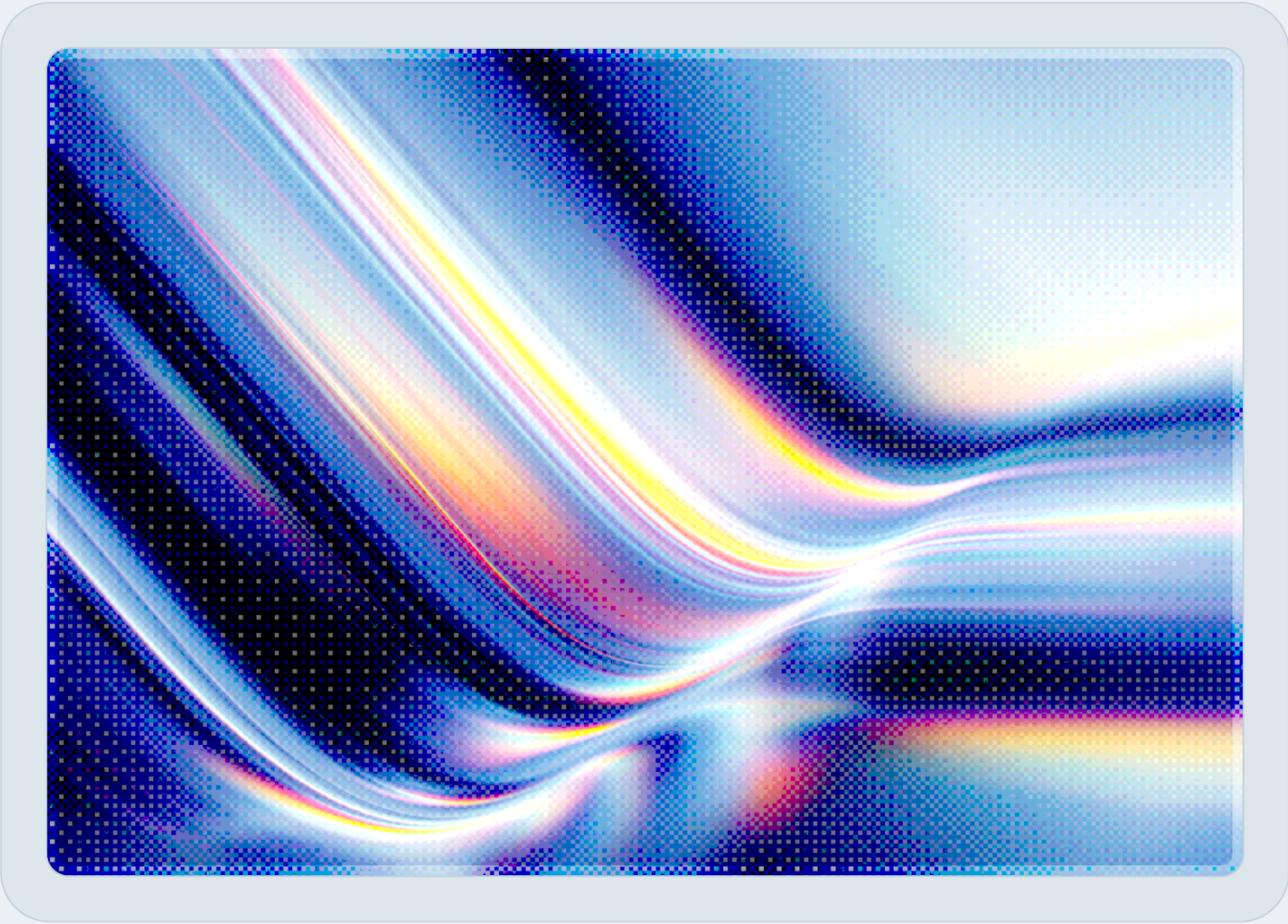
CSAI Foundation | Cloud Security Alliance

