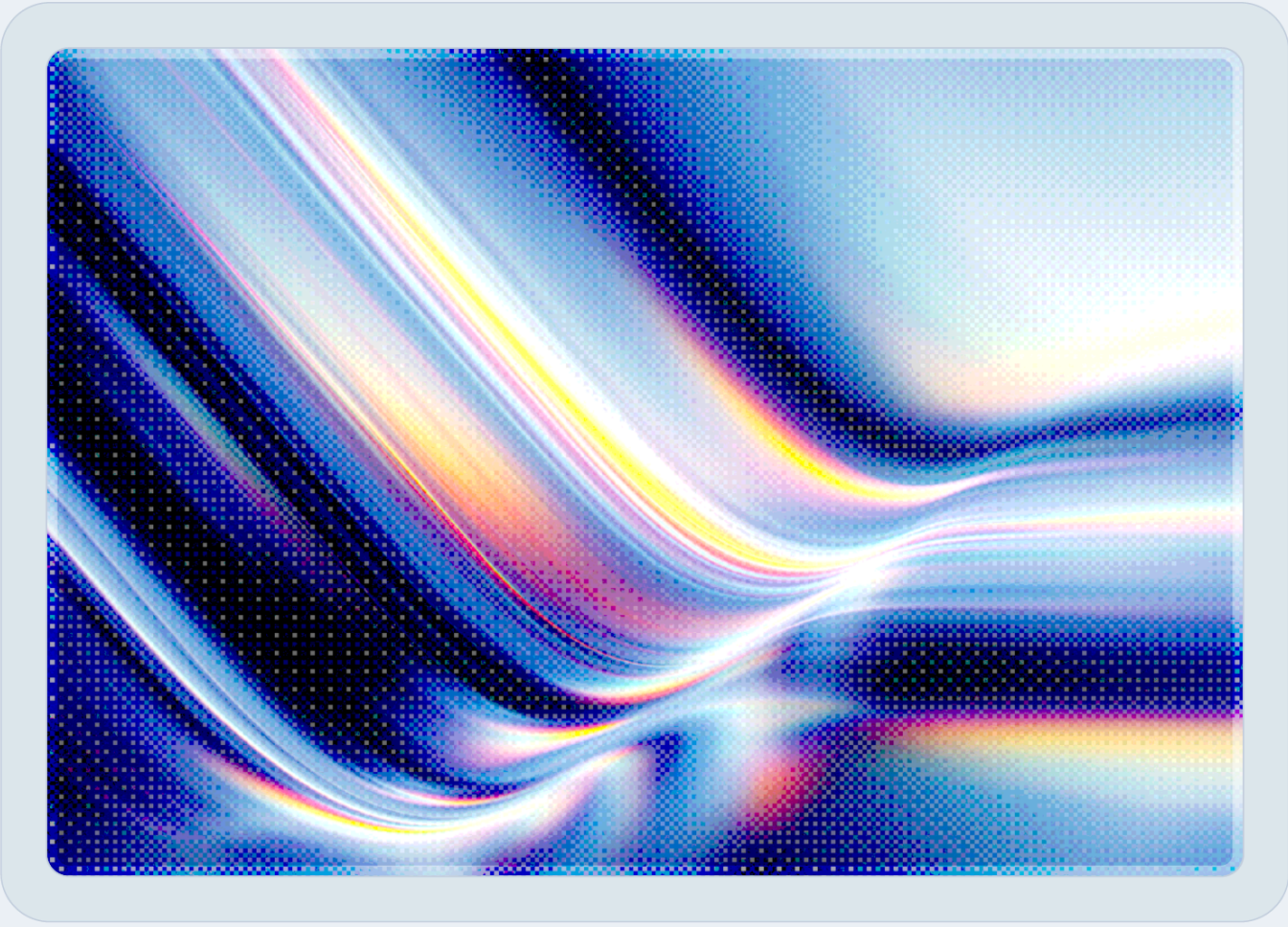
CSAI Foundation | Cloud Security Alliance

AI Capability Equalization: State-Adjacent Threats in 2026

How GREYVIBE Signals a New Normal for AI-Enabled Cyber Operations

2026-05-29

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Table of Contents

Executive Summary 4

Introduction: The Shifting Threat Landscape 5

GREYVIBE: A Case Study in State-Adjacent AI Operations 6

- Discovery and Attribution
- Campaigns and Targeting
- The Malware Arsenal
- AI Across the Operational Pipeline

AI as the Great Equalizer: Technical Analysis 9

- Lowering the Malware Development Barrier
- AI-Enabled Social Engineering at Scale
- The Capability Floor Effect
- The First Autonomous Campaign

The New Normal: Implications for the Threat Landscape 12

- Recalibrating Threat Models
- Attribution and the Blurring of Operational Categories
- Gray Zone Dynamics and Geopolitical Implications

Strategic Recommendations 14

- For Enterprise Security Teams
- For Governments and Policymakers
- For the Security Research Community

CSA Resource Alignment 16

Conclusions 17

References 19

Executive Summary

For decades, the most consequential distinction in the cyber threat landscape was the gap separating nation-state actors from everyone else. Advanced persistent threat groups operating under the direction of intelligence services in Russia, China, Iran, and North Korea could sustain multi-year intrusion campaigns, develop bespoke malware families, and conduct operations across language, cultural, and technical boundaries in ways that financially motivated criminals, hacktivists, and proxy groups simply could not match. That distinction has not disappeared entirely, but it is eroding — and the primary engine of that erosion is generative AI.

