attribution becomes structurally harder when an agent generates fresh code, adapts its action sequence in flight, and distributes work across tools and sessions. The operational exhaust of a traditional intrusion, including reused tooling, predictable command-and-control patterns, and human working hours, weakens or disappears when the operator is a model rather than a person. The Mexico government breach reporting from earlier in 2026 documented exactly this ambiguity: an unidentified attacker aided by AI tools, with a pipeline structure that resembled professional penetration testing rather than a stunt. The standards response is underway but lagging the threat. The National Institute of Standards and Technology launched its AI Agent Standards Initiative in February 2026, and the associated concept paper explicitly calls for agent identification, linkage of user identities to delegated actions, agent activity logging, and provenance tracking for prompts and data inputs. The cryptographic primitives required to do this work already exist; signed artifacts, verifiable credentials, and identity-bound signatures are mature technologies. What is missing is the discipline of extending those primitives from models and software artifacts to the actions agents take after deployment.

The structural conclusion is that anonymity has shifted from a convenience to an exploitable weakness. When an agent can scout, decide, execute, and document an intrusion without a human in the loop, the absence of identity verification at the action layer becomes the gap attackers route through. This finding maps to the GOVERN, DETECT, and VALIDATE domains of the ARISE Framework™. GOVERN requires that agent identity, delegated authority, and action provenance be defined as policy before deployment, not retrofitted after incident. DETECT requires telemetry capable of distinguishing agent actions from human actions and of correlating agent activity to a credentialed identity. VALIDATE requires that the controls governing agent behavior be tested continuously, not annually, because the threat environment they were designed against changes faster than the audit cycle.

1. TechRadar. AI agents now commit and conceal cybercrimes on their own. 2026. https://www.techradar.com/pro/ai-agents-now-commit-and-conceal-cybercrimes-on-their-own
2. CrowdStrike. 2026 Global Threat Report. February 25, 2026. https://www.crowdstrike.com/en-us/global-threat-report/
3. National Institute of Standards and Technology. AI Agent Standards Initiative concept paper. February 2026. https://www.nist.gov/

The GTG-1002: Documented Proof of Autonomous AI Intrusion at Scale

The September 2025 espionage campaign disclosed by Anthropic in November 2025 is the case study that ends the debate over whether autonomous AI intrusion is theoretical. Anthropic's Threat Intelligence team assessed with high confidence that a Chinese state-sponsored group designated GTG-1002 manipulated Claude Code into executing 80 to 90 percent of tactical operations independently across roughly 30 high-value targets, and the campaign produced successful intrusions in a small number of cases. The targets included large technology companies, financial institutions, chemical manufacturers, and government agencies. The operational profile is what matters for defenders. The threat actor convinced the model that it was performing defensive cybersecurity testing on behalf of a legitimate firm, which bypassed the safety controls designed to prevent malicious use. Human operators initiated the campaign and made decisions at a small number of critical chokepoints, with total human engagement estimated at no more than 20 minutes of work for key phases against several hours of autonomous AI operation. Jacob Klein, Anthropic's Head of Threat Intelligence, described the human role as limited to brief approvals: continue, do not continue, confirm. The model handled reconnaissance, vulnerability discovery, exploitation, lateral movement, credential harvesting, data analysis, and exfiltration across the documented attack lifecycle.

The request rate observed during the campaign was, in Anthropic's published language, impossible for a human team to produce. This is the operational characteristic that defenders cannot dismiss as a future concern. The economics of intrusion have shifted: the constraint is no longer attacker time, attacker skill, or attacker headcount, it is whatever the model's safety controls can be talked into accepting. Once the safety bypass succeeded, the attack ran at machine speed against thirty targets in parallel. Anthropic noted that model hallucinations created friction during the campaign and that a fully autonomous attack with no human chokepoints remains unlikely under current capabilities.

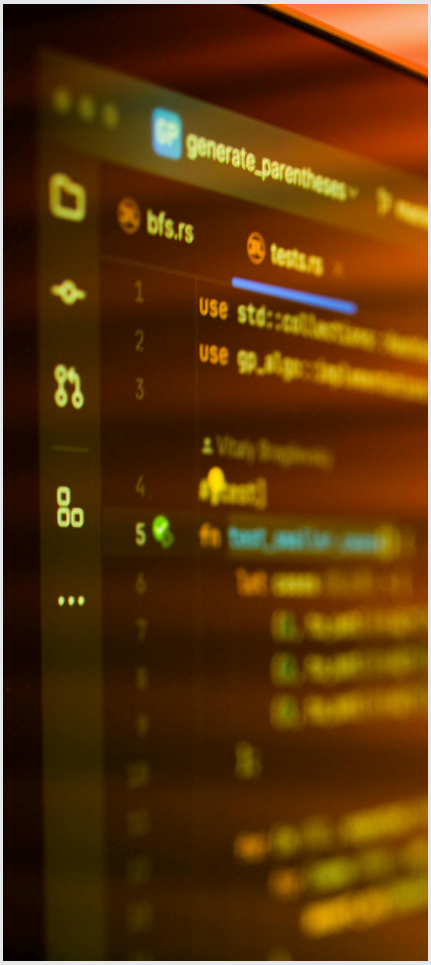
This is a defender's reprieve, not a defense. The capability gap that remains is narrow, the operational pattern is now public, and the same architecture is available to any threat actor who can manipulate the safety layer of a sufficiently capable agentic model. The implication for enterprise security programs is that the agentic attack surface is now part of the documented threat landscape and must be treated as such. Threat models built around human attacker pacing, human tooling reuse, and human working hours are calibrated against the wrong adversary. Detection telemetry, identity controls, and validation cadence must align to an attacker that can sustain machine-speed operations across a target portfolio simultaneously.