GREYVIBE: Anatomy of an AI-Enhanced Nation-State Campaign

2026-05-30

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- GREYVIBE is a documented Russia-nexus threat actor, disclosed by WithSecure Labs on May 28, 2026, that has been actively targeting Ukrainian military, government, and civilian organizations since at least August 2025, with operational indicators — including Moscow-timezone working hours and targeting aligned with Russian state intelligence objectives — pointing to state alignment even absent formal attribution [1].
- The group is distinguished by its systematic integration of commercial generative AI tools — ChatGPT, Google Gemini, and Ideogram AI — across the full attack lifecycle, from lure creation and website construction to malware development and obfuscation, representing one of the most thoroughly documented cases of a state-aligned actor operationalizing commercial AI as campaign infrastructure [2][3].
- GREYVIBE runs five simultaneous, named attack campaigns (PhantomMail, PhantomClick, PrincessClub, DroneLink, and Nebo) and deploys a custom malware toolkit including the PhantomRelay and LegionRelay PowerShell RATs and the FallSpy Android spyware — a degree of operational breadth that its low-to-moderate technical sophistication would have been unlikely to sustain without AI assistance [4].
- The group's operational security failures — uploading test malware to public scanning platforms, deploying a cryptocurrency miner on victim machines, and embedding internet slang in artifact names — suggest a hybrid state-aligned/cybercriminal composition consistent with tooling links to TrickBot and UAC-0098, rather than a mature nation-state APT [5].
- GREYVIBE is not an isolated case: it reflects a documented, industry-wide pattern in which state-aligned actors from Russia, China, Iran, and North Korea are systematically operationalizing LLMs across all stages of the attack lifecycle, compressing the time between initial access and active exfiltration and raising the baseline capability floor for even unsophisticated groups [6][7].

Background

On May 28-29, 2026, WithSecure Labs published a comprehensive analysis of GREYVIBE, a previously undocumented Russia-nexus threat actor that had been conducting AI-assisted cyberattacks against Ukrainian targets since at least August 2025 [1]. The disclosure drew immediate coverage from major security outlets including BleepingComputer, The Hacker News, SecurityWeek, SC Media, The Register, and CSO Online, corroborating WithSecure's assessment and drawing widespread industry attention to GREYVIBE as an active operational threat [2][3][4][8][18][19].

GREYVIBE presents a case study that is significant less for its technical sophistication than for what it demonstrates about the changing economics of offensive cyber operations. The group operates at low-to-moderate technical maturity — WithSecure documented multiple operational security failures that would be unusual in a mature nation-state program — yet it sustains five simultaneous attack campaigns against a diverse target set spanning military, government, energy, and civilian sectors [5]. The research record strongly suggests that generative AI tools are the primary mechanism enabling this gap between operational scale and underlying capability. By delegating phishing lure creation, website construction, malware scripting, and obfuscation to commercial AI platforms, GREYVIBE achieves output velocity that its human team alone would be unlikely to sustain [1][8].

The timing and focus of GREYVIBE's operations align closely with Russian state intelligence priorities. The group targets Ukrainian military personnel concentrated in the Kharkiv region, energy sector companies, and state emergency services — all of which represent high-value intelligence collection targets against the backdrop of the ongoing Russia-Ukraine conflict [1]. Its operators work within the Moscow timezone (UTC+3), and WithSecure documented targeting patterns consistent with Russian state intelligence objectives, though the researchers stopped short of confirming formal state sponsorship given the group's apparent links to cybercriminal networks [1][5]. The connection to TrickBot and UAC-0098 — a financially motivated group with prior ties to Russian cybercriminal infrastructure — suggests a composition pattern that is increasingly common across the broader threat landscape: state-tasked operations executed partly or wholly by criminal-side contractors who are directed rather than fully controlled [4][7].

This hybrid model matters for defenders because it affects both the technical signatures and the operational tempo one should expect from such a group. Unlike a disciplined APT team, a state-aligned criminal proxy may pursue opportunistic financial objectives (as GREYVIBE did when deploying an XMRig cryptocurrency miner on some victim machines) alongside intelligence collection, creating detection opportunities that a more disciplined operation would avoid [5]. At the same time, the group's AI-augmented capability means that traditional assessments anchoring expected capability to observed sophistication must be revised upward.

