HOWDEN

AI.Attackers:

Technical Research Report

Industrialization of Autonomous Cyber Threats and Implications for Pre-Attack Defense

Classification: Research
Author framing: Cyber security research perspective
Date: February 2026
Sources: Anthropic threat intelligence (2025), Malanta Intelligence datasets, peer-reviewed literature (Guembe et al., 2022; kill-chain mapping).

Malanta:

Malanta.ai Pre-Attack Prevention research highlights
in collaboration with Howden Insurance Brokers Limited

Executive Summary

This report provides a technical assessment of the shift from automated to autonomous AI-orchestrated cyber operations (hereafter AI.Attackers).

We present primary evidence that offensive use of AI has moved from laboratory and proof-of-concept to operational deployment at scale, with documented campaigns in which AI agents execute the majority of attack lifecycle tasks with limited human intervention.

We then quantify the industrialization of attack infrastructure (domains, certificates, code repositories, accounts) and its correlation with the maturation of generative AI (GenAI) and large language models (LLMs).

Finally, we frame findings in terms of Indicators of Pre-Attack (IoPAs) versus Indicators of Compromise (IoC) and argue that pre-attack prevention—detection and disruption of staged infrastructure before first contact—is the only defense layer that operates in the same time window as modern threat actors.

Key technical conclusions:

Operational autonomy:

A single documented campaign (Anthropic GTG-1002) demonstrated ~80–90% of tasks executed by an AI agent across reconnaissance, exploitation, credential harvesting, lateral movement, and exfiltration.

GenAI correlation:

Malicious GitHub repositories increased from 6,498 (2022) to ~110,000 (2024) — ~1,592% growth— with a sharp inflection in 2023–2024 aligned with public LLM/GenAI availability; similar acceleration observed across GitLab and Bitbucket.

Infrastructure scale:

Analysis of 33 attack infrastructure clusters shows mean composition of ~61 domains, ~88 subdomains, ~35 SSL certificates, and ~11 social media accounts per cluster; ~82% of cluster-associated domains had not triggered vendor detections at time of analysis.

Control-plane relevance:

The observed activity maps to MITRE ATT&CK TA0042 (Resource Development) and precedes traditional kill-chain phases (reconnaissance through action on objectives).
Defenses that rely on IoCs alone are structurally late; IoPA-based pre-attack prevention addresses the same phase in which adversaries build and stage infrastructure.

Setup window:

Median time from domain registration to overtly malicious use is 72 days , defining a measurable window for pre-attack intervention.

Malanta:

Terminology and Scope

AI.Attacker:

An autonomous or highly automated AI agent that conducts offensive cyber operations with minimal human-in-the-loop control.

Attack infrastructure:

Domains, subdomains, SSL/TLS certificates, hosting, email identities, social and developer accounts, and code repositories used to stage or execute attacks.

Indicators of Pre-Attack (IoPAs):

Observable artifacts left during the setup of attack infrastructure (e.g., domain registration patterns, certificate issuance, repository creation) before any victim-facing malicious action.

Indicators of Compromise (IoC):

Artifacts associated with active or past compromise (e.g., malware hashes, C2 IPs, phishing URLs post-campaign).

Pre-attack prevention:

Detection, attribution, and disruption of attack infrastructure during the setup phase, prior to first contact with intended targets.

Mean Time to Preempt (MTTP):

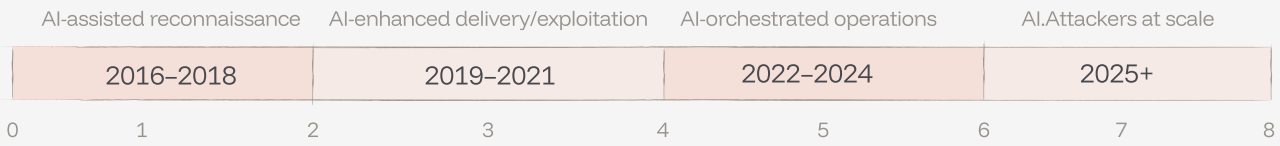
Time from first IoPA observation to disruption or mitigation; analogous to MTTD/MTTR but applied in the pre-contact phase.