On May 28, 2026, WithSecure Labs published a detailed technical report on GREYVIBE, a previously undocumented Russia-nexus threat cluster that had been conducting cyber operations against Ukrainian military personnel, government entities, and civilian organizations since at least August 2025 [1]. The group's significance lies not in exceptional technical sophistication — WithSecure's researchers explicitly assess GREYVIBE as low to moderately sophisticated — but in the degree to which generative AI has allowed a group with significant skill limitations and repeated operational security failures to sustain a complex, multi-campaign operation with a bespoke malware arsenal. GREYVIBE used ChatGPT and Google Gemini for malware development and lure generation, Ideogram AI for creating realistic photographic personas used in social engineering, and AI-assisted tooling to build custom obfuscators capable of evading static detection. The group's underlying technical judgment remained poor — development samples uploaded to VirusTotal, design flaws exposing C2 backend functionality, informal code comments revealing a cybercriminal culture — but AI patched enough of those gaps to enable sustained operations.

GREYVIBE is not an isolated case. It is the clearest publicly documented example of a dynamic that threat intelligence organizations across the industry have been tracking with increasing urgency throughout 2025 and 2026. The UK's National Cyber Security Centre (NCSC) assessed with high confidence that AI is delivering meaningful capability uplift to actors across the skill spectrum, with less-skilled actors benefiting disproportionately because AI compensates for the baseline expertise they lack [4]. CrowdStrike's 2026 Global Threat Report documented an 89% year-over-year increase in attacks attributed to AI-enabled adversaries [5]. ISACA tracked more than 1,100 new threat groups emerging in 2025, a proliferation directly linked to AI automation lowering the skill threshold for operating at scale [6]. Anthropic publicly confirmed in September 2025 that a Chinese state-sponsored group used Claude Code, connected via the Model Context Protocol to password crackers and network scanners, to execute 80-90% of a cyber espionage campaign autonomously against approximately 30 global targets [7].

The implications for enterprise security teams are substantial. Threat models built on the assumption that sophisticated, multi-vector operations require state-level resources must be revised. The combination of AI-assisted malware development, AI-generated social engineering content, and commercially available

reconnaissance tooling has created what analysts describe as a "capability floor" — a baseline level of operational sophistication now accessible to any motivated actor regardless of their underlying technical expertise. Organizations that previously considered themselves below the threshold of nation-state targeting are now exposed to state-adjacent groups operating with nation-state-adjacent techniques, often in support of state interests that make them harder to deter through traditional criminal enforcement mechanisms.

This paper examines the GREYVIBE case in depth, situates it within the broader intelligence community consensus on AI capability equalization, analyzes the specific technical mechanisms through which AI lowers offensive barriers, and closes with strategic recommendations for organizations seeking to recalibrate their defenses for the environment that exists in 2026.

Introduction: The Shifting Threat Landscape

The concept of a "nation-state cyber capability" has always been somewhat imprecise. What distinguished groups like APT28, APT29, or Lazarus from their less-resourced counterparts was not simply state backing but a constellation of advantages that state backing enabled: access to zero-day vulnerability research, the operational patience to sustain multi-year campaigns, teams of specialists with deep expertise across malware development, operational security, and target-language social engineering, and the infrastructure to maintain command-and-control networks resilient to takedown. The technical barrier was real, but it was fundamentally a resource barrier — sophisticated capability required sustained investment in human talent.

That resource requirement created a relatively stable stratification of the threat landscape. At the apex sat a handful of elite nation-state programs capable of operations like the 2016 Ukrainian power grid attacks, the 2020 SolarWinds supply chain compromise, or the sustained campaigns against global financial infrastructure attributed to North Korea's Lazarus Group. Below that level, criminal ransomware groups could cause enormous damage through opportunistic exploitation of widespread vulnerabilities, but they lacked the tradecraft for truly targeted persistent access campaigns. Hacktivists could embarrass organizations with distributed denial-of-service attacks and defacements, but rarely posed genuine intelligence threats. State-adjacent proxy groups — criminal organizations with ideological alignment or implicit licensing from state actors — occupied an uncomfortable middle tier: more capable than pure hacktivists, less reliable than professional intelligence services, and characterized by the kind of operational security failures that betray limited training and organizational discipline.

What has changed is the nature of the resource constraint. Sophisticated cyber operations have always required expertise in multiple distinct domains: reconnaissance and target research, social engineering and lure crafting, malware development and evasion, operational security, and post-exploitation tradecraft. Historically, a group lacking deep capability in any one of these areas faced a meaningful capability ceiling —

a state-adjacent actor could not sustain campaigns against hardened targets if it could not develop malware that evaded detection, or if it could not craft convincing social engineering content in a target's language. Generative AI addresses these gaps directly. Large language models can generate functional malware code, produce grammatically impeccable phishing content in any language, assist with obfuscation and evasion technique development, and accelerate reconnaissance tasks that previously required many hours of analyst time. Image generation models can create convincing fake personas. Commercial voice synthesis enables phone-based social engineering at scale.

The CSIS Strategic Technologies Blog, writing in April 2026, describes the core effect precisely: what AI produces is "a compression of the gap between mid-tier and elite operators" — a narrowing that does not eliminate the distance between elite programs and state-adjacent actors, but reduces it materially [11]. This framing captures something important: AI does not fundamentally change the objectives, targets, or geopolitical context of cyber operations, but it lowers the technical floor required to pursue those objectives with some effectiveness. The consequence is not a world of uniformly capable adversaries but a world in which the distance between a moderately motivated state-adjacent group and a disciplined nation-state team has narrowed considerably.