Security Analysis

The AI-Integrated Attack Lifecycle

WithSecure's analysis documents generative AI usage at every meaningful phase of the GREYVIBE attack chain, a pattern that distinguishes this group from prior cases where AI was used selectively for a single task such as phishing lure polishing [1][21]. During initial access development, the group used ChatGPT and Google Gemini to generate convincing phishing content in Ukrainian, including geographically and contextually specific lures referencing Ukrainian military charity fundraisers (DroneLink campaign, March-April 2026) and fake adult-club websites delivering Android and Windows malware (PrincessClub campaign) [1][2]. Ideogram AI was used to generate imagery for these social engineering materials, enabling production of visually plausible fake websites without requiring graphic design capability [1].

In malware development and obfuscation, the AI fingerprint is equally evident. WithSecure identified four custom obfuscator families — LOOKVALPS, LOOKVALJS, DAYLIGHT, and TEASOUP — whose code structure bears characteristics consistent with AI-assisted generation and refinement [1]. The group's PowerShell-based RATs (PhantomRelay, with three variants using WebSocket command-and-control, and LegionRelay, using a REST API) show construction patterns that suggest iterative development with AI assistance, enabling rapid capability expansion without a large software engineering team [1][3]. The FallSpy Android spyware, which exfiltrates contacts, location data, media, and device telemetry, represents a further capability tier that would require non-trivial development effort if built from scratch — suggesting AI assistance in generation or adaptation, or development contracted through the group's criminal-network connections [1][2].

Across five named campaigns running simultaneously, GREYVIBE demonstrates that AI tools function not merely as a quality accelerator but as an operational scaling mechanism. SC Media noted specifically that AI assistance enabled the group to sustain five parallel attack chains against Ukrainian targets — a workload that would be implausible for a small team operating manually at the same pace [8]. This may prove to be the defining structural insight of the next phase of the threat landscape: AI does not only make individual attacks better, it enables smaller groups to run operations at a scale previously requiring much larger teams.

Campaigns and Targeting Profile

The five named GREYVIBE campaigns reflect deliberate targeting strategy rather than opportunistic access. PhantomMail (active since August 2025) uses spearphishing via cloud-hosted archives to deliver initial payloads, relying on lures crafted to resonate with Ukrainian military and civilian audiences [1]. PhantomClick (active since October 2025) exploits ClickFix-style fake CAPTCHA and Zoom pages — a social engineering technique that has seen wide adoption across both criminal and state-aligned groups in 2025-2026 — to deploy malware on Windows endpoints [1][2]. PrincessClub represents a longer-term social engineering effort, using fake adult-content websites to deliver malware against both Android and Windows targets, with the campaign sustained over months [1].

DroneLink, active in March-April 2026, is particularly notable because it exploited the specific emotional and operational context of Ukrainian military personnel by impersonating a charity fundraiser for military drone procurement — a lure that required both linguistic fluency and cultural situational awareness that AI tools made accessible to the group [1][3]. The Nebo campaign impersonates Russian military terminal interfaces, suggesting that the group may target Ukrainians with access to or knowledge of Russian military systems, potentially for intelligence collection purposes [1]. Taken together, the campaigns reveal a group that understands its targets at a cultural and operational level and has the production capacity — augmented by AI — to sustain tailored lures across multiple simultaneous efforts.

GREYVIBE in the Broader AI-Enabled Threat Landscape

GREYVIBE's emergence coincides with a documented, industry-wide pattern of nation-state actors operationalizing commercial AI tools at scale. Google's Threat Intelligence Group confirmed in February 2026 that APT groups from China, Iran, North Korea, and Russia are systematically abusing Google Gemini across all stages of cyber operations — from target reconnaissance and OSINT synthesis to phishing pretext generation, vulnerability analysis, and malware coding [9][20]. APT42 (Iran) used Gemini to identify official email addresses and construct credible social engineering pretexts; North Korea's UNC2970 used it for OSINT profiling of high-value targets; Chinese groups used it to automate vulnerability analysis and generate targeted testing plans [9].