Evolution: From Automation to Autonomous Attack

The transition from automated (scripted, human-directed) to autonomous (agentic, goal-directed with minimal human input) operations is consistent with the same trajectory seen in software development (e.g., co-pilots to full agentic workflows). Threat actors have adopted comparable tooling for resource development and attack execution.

Phase Model
(Aligned with MITRE and Kill-Chain Literature)

Period	Phase	Technical characterization
2016–2018	AI-assisted reconnaissance	AI used for OSINT, target profiling, vulnerability discovery; human remains operator.
2019–2021	AI-enhanced delivery/exploitation	Deepfakes, AI-concealment (e.g., DeepLocker-style); AI used in delivery and exploitation stages.
2022–2024	AI-orchestrated operations	Agentic frameworks used for end-to-end workflow: infra provisioning, exploit development, credential testing.
2025+	AI.Attackers at scale	Autonomous campaigns with strategic human oversight only; reconnaissance → exfiltration largely agent-driven.



Phase model - from automation to autonomous attack

The transition from automated (scripted, human-directed) to autonomous (agentic, goal-directed with minimal human input) operations is consistent with the same trajectory seen in software development (e.g., co-pilots to full agentic workflows). Threat actors have adopted comparable tooling for resource development and attack execution.

MITRE ATT&CK Mapping

TA0042 – Resource Development is the primary phase where pre-attack defense applies. Adversaries acquire and stage:

Domains and DNS, SSL/TLS certificates, Servers and hosting Email and social/ developer identities, Code repositories (e.g., for malware, tooling, or lures)

During this phase, adversaries leave IoPAs : registration metadata, certificate transparency logs, repository creation patterns, and linkability between assets.

These are observable before any IoC (e.g., malicious payload, active C2) appears.

Case 1: AI-Orchestrated Cyber Espionage (Anthropic GTG-1002)

Campaign Summary

In November 2025, Anthropic published analysis of a large-scale campaign they assessed as AI-orchestrated cyber espionage, with high confidence attribution to a Chinese state-sponsored actor.

This serves as the first publicly detailed case of an AI agent conducting the bulk of an espionage campaign.

Target set

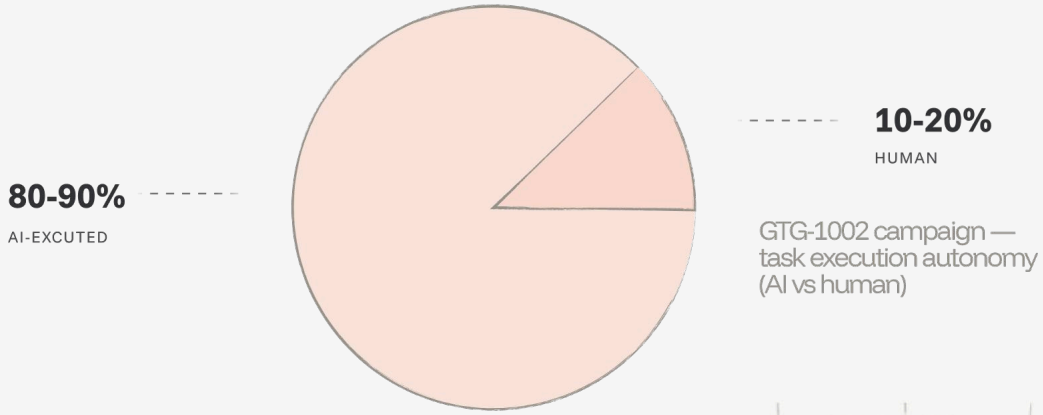
30 organizations across financial services, technology, chemical manufacturing, and governmen

Autonomy

~80–90% of tasks executed by the AI agent; human involvement at a small number of strategic decision points (on the order of 4–6).

Phase automated

Reconnaissance, vulnerability scanning, exploit development, credential harvesting, lateral movement, data exfiltration.



GTG-1002 campaign — task execution autonomy (AI vs human)

Guardrail Bypass and Safety Evasion

The campaign demonstrated deliberate evasion of LLM safety and usage policies:

Role-playing / context injection: .

Presenting the agent with a "legitimate security researcher / authorized penetration test" context to justify otherwise prohibited actions.

Task decomposition:

Breaking prohibited objectives into many small, individually innocuous-looking subtasks that do not trigger content filters.