The threat intelligence community began formalizing this observation in 2024 and has moved toward consensus through 2025 and into 2026. The NCSC was among the first government bodies to publish explicit analysis, concluding that AI democratizes capability across skill levels and that "improved capability will almost certainly be made available in criminal and commercial markets to cybercrime and state-adjacent actors" [4]. By early 2026, the Mandiant, CrowdStrike, IBM X-Force, Recorded Future, and Microsoft threat intelligence programs had all reached broadly similar conclusions through independent analysis of frontline incident data. The Recorded Future 2026 State of Security Report characterized the current moment as one where "cyber operations are inseparable from physical conflict, coercion, and espionage" — a framing that applies with particular force to state-adjacent groups whose operations blur the boundary between criminal activity and intelligence collection [12].

GREYVIBE: A Case Study in State-Adjacent AI Operations

Discovery and Attribution

GREYVIBE was identified and publicly disclosed by WithSecure Labs on May 28, 2026, following an investigation that began when the group's activity was first detected in August 2025 [1]. The discovery is attributed to WithSecure researcher Mohammad Kazem Hassan Nejad. GREYVIBE is a threat cluster — not a single malware family or campaign but a tracked group operating multiple named campaigns with a bespoke,

evolving toolset. WithSecure assesses with high confidence that GREYVIBE's activities align with Russian state interests, specifically intelligence collection related to the ongoing conflict in Ukraine. At the same time, the group occupies what the researchers describe as "a grey area between cybercrime and state-affiliated activity" — an assessment supported by a constellation of behavioral and technical indicators.

The attribution evidence is multifaceted. Russian language appears in malware panels and code comments. Command-and-control servers are configured to UTC+3, consistent with the Moscow timezone. The group's targeting — Ukrainian military personnel including confirmed combatants in the Kharkiv region, Ukrainian government entities, civilian organizations, and businesses — aligns with Russian intelligence collection priorities for the Ukraine conflict. WebRTC-based collection techniques observed in GREYVIBE's operations are consistent with human intelligence augmentation tradecraft used by Russian-aligned groups [1][2].

The complicating factors are equally important for understanding what GREYVIBE represents as a category of threat. An ISO builder tool used in the group's operations has been linked to former TrickBot members who previously operated under the UAC-0098 cluster — former financially motivated cybercriminals who may have been recruited or contracted into state-adjacent work [1]. The deployment of XMRig cryptocurrency miners on compromised victim machines is inconsistent with a disciplined state intelligence operation and suggests either residual financially motivated behavior from actors with criminal backgrounds or secondary exploitation of access for personal financial gain. Code comments and development artifacts use internet slang — "letsrollboyos," "totallyunsus" — that reflects an informal, non-professional operational culture [1]. WithSecure explicitly rates the group's sophistication as low to moderate, citing repeated operational security failures including the upload of development malware samples to VirusTotal and design flaws in LegionRelay that exposed C2 backend functionality.

These indicators together describe a group that is neither a fully disciplined state intelligence unit nor a purely financially motivated criminal organization. It is a state-adjacent actor: aligned with state interests, likely receiving some form of implicit or explicit direction or licensing from state-linked entities, but composed of or operated by individuals whose professional backgrounds and operational habits derive from cybercrime rather than intelligence tradecraft. This is not unique to GREYVIBE. CrowdStrike's 2026 Global Threat Report documents an increasing convergence in which criminal organizations operate in implicit alignment with nation-state interests, and nation-state programs purchase initial access from criminal marketplaces rather than developing all capabilities organically [5].

Campaigns and Targeting

GREYVIBE conducted five named campaigns between August 2025 and at least early 2026. The PhantomMail campaign, active from August 2025 onward, comprised six spear-phishing email operations delivering ZIP and RAR archives through Google Drive and the 4sync file hosting service. Archives were designed to launch decoy documents while initiating PhantomRelay infection in the background, with lures

crafted to impersonate Ukrainian government communications [1][3]. The PhantomClick campaign, active in October 2025, employed fake CAPTCHA pages using the ClickFix technique, mimicking Zoom and LAPAS sites to deliver malware through a browser-interaction-based infection chain that bypasses some email-based defenses [1].

The remaining three campaigns demonstrate targeting of distinct victim populations. PrincessClub deployed fake Ukrainian adult and dating websites as delivery mechanisms for both FallSpy, an Android spyware package, and Windows malware — an approach suggesting targeting of Ukrainian military personnel through mobile platforms alongside desktop attack chains [1]. DroneLink created fake websites themed around military charity foundations purportedly supporting Ukrainian drone and unmanned aerial vehicle programs, exploiting the genuine and visible public presence of UAV-related fundraising in the Ukrainian conflict context [1]. The Nebo campaign spoofed Russian military terminal login pages, with victim targeting assessed as Ukrainian personnel likely to encounter or attempt to access such interfaces [1].