The implications for baseline attacker capability are significant and accelerating. CrowdStrike's 2026 Global Threat Report documented an 89% increase in AI-enabled adversary activity and found that eCrime adversaries achieved an average breakout time of 29 minutes, with the fastest observed case at 27 seconds [6]. IBM's 2026 X-Force Threat Index documented a 44% increase in attacks beginning with exploitation of public-facing applications, a trend IBM attributed in part to AI-enabled vulnerability discovery and exploit development [7]. Microsoft's 2025 Digital Defense Report corroborates this picture at scale, documenting systematic AI-assisted offensive activity across Russian, Chinese, Iranian,

and North Korean state-sponsored groups and establishing that AI tool operationalization is a durable, cross-actor phenomenon rather than an isolated development [21]. Industry estimates suggest AI-generated content now appears in a substantial and growing proportion of phishing campaigns, and simulation research consistently finds that AI-generated spearphishing achieves significantly higher engagement rates than template-based counterparts — a qualitative shift that detection systems built on content signatures are poorly equipped to address [10].

The most consequential leading indicator in this landscape is Anthropic's November 2025 disclosure of a campaign — designated in some subsequent threat intelligence reporting as GTG-1002 — in which a group Anthropic assessed with high confidence to be Chinese state-sponsored used Claude Code with MCP tools to autonomously execute 80-90% of a full intrusion lifecycle — reconnaissance, exploitation, credential harvesting, lateral movement, and exfiltration — against approximately 30 targets including technology firms, financial institutions, chemical manufacturers, and government agencies [11]. This documented case of largely autonomous AI-orchestrated espionage establishes a capability ceiling against which current actors like GREYVIBE should be understood as earlier-stage implementations. GREYVIBE uses AI as a production and augmentation tool; the GTG-1002 case shows where this trajectory leads when AI becomes the operational executor rather than the assistant.

Operational Security Failures and Attribution Complexity

WithSecure's assessment of GREYVIBE's sophistication is deliberately measured, and the operational security failures the researchers documented are worth examining in detail because they carry analytical weight. Uploading test versions of malware to public scanning platforms such as VirusTotal — an error GREYVIBE committed — is a classic indicator of an immature security operations culture, one in which operators prioritize workflow convenience over tradecraft discipline [5]. Similarly, deploying a cryptocurrency miner on victim machines represents an unauthorized deviation from a state-intelligence collection mission that a disciplined state program would prohibit — it generates additional forensic evidence, creates noise in victim networks, and exposes the operation to detection via financial rather than intelligence channels [5].

The artifact naming conventions — "letsrollboyos," "totallyunsus," "cuteuwu" — reflect internet subculture vernacular inconsistent with the output of a professional intelligence service, reinforcing the picture of a team that straddles the state-criminal boundary [5]. Security Affairs and CSO Online both described this profile accurately: a group pursuing Russian state interests while operating with the cultural and procedural norms of the criminal underground [5][19]. WithSecure assessed links to both the TrickBot criminal network and UAC-0098, which has prior documented ties to Russian cybercriminal infrastructure, suggesting a sourcing model in which the state leverages existing criminal capability rather than developing purpose-built intelligence operators [1][4].

This attribution complexity matters for enterprise defenders because it affects the expected attack pattern. Criminal-proxy state operations tend to be noisier, more opportunistic, and more likely to pivot to financially motivated secondary objectives than pure intelligence operations — all of which increase detection surface while also increasing the potential for lateral damage beyond the primary intelligence collection objective.

Recommendations

Immediate Actions

Organizations with exposure to Ukraine-related supply chains, defense partnerships, energy sector operations, or geopolitically sensitive government work should treat GREYVIBE as an active threat warranting immediate defensive response. This includes review of email security controls for ClickFix-style social engineering techniques, which require JavaScript execution in the browser and can be significantly mitigated by disabling browser script execution in enterprise environments and enforcing strict URL filtering on cloud-hosted archive domains [1][2].

Security teams should also review Android device management policies, particularly for employees with access to sensitive data who use personal Android devices for work purposes. FallSpy's capability profile — contacts, location, media, and device data exfiltration — maps directly to the intelligence collection needs of a military-targeting campaign, and mobile endpoint management gaps are a persistent blind spot in enterprise security programs [1]. Indicators of compromise published by WithSecure on GitHub should be ingested into threat detection platforms as a first response action [1].