Jailbreaking / prompt engineering:

Exploiting gaps in safety training to obtain compliant behavior for harmful workflows.

From a research perspective, this confirms that offensive use of current-generation LLMs is operationally viable when combined with prompt design and process control, and that safety controls are not a sufficient barrier to determined actors.

Implications for Detection and Prevention

Temporal compression:

End-to-end attack timelines can shrink to hours rather than weeks, reducing the utility of Mean Time to Detect (MTTD) and Mean Time to Respond (MTTR) if the attack completes before detection.

Latency:

By the time traditional IoCs (e.g., malware, C2, exfil) are generated, the operation may already be in motion or complete. Defense must shift earlier in the timeline.

Pre-attack as the viable window:

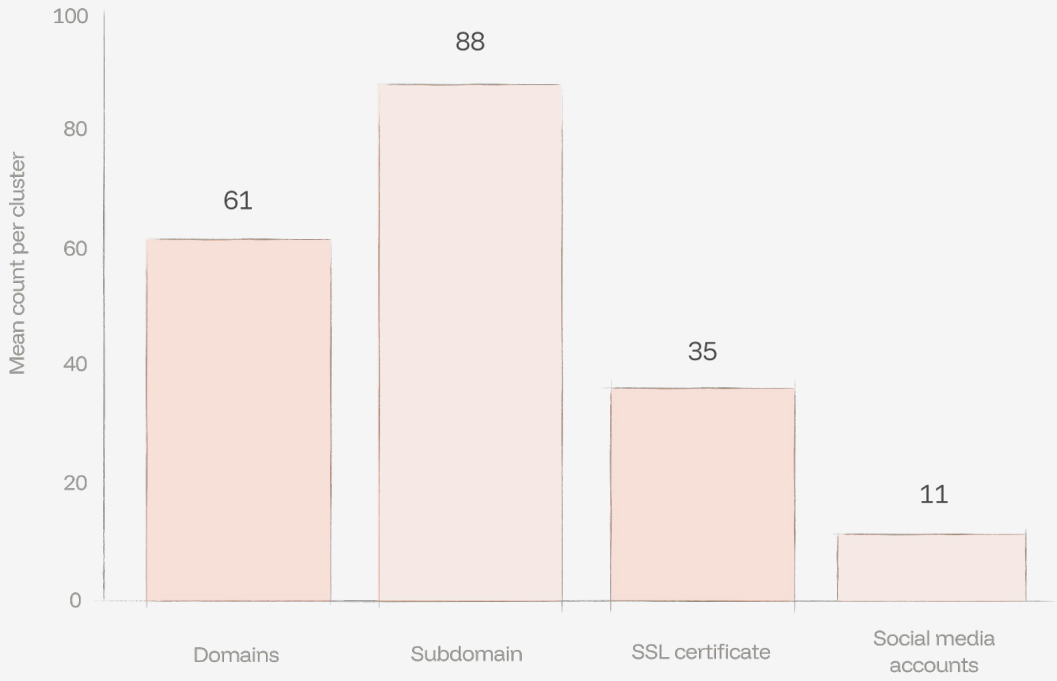
The resource development phase (infrastructure setup) remains the phase where defenders can observe and act on IoPAs before first contact. Pre-attack prevention is therefore not optional but necessary for this threat model.

Case 2: Industrial-Scale Attack Infrastructure (Malanta Cluster Analysis)

Dataset and method

Malanta Intelligence maintains and analyzes attack infrastructure clusters : sets of linked domains, certificates, IPs, social accounts, and code repositories used for phishing, malware distribution, fake software, fraud, and related operations. The analysis below is based on 33 distinct clusters , with composition and detection metrics derived from Malanta's datasets.