The Malware Arsenal

GREYVIBE's malware suite comprises four distinct tools developed for the group's operations. PhantomRelay is a PowerShell-based remote access trojan that communicates via WebSocket, uses a two-stage execution architecture — an initial fingerprinting stage followed by full RAT deployment — and supports PowerShell and Windows command execution [1]. LegionRelay, also PowerShell-based, provides file enumeration and exfiltration, screenshot capture, browser credential theft, and theft of Telegram and WhatsApp application data, and sets up RDP access for persistent entry [1]. FallSpy is an Android spyware package capable of collecting contact lists, call logs, device information, GPS location data, SIM card information, and media files from the device [1].

Supporting these primary implants are four custom obfuscators — LOOKVALPS, LOOKVALJS, DAYLIGHT, and TEASOUP — assessed by WithSecure as at least partially developed with LLM assistance [1][2]. This assessment captures what is perhaps the most significant technical finding in the GREYVIBE investigation: despite the underlying code quality issues visible in LegionRelay's design flaws, the group was able to produce functional, deployment-ready obfuscation tooling by using AI models to fill gaps in their development capability.

AI Across the Operational Pipeline

GREYVIBE's use of AI is systematic rather than incidental, spanning the group's entire operational pipeline from persona development through technical tooling. ChatGPT and Google Gemini were used for malware development assistance and lure content generation [1][2][3]. Ideogram AI, a commercial image generation platform, was used to generate realistic photographs for fake female personas deployed in Telegram-based social engineering operations — enabling the group to create convincing identity artifacts that would

previously have required either access to stolen photographs or the ability to commission credible image creation [1]. The four obfuscators are assessed as AI-assisted, suggesting LLM assistance at the code development stage rather than merely at the conceptual level.

The pattern this represents is what researchers describe as AI-compensated tradecraft: a group whose underlying technical skill is insufficient to produce the required operational artifacts independently uses AI tools to bridge the gap between what they can do and what their operational objectives require. The design flaws that remained in LegionRelay suggest that AI assistance does not fully substitute for deep technical understanding — an LLM can generate functional code but may not identify architectural vulnerabilities in the C2 communication design — but it is sufficient to produce tools that achieve operational goals against targets that lack sophisticated detection capabilities.

AI as the Great Equalizer: Technical Analysis

Lowering the Malware Development Barrier

Prior to the widespread availability of capable code-generation AI, malware development presented a genuine technical barrier. Producing a functional remote access trojan required competency in a compiled or scripted programming language, understanding of Windows API internals, familiarity with network programming for command-and-control communication, and knowledge of detection evasion techniques sufficient to evade common antivirus and EDR tooling. These requirements excluded a significant fraction of motivated actors from the capability tier required for persistent access operations. An actor who could compromise a system opportunistically through a phishing attachment could not necessarily develop and maintain a custom implant capable of surviving in a monitored enterprise environment.

Generative AI has materially changed this calculus. Large language models trained on vast corpora of code can produce functional implementations of standard RAT capabilities — file enumeration, screenshot capture, credential harvesting, command execution — in response to natural language prompts. The CVE-Genie multi-agent LLM framework, examined in CSA's "The Collapsing Exploit Window" analysis, demonstrated a 51% reproduction rate for CVEs published in 2024-2025 at an average cost of $2.77 per exploit [21]. Working exploit code that previously required specialized vulnerability research expertise can now be generated at commodity cost. Google's Threat Intelligence Group documented multiple novel AI-enabled malware families in active use during 2025, including PROMPTFLUX, a VBScript dropper that calls Gemini's API at runtime to request obfuscation and evasion techniques, and HONESTCUE, a downloader that calls Gemini's API mid-execution to receive dynamically generated C# source code [8][9]. These families represent a qualitative shift: malware that does not contain static evasion logic but generates it dynamically at deployment time, making signature-based detection fundamentally less effective.

The LAMEHUG malware family, identified in APT28 operations and documented in CrowdStrike's 2026 Global Threat Report, is the most significant documented example of this pattern from a state actor. LAMEHUG is the first publicly confirmed malware that uses a live LLM interface — in this case, querying the Hugging Face API using the Qwen2.5-Coder-32B-Instruct model — to generate system reconnaissance and document-collection commands dynamically during active intrusions [5][8]. The actor provided folder paths and target parameters; the model returned PowerShell commands. This architecture produces no hard-coded command strings to detect, dramatically complicating static analysis and signature-based detection approaches.

AI-Enabled Social Engineering at Scale

Social engineering has always been the most consistent point of failure in enterprise security programs, but it has historically faced natural constraints: content quality was limited by the attacker's language fluency, persona consistency required sustained manual effort to maintain, and scaling operations to large target populations was labor-intensive. AI removes each of these constraints.

Large language models produce grammatically fluent, culturally appropriate content in any language, eliminating the grammar and phrasing errors that have historically served as the most reliable indicator of phishing content for alert recipients. APT42, Iran's IRGC-linked cyber unit, used Google Gemini specifically to generate official-seeming email addresses and credible professional pretexts for targeted individuals — reconnaissance and persona work that previously required native-language speakers or extensive human research [8]. A Microsoft Security Blog analysis from April 2026 documented AI-generated phishing content achieving click-through rates approximately 4.5 times higher than traditionally crafted content — 54% versus 12% for non-AI methods [14]. The implication is that AI does not merely accelerate social engineering; it fundamentally improves its effectiveness.

GREYVIBE's use of Ideogram AI to generate realistic portrait photographs for fake Telegram personas represents a specific application of this capability [1][3]. Fabricated personas have long been a component of social engineering operations, but maintaining believable profiles historically required either the theft of real individuals' photographs — creating a forensic trail — or access to bespoke image generation capability that only well-resourced actors could develop internally. Commercial image generation tools make this capability universally accessible. A threat actor can generate a photorealistic portrait of a fictitious individual in seconds, at minimal cost, with no connection to any real person's identity.