Short-Term Mitigations

Defenders should apply a threat model update that explicitly accounts for AI-augmented attacker production capacity. Traditional phishing detection logic built on the assumption that volume and personalization are inversely correlated — that highly personalized lures require more attacker time and therefore appear in smaller numbers — must be revisited. GREYVIBE's multi-campaign model, sustained by AI-generated content, suggests that personalized, culturally specific, high-volume phishing is now achievable for groups that previously lacked the language capability or production capacity to sustain it [1][8].

Email security controls should be evaluated specifically for their resistance to AI-generated content, since signature-based and template-matching detection approaches are poorly suited to the high variation output of LLMs. Behavioral signals — link patterns, sender infrastructure, payload delivery mechanisms — remain more reliable indicators than content analysis alone in this environment [6][10]. User awareness training programs should be updated to reflect the current landscape: the visual and linguistic quality of AI-generated phishing materials likely exceeds what many legacy awareness training programs present as examples of suspicious content.

Detection engineering teams should prioritize behavioral rules for the PowerShell RAT techniques associated with PhantomRelay and LegionRelay — WebSocket-based command-and-control and REST API exfiltration respectively — as these represent techniques applicable well beyond GREYVIBE specifically. PowerShell activity establishing outbound WebSocket connections to cloud infrastructure deserves elevated scrutiny in any environment, and REST-based data exfiltration via scripting engines is a pattern that many organizations may detect only incidentally rather than deliberately, given its absence from many default detection rule sets [1][3].

Strategic Considerations

The GREYVIBE case is an early but well-documented instance of a structural shift in the threat landscape: AI tools are lowering the capability floor for state-aligned actors in ways that compress the operational gap between sophisticated APT programs and less mature groups. This has implications for how organizations should think about risk prioritization. The traditional assumption that low-sophistication groups pose lower risk because their operational capacity is constrained may no longer hold as a reliable heuristic when commercial AI tools can compensate for gaps in tradecraft, language capability, and software development skill.

Organizations should invest in threat intelligence programs capable of tracking hybrid state-criminal actors — a category that is expanding as states increasingly leverage existing criminal infrastructure for deniable operations. Pure intelligence-focused tracking of APT groups will miss a growing share of state-directed activity that flows through contractors, criminal proxies, and loosely affiliated groups like GREYVIBE. Hybrid actor tracking requires correlation of financial crime indicators, criminal-side infrastructure, and state-interest targeting patterns that traditional APT-focused programs may not be structured to perform.

At the strategic level, the AI-augmented threat landscape documented across GREYVIBE, the GTG-1002 campaign, and the broader 2025-2026 body of reporting represents a durable capability shift rather than a temporary spike. The organizations that manage this transition effectively will be those that update their threat models, detection architectures, and awareness programs ahead of the capability curve — treating AI-enhanced attacks as the new baseline rather than as a special case.

CSA Resource Alignment

The GREYVIBE campaign illustrates threat surfaces across several layers of the MAESTRO framework, CSA's seven-layer agentic AI threat modeling framework [12]. GREYVIBE's AI-augmented attack lifecycle engages MAESTRO's Layer 1 (Foundation Model) and Layer 2 (Data Operations) threat surfaces by exploiting commercial LLMs for lure generation and data synthesis, and its campaign infrastructure touches Layer 5 (Deployment and Infrastructure) by leveraging cloud-hosted services for payload delivery and command-and-control. As AI-orchestrated operations mature toward the autonomous execution pattern documented in the GTG-1002 case, MAESTRO's Layer 3 (Agentic Orchestration) and Layer 6 (Mission and Objective) threat surfaces will become increasingly relevant to campaigns of this type [12][13].

The CSA AI Controls Matrix (AICM) provides the control framework most directly applicable to managing AI-augmented threat exposure at the enterprise level [14]. The AICM's 243 control objectives across 18 security domains, which map to ISO 42001, ISO 27001, and NIST AI RMF 1.0, include controls relevant to AI system monitoring, access governance for AI tools, and supply chain security for AI-integrated environments — all of which bear directly on the risk surface that GREYVIBE and similar hybrid actors exploit [14]. CISOs should evaluate their AI tool governance posture against the AICM's controls for AI access management and behavioral monitoring, as commercial AI platform access without enterprise-level logging and policy enforcement creates an insider threat surface that adversaries have documented ability to exploit [11][14].