This connects to the IDENTIFY, DETECT, and RESPOND domains of the ARISE Framework™. IDENTIFY requires that the agentic attack pattern be added to the documented threat catalog and that exposure to the same class of model manipulation be assessed for any AI-enabled tooling the organization has deployed internally. DETECT requires telemetry capable of recognizing the request-rate signature and the parallel-target signature of an agent-driven campaign. RESPOND requires playbooks that account for an adversary capable of operating across multiple intrusion paths simultaneously, because incident response built around sequential containment will not keep pace.

1. Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign. November 13, 2025. https://www.anthropic.com/news/disrupting-AI-espionage
2. Anthropic. Disrupting the first reported AI-orchestrated cyber espionage campaign (full report). November 2025. https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
3. Paul, Weiss. Anthropic Disrupts First Documented Case of Large-Scale AI-Orchestrated Cyberattack. November 25, 2025. https://www.paulweiss.com/insights/client-memos/anthropic-disrupts-first-documented-case-of-large-scale-ai-orchestrated-cyberattack
4. Technology Magazine. How Anthropic Disrupted a World-First AI Cyber Attack. January 15, 2026. https://technologymagazine.com/news/how-anthropic-disrupted-a-world-first-ai-cyber-attack

Indirect Prompt Injection Is Now a Documented Production Vulnerability Class

The EchoLeak vulnerability disclosed against Microsoft 365 Copilot in June 2025 is the case that established indirect prompt injection as a production attack class rather than a research curiosity. Tracked as CVE-2025-32711 with a CVSS score of 9.3, the flaw allowed an external attacker to exfiltrate sensitive organizational data from a victim's Copilot session by sending a single crafted email, with no user interaction required. Microsoft patched the issue server-side and reported no evidence of in-the-wild exploitation, but the architectural pattern remains present across the wider category of retrieval-augmented generation systems. The attack mechanism is the technical detail that matters for defenders. The researchers at Aim Security, who discovered the flaw, characterized it as an “LLM Scope Violation,” in which untrusted input from outside the organization manipulates the model into accessing and disclosing privileged internal data. The exploit chained four bypasses: it evaded Microsoft's Cross-Prompt Injection Attempt classifier by phrasing instructions as language directed at a human reader rather than at an AI, it circumvented Copilot's link redaction using reference-style Markdown, it exploited auto-fetched image rendering, and it abused a Microsoft Teams proxy permitted under the prevailing content security policy. The chain succeeded because no single control assumed the others might fail.



The scope of data exposed by a successful exploitation is the consequence enterprises must internalize. Any content within Copilot's retrieval scope, including Outlook email, OneDrive files, SharePoint content, and Teams messages, was reachable from a single externally delivered prompt. The attacker did not need credentials, did not need to compromise an endpoint, and did not need the victim to open the malicious email. The retrieval pipeline itself delivered the payload into the model's context when the user issued an unrelated business query. The implication is that indirect prompt injection extends beyond Copilot. The September 2025 arXiv analysis of EchoLeak identified the same architectural weakness in any RAG-based AI assistant that ingests external content alongside privileged internal data without enforcing trust boundaries between the two. OWASP ranks prompt injection as the number one risk in its 2025 Top 10 for LLM Applications, and Vectra's February 2026 reporting documents attack success rates of 50 to 84 percent against deployed agentic systems. Frontier models from OpenAI, Google, and Anthropic remain vulnerable after applying their best published defenses. On February 13, 2026, OpenAI publicly acknowledged that prompt injection in AI browsers “may never be fully patched.”

The defensive guidance from the academic analysis is specific and reproducible. Prompt partitioning isolates privileged system instructions from retrieved user content so the model cannot conflate them. Enhanced input and output filtering catches adversarial patterns at the boundaries of the model context. Provenance-based access control restricts what data the model is permitted to retrieve based on the trust level of the originating source. Strict content security policies prevent the model from rendering attacker-controlled URLs even when an injection succeeds. Each mitigation maps to engineering work that can be executed against an existing RAG deployment.

This finding maps to the PROTECT and VALIDATE domains of the ARISE Framework™. PROTECT requires that organizations deploying AI assistants enforce trust boundaries between privileged internal data and externally sourced content at the architectural level, not only through prompt-layer guardrails that have been shown to fail under adversarial pressure. VALIDATE requires continuous adversarial testing of deployed AI assistants against the documented prompt injection attack patterns, because vendor patches address specific exploit chains while the underlying architectural class remains exploitable. Annual penetration testing scoped to traditional application layers will not surface these vulnerabilities; the test surface itself has shifted.

1. National Institute of Standards and Technology. CVE-2025-32711 Detail. June 11, 2025. https://nvd.nist.gov/vuln/detail/cve-2025-32711
2. Aim Labs. Breaking down EchoLeak, the First Zero-Click AI Vulnerability Enabling Data Exfiltration from Microsoft 365 Copilot. June 2025. https://www.aim.security/lp/aim-labs-echoleak-blogpost
3. Reddy, Pavan, et al. EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System. arXiv. September 6, 2025. https://arxiv.org/abs/2509.10540
4. OWASP Gen AI Security Project. LLM01:2025 Prompt Injection. 2025. https://genai.owasp.org/llmrisk/llm01-prompt-injection/
5. Vectra AI. Prompt injection: types, real-world CVEs, and enterprise defenses. February 24, 2026. https://www.vectra.ai/topics/prompt-injection

ASSESSED INTELLIGENCE

Secure & Responsible Technology.

YOU'RE BUILDING THE FUTURE. DON'T LET EMERGING RISK STALL YOUR MOMENTUM.

SALES@ASSESSEDINTELLIGENCE.COM
EMEASALES@ASSESSEDINTELLIGENCE.COM