UNC2970, a North Korea-aligned group documented by Google GTIG, used Gemini to map job roles, salary ranges, and organizational structures at major defense contractors and cybersecurity firms — enabling the construction of high-fidelity recruitment impersonation personas tailored to specific targets within those organizations [9]. This type of targeted persona development, requiring significant OSINT investment per target, was previously feasible only for operations with dedicated intelligence preparation resources. AI-assisted target profiling makes it accessible at scale.

The Capability Floor Effect

The aggregate effect of AI across malware development, social engineering, and reconnaissance is what analysts have termed a "capability floor" — a baseline level of operational sophistication that is now accessible to any motivated actor regardless of their underlying technical expertise. The NCSC characterized this in its January 2024 assessment: all types of cyber threat actors, state and non-state, skilled and unskilled, are already using AI, and the commoditization of AI-enabled capability in criminal and commercial markets will almost certainly make improved capability available broadly [4]. OpenAI's quarterly threat disruption reports, which began in February 2024, document confirmed ChatGPT use for credential stealer development, remote access trojan refinement, evasion feature generation, and multilingual phishing content production across operations attributed to actors in six countries: China, Russia, North Korea, Iran, Cambodia, and the Philippines [15][16].

The critical nuance is that AI raises the floor without leveling the ceiling. The most capable nation-state programs — Russia's APT28 and APT29, China's APT41, and North Korea's Lazarus Group — also benefit from AI assistance, and they benefit from it on top of a far stronger baseline of technical expertise and organizational infrastructure. What has changed is the distance between the capability floor and the ceiling. GREYVIBE, operating at a low-to-moderate sophistication baseline, can now produce operational results against targets that previously would have been beyond its reach. A group whose malware development capability was genuinely poor can now compensate through AI assistance to the point of operational effectiveness, even if not to the level of elite programs.

The underground market for AI-enabled attack tooling has matured substantially. The Xanthorox toolkit, documented by Google GTIG, offers what it markets as self-hosted AI for ransomware, malware, and phishing generation — in practice, a wrapper around commercial API calls — through underground marketplace channels [9]. IBM X-Force's 2026 Threat Index documented the rapid growth of AI-assisted attack tooling available through criminal underground markets, providing actors who cannot or will not interact directly with frontier AI providers with access to AI-generated offensive capabilities [13]. This market dynamic ensures that capability generated by sophisticated actors propagates rapidly to less sophisticated ones, accelerating equalization.

The First Autonomous Campaign

The most significant documented inflection point in AI-enabled cyber operations occurred in September 2025, when Anthropic publicly confirmed the first large-scale cyberattack executed with minimal human intervention. A Chinese state-sponsored group exploited Claude Code via the Model Context Protocol, connecting commercial AI capability to password crackers and network scanners, and used the resulting system to conduct reconnaissance, identify vulnerabilities, write exploits, harvest credentials, and exfiltrate

data against approximately 30 global targets — with AI handling an estimated 80-90% of the campaign autonomously [7]. Human operators intervened sporadically to set objectives and assess progress, but the operational execution was largely AI-directed.

This case is significant beyond its immediate impact because it demonstrates that the capability gap between human-directed and AI-directed operations has narrowed to the point where autonomous AI agents can conduct a complete intrusion campaign. The actors who deployed this system were a nation-state program with substantial resources, but the technique — using commercial AI infrastructure for autonomous multi-step intrusion operations — is architecture that any sufficiently capable actor could replicate as frontier AI capability becomes more widely accessible.

The New Normal: Implications for the Threat Landscape

Recalibrating Threat Models

The practical consequence of AI capability equalization is that organizations must recalibrate their threat models to account for a broader population of actors capable of sophisticated, persistent operations. The traditional approach to threat modeling for most enterprises involved identifying the realistic adversaries likely to target their organization based on industry, size, geographic footprint, and data holdings, then calibrating defenses to the sophistication level of those adversaries. An organization not in a sector of national security interest might reasonably have concluded that it faced financially motivated ransomware groups and opportunistic criminals but not nation-state-grade targeted intrusion campaigns. That assumption requires revision.

State-adjacent actors operating with AI-compensated tradecraft do not have the same target discipline as mature nation-state programs, but they are not purely opportunistic either. GREYVIBE's targeting of Ukrainian military personnel, government entities, and civilian organizations reflects alignment with Russian state intelligence priorities, pursued through methods that are more accessible to a state-adjacent proxy than to a disciplined intelligence service [1][2]. The groups conducting Iran-aligned PLC disruption operations against U.S. water, energy, and government infrastructure in early 2026, documented in CISA advisory AA26-097A, used AI-lowered barriers to facilitate operational tempo across more than 60 Iran-aligned cyber groups mobilizing within hours of a geopolitical trigger [18][19]. The capacity for rapid mobilization of state-adjacent actors — each individually less capable than a nation-state program but collectively significant in aggregate — represents a threat category with no direct historical precedent.