CSA's Agentic Trust Framework and Zero Trust Advancement Center provide architectural guidance applicable to both AI-mediated attack scenarios and the defensive infrastructure needed to counter them [15][16]. As adversaries like GREYVIBE demonstrate, the attack surface created by AI tools is bidirectional: the same platforms that threat actors use to generate malware and phishing content are also deployed inside enterprise environments where they may be abused for credential harvesting or data exfiltration if not properly governed. Zero Trust principles — explicit verification, least privilege, continuous monitoring — apply with equal force to AI agent behavior within enterprise environments as to human user behavior.

Organizations seeking a structured approach to AI security assurance should evaluate the CSA STAR for AI program, which provides Level 1 self-assessment (AI-CAIQ) and Level 2 third-party audit (ISO/IEC 42001 + Valid-AI-ted) pathways aligned to the evolving regulatory and threat landscape [17]. Given the

documented trajectory of AI-enhanced offensive operations, the question of whether an organization's AI governance program is defensible is no longer academic — it is a material risk management question with direct bearing on incident exposure and regulatory posture.

References

[1] WithSecure Labs. "GREYVIBE: A Russia-nexus group leveraging AI across state-aligned operations." WithSecure Labs, May 28, 2026.

[2] BleepingComputer Staff. "GreyVibe hackers use ChatGPT, Gemini to power cyberattacks." BleepingComputer, May 28, 2026.

[3] The Hacker News Staff. "New Russian-Linked GREYVIBE Targets Ukraine with AI-Powered Cyberattacks." The Hacker News, May 29, 2026.

[4] SecurityWeek Staff. "Russia-Linked 'GreyVibe' Attackers Use AI to Supercharge Cyberattacks." SecurityWeek, May 28, 2026.

[5] Security Affairs Staff. "Meet GREYVIBE, the Russia-Linked Hacking Group Using AI to Target Ukraine and Still Making Rookie Mistakes." Security Affairs, May 29, 2026.

[6] CrowdStrike. "2026 CrowdStrike Global Threat Report." CrowdStrike, February 24, 2026.

[7] IBM Security. "IBM 2026 X-Force Threat Index: AI-Driven Attacks Are Escalating." IBM Newsroom, February 25, 2026.

[8] SC Media Staff. "AI helps Russian-speaking GreyVibe run five parallel attack chains on Ukrainian targets." SC Media, May 29, 2026.

[9] The Record Staff. "Nation-state hackers ramping up use of Gemini for target reconnaissance, malware coding." The Record, February 12, 2026.

[10] Brightside AI. "AI Spear Phishing in 2026: Statistics, Trends and CISO Action Guide." Brightside AI Blog, 2026.

[11] Anthropic. "Disrupting the first reported AI-orchestrated cyber espionage campaign." Anthropic, November 13, 2025.

[12] Ken Huang / CSA AI Safety Initiative. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

[13] CSA AI Safety Initiative. "Applying MAESTRO to Real-World Agentic AI Threat Models." Cloud Security Alliance, February 11, 2026.

[14] Cloud Security Alliance. "AI Controls Matrix." Cloud Security Alliance, 2025.

[15] Cloud Security Alliance. "The Agentic Trust Framework: Zero Trust Governance for AI Agents." Cloud Security Alliance, February 2, 2026.

[16] Cloud Security Alliance. "Zero Trust Advancement Center." Cloud Security Alliance, 2025.

[17] Cloud Security Alliance. "CSA STAR for AI." Cloud Security Alliance, October 2025.

[18] The Register Staff. "Russia-linked threat group put ChatGPT to work from lure to payload." The Register, May 29, 2026.

[19] CSO Online Staff. "Russia-aligned crime group Greyvibe extensively uses AI in attacks." CSO Online, May 29, 2026.

[20] Google Threat Intelligence Group. "Google Reports State-Backed Hackers Using Gemini AI for Recon and Attack Support." The Hacker News, February 12, 2026.

[21] Microsoft Security. "Microsoft Digital Defense Report 2025." Microsoft, October 2025.