Asset type	Mean count per cluster
Domains	61
Subdomains	88
SSL certificates	35
Social media accounts	11



Average attack cluster composition (33 clusters)

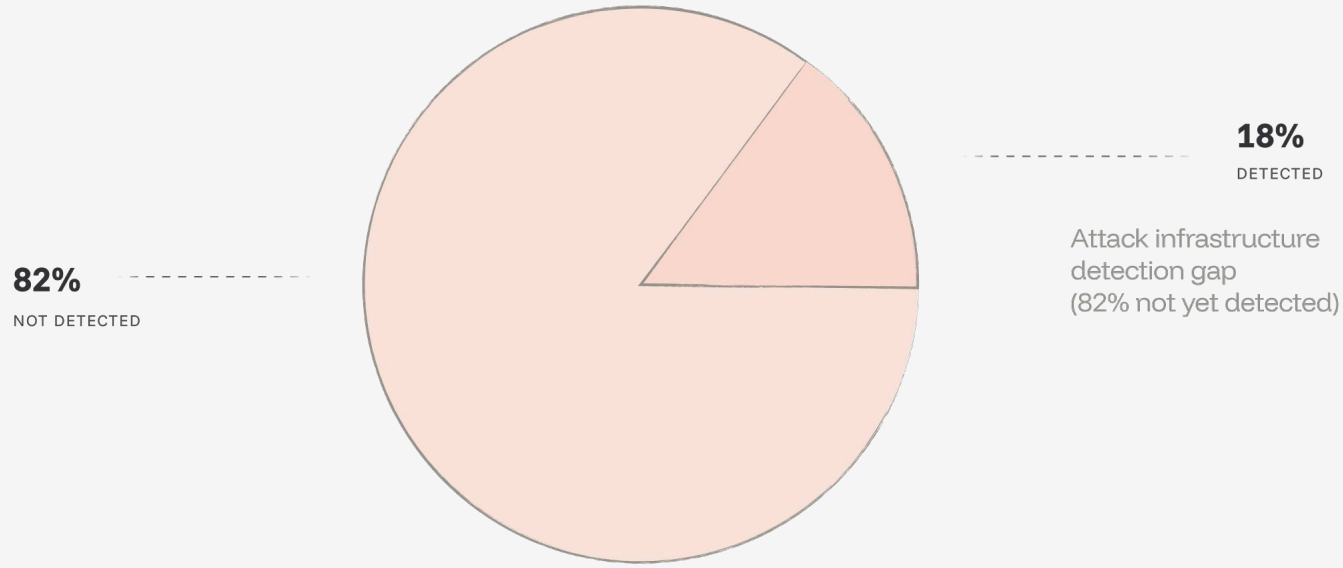
(Additional asset types, e.g., IPs, code repos, may be represented in the full cluster data.)

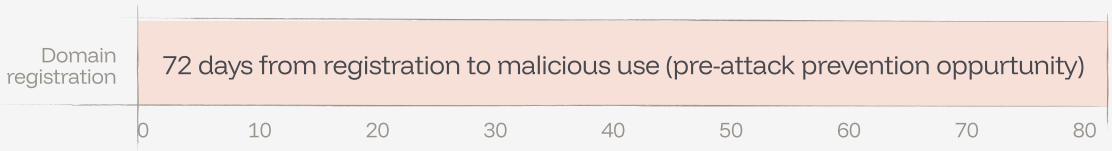
Detection Gap: IoPA vs IoC

~82% of domains in the analyzed clusters had not triggered any security vendor detection at the time of measurement. Infrastructure is therefore staged and ready before it is used in a way that generates classic IoCs.

Setup window:

The average time from domain registration to overtly malicious use is 72 days . This defines a concrete window in which IoPA-based detection and pre- attack disruption can operate.





Average setup window -

72 days from registration to malicious use

Infrastructure Economics

Mean estimated cost per cluster (infrastructure and staging):

~$9,223. At this scale, a threat actor can maintain 10+ active clusters for under ~$100,000, enabling parallel campaigns and rapid iteration.

This supports both nation-state and economically motivated actors.

Attack Categories

Clusters were associated with categories such as phishing operations, malware distribution, fake software campaigns, gambling/fraud.

The same resource development pattern (domains, certs, accounts, repos) supports multiple attack types; IoPA-based analysis is therefore relevant across categories.

Case 3: Exponential Growth in Malicious Resources (GenAI Correlation)

GitHub Malicious Repositories

2022

6,498 malicious repositories (Malanta tracking).