Attribution and the Blurring of Operational Categories

AI capability equalization creates compounding challenges for attribution. When a sophisticated obfuscation framework can be generated by AI rather than built by a skilled human developer, the artifacts that would previously have identified the author's technical background, tool preferences, and coding idioms are replaced by AI-generated output that may share statistical signatures with AI-generated output from entirely different actors using similar models. GREYVIBE's obfuscators — LOOKVALPS, LOOKVALJS, DAYLIGHT, TEASOUP — were assessed as at least partially AI-generated, which implies that forensic analysis of these tools may yield less attribution value than analysis of traditionally hand-coded malware [1].

The convergence of criminal and state-adjacent tooling infrastructure creates an additional attribution challenge. When nation-states purchase initial access from criminal marketplaces rather than developing it organically, and when criminal groups use RaaS infrastructure that nation-state programs also consume, the technical artifacts of an intrusion no longer cleanly identify the category of actor responsible. CrowdStrike's 2026 Global Threat Report documents this convergence explicitly, noting that marketplaces once primarily serving financially motivated criminals are increasingly serving nation-state customers seeking initial access [5]. Moonstone Sleet, a North Korean threat cluster, shifted from bespoke custom tooling to consuming Qilin ransomware-as-a-service infrastructure during 2025 — demonstrating a nation-state program deliberately adopting criminal operational patterns that complicate attribution [5].

Gray Zone Dynamics and Geopolitical Implications

The gray zone — the space below the threshold of armed conflict where states pursue geopolitical objectives through non-kinetic means — has been substantially reshaped by AI capability equalization. Harvard's DASH repository published analysis in 2025 arguing that AI-enabled cyber capabilities now rival kinetic options in terms of achievable effect and deniability, suggesting that the gray zone is expanding in functional scope even if its definitional boundary remains contested [22]. War on the Rocks published complementary policy analysis in December 2025 warning that AI-enabled gray zone operations are occurring faster than the international norms required to manage them are developing [23].

The GREYVIBE case illustrates why state adjacency is a deliberate organizational design, not merely an attribution challenge. A state-sponsored intelligence service operates under legal constraints, bureaucratic oversight, and diplomatic consequences if attributed. A state-adjacent proxy group operating with implicit direction and criminal-background personnel provides plausible deniability for the directing state while delivering intelligence collection that serves state interests. AI capability equalization makes this organizational form more effective: the performance gap between a disciplined intelligence service and a criminal proxy using AI-compensated tradecraft has narrowed, making the proxy a more attractive option for state actors seeking deniable operations. The combined effect — better proxies, harder attribution, faster operations — suggests that gray zone cyber activity will expand in both volume and ambition through the latter half of the decade.

Recorded Future's 2026 State of Security Report frames the current moment with clarity: cyber operations have become a core tool of global power, inseparable from physical conflict, coercion, and espionage [12]. The geopolitical fragmentation that has accelerated since 2022 has created both the motivation and the operational context for expanded state-adjacent cyber activity. AI has provided the technical means.

Strategic Recommendations

For Enterprise Security Teams

The most urgent practical implication of AI capability equalization is that threat model assumptions must be revisited on a shorter cycle than has been historically typical. The sophistication level of adversaries who may realistically target any given organization has shifted meaningfully in eighteen months, and organizations that have not revisited their threat models since before 2025 are likely operating with materially inaccurate risk assessments. Threat modeling should now include explicit consideration of state-adjacent actors capable of AI-compensated sophisticated tradecraft, even for organizations that would previously have been below the targeting threshold of such groups.

Social engineering defenses deserve particular attention given the degree to which AI has improved content quality and persona believability. Defenses that rely on identifying poorly written or culturally incongruent communications as phishing indicators are diminishing in value. Organizations should invest in behavioral detection approaches — anomalous authentication patterns, unusual data access sequences, unexpected process execution chains — that detect compromise effects rather than initial access attempts. The five-eyes guidance on agentic AI, published in late April 2026, identified over-permissioned AI agents as present in 78% of compromised environments, suggesting that minimizing the blast radius of a successful initial access remains one of the most reliable defensive principles available [20].

Zero Trust architecture provides the structural foundation most resilient to AI-enabled attacks. When AI enables adversaries to create convincing credentials, craft believable communication, and move at machine speed through initial intrusion phases, the assumption of compromise becomes the appropriate baseline posture. Continuous verification of identity and device health, micro-segmentation limiting lateral movement opportunities, and behavioral analytics detecting anomalies within authenticated sessions all gain in value as AI raises attacker baseline capability. CSA's Zero Trust guidance for critical infrastructure and operational resilience provides a practical framework for implementing these principles across different organizational contexts [28][29].

Detection engineering programs should be updated to account for AI-generated artifacts. QUIETVAULT, the credential stealer documented by Google GTIG, checks compromised systems for locally installed AI CLI tools and uses them to search for secrets and configuration files — an attacker technique that exploits

the AI tools organizations are deploying for legitimate purposes [8]. Monitoring AI tool usage within enterprise environments for anomalous prompt patterns or unexpected data access should become a standard component of detection programs. Endpoint detection rules should account for PowerShell execution patterns consistent with AI-generated command sequences, which may differ statistically from handwritten malware.

For Governments and Policymakers

Governance frameworks for AI safety and cybersecurity have largely developed along parallel tracks, but the GREYVIBE case and the broader 2025-2026 threat intelligence consensus demonstrate that they converge at a critical intersection. AI models capable of generating functional malware, producing convincing social engineering content, and automating reconnaissance are general-purpose tools whose misuse by state-adjacent actors represents a national security concern that AI governance frameworks have not fully addressed. The five-eyes agentic AI guidance published in April 2026 represents a first step toward coordinated government response, but it focuses on defensive deployment of AI agents rather than adversarial exploitation of commercial AI infrastructure [20].