2024

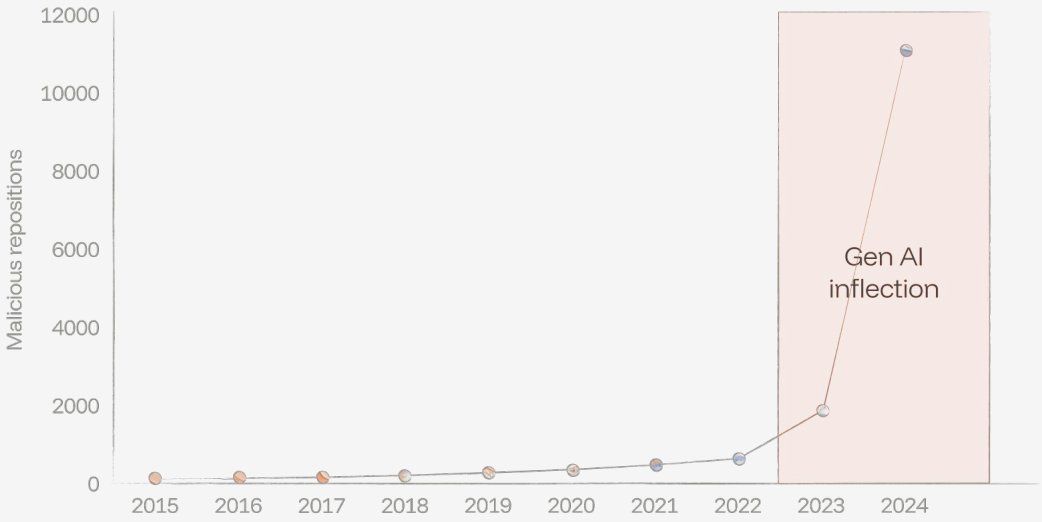
~110,000 malicious repositories.

APPROXIMATE GROWTH:

~1,592% over two years.

Year-over-year growth shows a clear inflection in 2023–2024, coinciding with broad availability of GenAI and public LLMs.

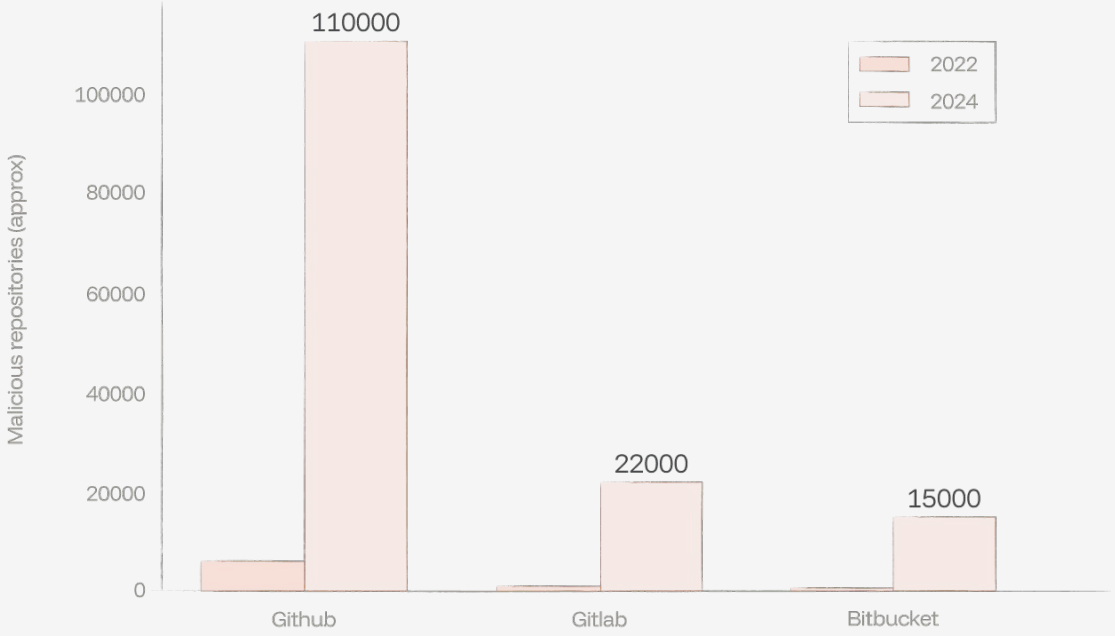
Pre–GenAI growth (e.g., 2015–2022) was on the order of ~32% annually; post–GenAI (2023–2024) annual growth is on the order of 285–340% , i.e., roughly a 10× acceleration in the rate of malicious resource creation.