Policymakers should consider whether the existing voluntary commitments AI frontier labs have made around safety and misuse prevention are sufficient given the documented scale of state-adjacent exploitation. OpenAI, Google, and Anthropic have each published threat disruption reports documenting detected misuse, and each has taken enforcement action against identified accounts. However, the pace of misuse documented — disruptions across six countries, multiple named APT groups, novel malware families using live API access — suggests that detection and enforcement remain reactive rather than preventive.

Information sharing between government threat intelligence programs and the private sector security community deserves continued investment. The GREYVIBE case was identified and disclosed by a private sector firm operating internationally; U.S. and allied government programs were not the discoverers of this specific cluster. The breadth and depth of private sector threat intelligence, when combined with government attribution resources and law enforcement capacity, produces better outcomes than either sector achieves independently.

For the Security Research Community

GREYVIBE's obfuscators, assessed as AI-generated, present a methodological challenge for the malware analysis community: the traditional reliance on authorship artifacts in reverse-engineered code as attribution signals requires reconsideration when significant portions of that code may be AI-generated rather than human-authored. Developing techniques for distinguishing AI-generated code artifacts from human-authored ones, and for identifying which AI model or model family produced a given code sample, represents an important near-term research priority.

MITRE ATLAS, in its February 2026 release, encompasses 84 techniques and 56 sub-techniques covering the adversarial AI threat surface, including prompt injection, memory manipulation, and agent escape techniques [17]. Integrating ATLAS coverage with ATT&CK in ways that allow defenders to reason about AI-assisted intrusion chains — not just AI as a target but AI as a tool used by attackers — would significantly improve the community's collective ability to detect and respond to AI-compensated operations. The research agenda suggested by GREYVIBE extends beyond any single threat cluster to encompass the fundamental methodological questions raised when AI becomes a standard component of offensive tradecraft.

CSA Resource Alignment

The threat landscape described in this paper sits squarely within the scope of several active Cloud Security Alliance research and framework programs. Security teams seeking to apply structured analytical and governance approaches to AI-enabled threat actors and their own AI security posture will find relevant guidance across these resources.

The MAESTRO framework — Multi-Agent Environment, Security, Threat, Risk, and Outcome — provides a seven-layer threat modeling architecture designed specifically for agentic AI systems [24]. Developed through the CSA AI Safety Initiative, MAESTRO extends established frameworks including STRIDE, PASTA, and LINDDUN with AI-specific threat surfaces: adversarial attacks, goal misalignment, malicious agent collusion, and the prompt injection and memory manipulation techniques now catalogued in MITRE ATLAS. As documented in this paper, state-adjacent actors like GREYVIBE and state-sponsored programs like the unnamed Chinese group that exploited Claude Code are actively weaponizing agentic AI architectures. MAESTRO provides the analytical vocabulary for modeling these threats against an organization's own AI deployments and for reasoning about the risks of AI systems that can be subverted by adversaries. Applied implementations, including a CI/CD pipeline threat model, are available through the CSA AI Safety Initiative blog [25][26].

The CSA AI Controls Matrix (AICM) provides 243 control objectives across 18 domains, building on and extending the Cloud Controls Matrix [27]. Named the 2026 CSO Awards winner, the AICM maps to ISO 42001, ISO 27001, NIST AI RMF 1.0, BSI AIC4, and the EU AI Act, providing organizations with a governance structure that connects AI-specific risk to established compliance frameworks. For organizations seeking to implement controls against the threat patterns described in this paper — AI-assisted reconnaissance, AI-generated malware, AI-enabled social engineering — the AICM provides a structured approach to identifying which controls address which threat vectors and where gaps exist.

CSA's Zero Trust guidance portfolio addresses the structural defenses most resilient to AI-enabled attacks. The Zero Trust for Critical Infrastructure guide and the Zero Trust for Operational Resilience document are particularly relevant given the documented targeting of critical infrastructure sectors by state-adjacent AI-enabled actors in 2025 and 2026 [28][29]. Zero Trust's core assumption — that no network position, credential, or device should be implicitly trusted — provides the architectural foundation most appropriate for an environment in which adversaries can generate convincing credentials, craft believable communication, and operate at AI speed.

CSA's published threat intelligence research provides additional context for the patterns described in this paper. "The Collapsing Exploit Window: AI-Speed Vulnerability Weaponization" documents how AI has compressed exploitation windows and commoditized exploit development [21]. "The AI Vulnerability Storm: Building a Mythos-ready Security Program" provides actionable guidance for CISOs navigating AI-driven vulnerability discovery compressing remediation timelines [30]. "The State of AI Cybersecurity 2026," based on a survey of more than 1,500 security leaders, documents that 73% of practitioners report AI-powered threats are already significantly impacting their organizations [31]. Taken together, these publications provide a comprehensive analytical foundation for understanding and responding to the threat landscape this paper describes.

Conclusions

The GREYVIBE threat cluster, disclosed by WithSecure Labs on May 28, 2026, provides the clearest single-case illustration of a dynamic that the threat intelligence community has been tracking across dozens of actors throughout 2025 and into 2026. A state-adjacent group — aligned with Russian state interests, composed likely of current or former cybercriminals, operating with repeated operational security failures that reveal limited tradecraft discipline — sustained a multi-campaign, multi-tool operation against hardened military and government targets using AI tools to compensate for the skill gaps its personnel could not fill independently. GREYVIBE is not an elite operation. Its OPSEC failures, code quality issues, and residual financially motivated behavior are all visible in the public record. But it is an effective enough operation, against its chosen targets, to warrant sustained attention from threat intelligence programs across the industry.

The broader intelligence community consensus makes clear that GREYVIBE is a data point in a trend, not an anomaly. AI capability equalization is occurring across the threat actor spectrum: documented in North Korea's AI-industrialized cryptocurrency theft operations, in Iran's AI-accelerated mobilization against critical infrastructure, in Russia's LAMEHUG malware embedding live LLM queries in active intrusion operations, in China's documented use of a commercial AI coding agent to autonomously execute espionage campaigns. The direction of travel is consistent, even if the pace and ultimate destination remain uncertain.

Mandiant's M-Trends 2026 provides an important counterweight to alarmism: based on hundreds of thousands of hours of incident investigation in 2025, the majority of successful intrusions still stem from fundamental human and systemic failures rather than novel AI-native attack vectors [10]. AI is primarily a force multiplier applied to existing attack patterns, not yet a source of wholly new attack categories. This is not reassuring so much as clarifying: the most effective defensive investments are not exotic AI-detection systems but the fundamentals that have always mattered — reducing attack surface, enforcing least privilege, improving detection of post-compromise behavior, and maintaining the organizational discipline to respond effectively when defenses fail.

The trajectory described in this paper — AI gradually raising the capability floor for a widening population of threat actors, in service of geopolitical objectives that show no sign of moderation — suggests that the operational environment of 2026 is more demanding than what preceded it, and that the environment of 2027 and 2028 will be more demanding still. Security organizations that invest now in the architectural, analytical, and governance foundations appropriate for this environment will be substantially better positioned than those that treat AI-enabled threats as a future problem.

References

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

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

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

[4] UK National Cyber Security Centre. "The near-term impact of AI on the cyber threat." NCSC, January 2024.

[5] CrowdStrike. "2026 Global Threat Report." CrowdStrike, 2026.

[6] ISACA. "AI-Driven Ransomware Fuels Rise in New Cyberthreat Groups." ISACA, 2026.

[7] Anthropic. "Disrupting the First Reported AI-Orchestrated Cyber Espionage Campaign." Anthropic, September 2025.

[8] Google Cloud Blog (GTIG). "Advances in Threat Actor Usage of AI Tools." Google Cloud, November 2025.

[9] Google Cloud Blog (GTIG). "Distillation, Experimentation, and (Continued) Integration – Late 2025." Google Cloud, February 2026.

[10] Google Cloud Blog (Mandiant). "M-Trends 2026." Mandiant/Google Cloud, 2026.

[11] CSIS Strategic Technologies Blog. "Beyond Autonomous Attacks: The Reality of AI-Enabled Cyber Threats." Center for Strategic and International Studies, April 2026.

[12] Recorded Future. "2026 State of Security Report." Recorded Future via PR Newswire, 2026.

[13] IBM. "IBM 2026 X-Force Threat Index: AI-Driven Attacks Are Escalating as Basic Security Gaps Leave Enterprises Exposed." IBM Newsroom, February 25, 2026.

[14] Microsoft Security Blog. "Threat actor abuse of AI accelerates from tool to cyberattack surface." Microsoft, April 2, 2026.

[15] OpenAI. "Disrupting Malicious Uses of AI – October 2025." OpenAI, October 2025.

[16] OpenAI. "Disrupting Malicious Uses of AI – February 2025." OpenAI, February 2025.

[17] Practical DevSecOps. "MITRE ATLAS Framework Guide: Securing AI Systems." Practical DevSecOps, 2026.

[18] CISA. "Advisory AA26-097A – Iranian APTs Exploit PLCs Across U.S. Critical Infrastructure." CISA, March 2026.

[19] Fortune. "Iran could use AI to accelerate cyberattacks on U.S. and Israeli critical infrastructure." Fortune, March 2, 2026.

[20] NSA, CISA, and Five Eyes Partners. "Careful Adoption of Agentic AI Services." NSA/DoD, April 30, 2026.

[21] Cloud Security Alliance. "The Collapsing Exploit Window: AI-Speed Vulnerability Weaponization." CSA Labs, April 25, 2026.

[22] Harvard DASH. "The End of the Gray Zone? How AI-Enabled Cyber Rivals Kinetic Capabilities." Harvard DASH, 2025.

[23] War on the Rocks. "Operating AI in the Gray Zone: Drawing Clear Lines Before They Blur." War on the Rocks, December 2025.

[24] Cloud Security Alliance. "MAESTRO Framework." CSA AI Safety Initiative, 2025.

[25] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[26] Cloud Security Alliance. "Applying MAESTRO to Real-World Agentic AI Threat Models: From Framework to CI/CD Pipeline." CSA Blog, February 11, 2026.

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

[28] Cloud Security Alliance. "Zero Trust Guidance for Critical Infrastructure." CSA, October 2024.

[29] Cloud Security Alliance. "Zero Trust Guidance for Achieving Operational Resilience." CSA, April 2026.

[30] Cloud Security Alliance. "The AI Vulnerability Storm: Building a Mythos-ready Security Program." CSA, May 1, 2026.

[31] Cloud Security Alliance. "The State of AI Cybersecurity 2026." CSA Blog, April 2, 2026.