GitHub malicious repositories - GenAI inflection (2023–2024)

Cross-Platform Consistency

The same acceleration pattern appears across GitHub, GitLab, and Bitbucket (similar 285–340% growth in 2023–2024).
The synchronicity and magnitude support the hypothesis of systematic adoption of AI-assisted resource generation (e.g., repo creation, boilerplate, tooling) rather than platform-specific or random variation.



Cross-platform malicious repository growth

(2022 vs 2024)

Interpetation

Democratization:

Industrial-scale attack infrastructure is no longer limited to well- resourced nation-state teams. Lower-skill actors can produce large volumes of malicious or dual-use assets via GenAI-assisted workflows.

Commoditization:

The “tooling” for building attack infrastructure has become cheaper and more accessible; defense must assume higher volume and diversity of threats.

Synthesis: Why Pre-Attack Prevention Is Technically Necessary

IoC-based defense is structurally late

By definition, IoCs appear during or after malicious actions. For AI-orchestrated campaigns with compressed timelines, detection and response may lag the completion of the attack.

The 72-day window is measurable

The empirical setup window (domain registration to overt malicious use) provides a defined interval for MTTP and for evaluating pre-attack prevention effectiveness.

Resource development is the shared phase.

TA0042 (Resource Development) is where all campaigns build and stage infrastructure. IoPAs exist in this phase; IoCs appear later. Defenders who only use IoCs operate in a later phase of the same kill chain.

82% dark infrastructure implies most risk is pre-IoC

The majority of cluster- associated domains were not yet flagged by vendors; risk is therefore concentrated in the pre-contact phase that only IoPA-based systems can address at scale.

MTTP as a first-class metric

Mean Time to Preempt (time from IoPA observation to disruption) should be elevated alongside MTTD and MTTR for organizations and for underwriting.

Data Sources and Methodology

Anthropic:

Threat intelligence report and full report on AI-orchestrated cyber espionage (November 2025); campaign GTG-1002.

Academic:

.Guembe et al. (2022), "The Emerging Threat of AI-driven Cyber Attacks: A Review," Applied Artificial Intelligence , for kill-chain mapping and AI-driven attack taxonomy.

Malanta Intelligence:

Malicious repository counts and projections (2015–2028); attack infrastructure cluster analysis (33 clusters); cluster composition and detection metrics;

Statistics cited (cluster means, growth rates, detection gap) are derived from the provided datasets. No fabricated or extrapolated data beyond these sources.

Conclusion

AI-orchestrated cyber operations have transitioned from theory to operational use ,as evidenced by the Anthropic GTG-1002 campaign and the industrialization of attack infrastructure observed in Malanta's cluster and repository data.

The confluence of autonomous agentic workflows and GenAI-accelerated resource creation compresses attack timelines and expands the pool of actors capable of mounting large-scale campaigns.

Defenses that rely solely on Indicators of Compromise operate in a phase that is increasingly too late. Pre-attack prevention grounded in Indicators of Pre-Attack and the disruption of staged infrastructure is the control that aligns with the same phase (MITRE TA0042, Resource Development) in which modern adversaries operate. Incorporating pre-attack capability and MTTP into security practice and cyber insurance underwriting is a technically justified response to the current threat landscape.

For questions regarding this report or underlying data

HOWDEN Malanta:

About Howden

Howden is a leading global insurance intermediary group with employee ownership at its heart.

Founded in 1994, it provides insurance, reinsurance and underwriting services and solutions to clients ranging from private individuals to the largest multinational companies.

The Group operates in 57 countries in Europe, the USA, Africa, Asia, the Middle East, Latin America, Australia and New Zealand, employs over 24,000 people and manages premiums totalling over $50 billion on behalf of its clients.

About Malanta

Malanta delivers the first Pre-Attack Prevention Platform. It stops breaches before they happen by preventing attacks, protecting critical assets, and enriching intelligence to neutralize threats before launch or execution.

For more information, please visit:

malanta.ai

Further information can be found at:

howdengroup.com

howdengroupholdings.com