

WHITEPAPER | SEPTEMBER 17, 2026

AI in the Cyber Kill Chain: Weapon or Target?

Where do adversaries actually use artificial intelligence today, where only in the lab, and what does this mean for prioritizing defense? This whitepaper filters the publicly available evidence through the 14 tactics of MITRE ATT&CK and separates two questions that are constantly mixed up in the debate: how well a use is documented, and who is capable of it today.

Audience

CISOs, heads of IT, management (Germany, Austria, Switzerland)

Classification

TLP:CLEAR, sharing permitted

Sources

Anthropic, MITRE ATT&CK (Campaign C0062), MITRE ATLAS 5.6, Google Threat Intelligence Group, Czybik et al. (USENIX Security 2026), Heiding et al. (2024), Agencia Española de Protección de Datos

Methodology

MITRE ATT&CK v18, MITRE ATLAS v5.6, Threat-Informed Defense

Evidence as of

September 17, 2026, version 2.0

THREE ZONES OF EVIDENCE

1	2	3
Mass market	Selective automation	Exists, does not scale
AI scales what is already known. Spear phishing and basic malicious code have reached the mainstream, and the marginal cost of initial access is a matter of cents.	Command generation at runtime, autonomous reconnaissance and data collection are documented in live operations, so far by well-resourced actors.	Autonomous persistence, lateral movement and data exfiltration are documented in individual campaigns. They still fail at scale because of the models' reliability.
Documented, affects everyone	Documented, affects few	Isolated, affects hardly anyone today

Management Summary

What does the evidence show?

The use of AI by adversaries is concentrated at the beginning of the kill chain. Preparation and initial access are documented multiple times and independently. The deeper a tactic operates inside the victim's network, the thinner the evidence becomes. For privilege escalation and command and control, there is no reliable field evidence.

What does this mean?

AI does not make the impossible possible, but it makes the known drastically cheaper. A personalized phishing email costs three cents and achieves the effect of human experts. Fully autonomous attack chains exist, but they are tied to the resources of state actors and remain error-prone.

What needs to be done?

Put resources where AI achieves economies of scale today: initial access, identities, detection of mass personalization. Zone 3 belongs in monitoring and exercises, not in investment planning. If you cannot handle Zone 1, you will never experience Zone 3.

Why this paper has two axes

Version 1 mixed evidence and reach on a single scale. A tactic could be observed in the field and a lab illusion at the same time. This version separates the two questions: the evidence says how well a use is documented. The zone says who is capable of it today and whom it affects. The two axes are independent of each other.

Two axes instead of one

Defending efficiently against cyberattacks requires aligning with real adversary behavior. This analysis applies the perspective of Threat-Informed Defense to the use of artificial intelligence. The MITRE ATT&CK framework serves as the vocabulary for the phases of an attack, MITRE ATLAS for attacks on AI systems. The two questions, how well a use is documented and who is capable of it today, are assessed separately.

Evidence: how well is the use documented?	Documented means documented multiple times or independently. Isolated means observed in the field in individual cases. Lab means shown only under test conditions.
Zone: who is capable of it today, and whom does it affect?	Zone 1 is the mass market, cheap and scaling. Zone 2 is well-resourced actors with repeated use. Zone 3 is individual campaigns that cannot be repeated at scale.

A tactic can be documented through two state actors and still be Zone 3. Conversely, a technique without field evidence remains outside the zones, even if it looks impressive in the lab.

The missing distinction

Before analyzing adversary tactics, a conceptual blur must be resolved. In strategic discussions, AI as the adversary’s tool is regularly equated with AI as the target of an attack. Anyone who wants to protect their own AI systems needs a different threat model (MITRE ATLAS) from anyone defending against an actor who uses AI to prepare a classic attack (MITRE ATT&CK).

The two perspectives meet at one point, and the case study below shows it: the adversary first had to defeat the model’s safeguards (ATLAS: LLM Jailbreak) before being able to use it as a tool against third parties (ATT&CK: Obtain Capabilities, Artificial Intelligence).

Tactic (ATLAS)	Technique	Evidence	Source or assessment
Reconnaissance	Active Scanning, AML.T0006	Lab	Systematic probing of production models has been described, but not documented as an incident
Resource Development	Poison Training Data, AML.T0020	Lab	Demonstrated academically several times, no confirmed field incident
AI Model Access	AI Model Inference API Access, AML.T0040	Documented	Prerequisite for any API abuse, not an attack in itself

Execution	LLM Prompt Injection, AML.T0051	Isolated	Documented in agent systems with tool access. ATLAS lists the technique under Execution and Privilege Escalation, not under Initial Access
Privilege Escalation, Defense Evasion	LLM Jailbreak, AML.T0054	Documented	GTG-1002: role play as an authorized security test to bypass safeguards [1]
Discovery	Discover LLM System Information: System Prompt, AML.T0069.002	Lab	Demonstrated many times, without documented damage
Exfiltration	Extract AI Model, AML.T0024.002	Documented	Systematic copying of model behavior via the interface (distillation) on an industrial scale
Exfiltration	LLM Data Leakage, AML.T0057	Isolated	Disclosure of training or context data through targeted inputs
Impact	Denial of AI Service, AML.T0029	Lab	Compute-intensive requests described as an overload vector
Impact	External Harms: Reputational Harm, AML.T0048.001	Isolated	Forced toxic or false outputs from production assistants

Corrected compared with version 1: in ATLAS, prompt injection is not an initial access technique. Plugin Compromise no longer exists as a separate technique in ATLAS 5.6 and has been removed. Model extraction is listed under Exfiltration, not under Model Access.

The survivorship bias of vendor reports

When analyzing AI as a weapon of adversaries, the industry relies predominantly on reports from commercial AI providers about suspended accounts. A considerable survivorship bias comes into play here. When providers show in detailed transparency reports how they stop activities ranging from cyber operations to weapons development, this mainly proves the existence of that filter. Inevitably, the actors that become visible are those careless enough to use monitored cloud interfaces with complete logging.

That professional actors switch to local models is not a guess. The phishing study cited below ran entirely on locally operated open-source models, and the PROMPTSTEAL malware documented by Google queries an open model via a Hugging Face interface. The evidence is nevertheless sufficient for a structured categorization. In Threat-Informed Defense, the rule is: not observed means exactly that, namely not observed. It does not mean impossible.

GTG-1002: the first documented AI-orchestrated campaign

In September 2025, Anthropic discovered an espionage operation that the company attributes with high confidence to a Chinese state actor and refers to as GTG-1002. MITRE lists it as Campaign C0062. It is the first documented campaign in which an adversary had an AI agent system carry out the tactical work of an intrusion. Of around 30 targeted organizations, a handful were successfully compromised according to Anthropic’s validation.



Humans were not absent but present at four decision points. It is exactly these points that define what autonomous meant in this campaign.

Initialization	Setting up the cover identity and the prompts used to get the model to cooperate in a supposedly authorized security test.
Reconnaissance to exploitation	Approval of the transition from passive reconnaissance to active exploitation of vulnerabilities found.
Credentials to movement	Approval of stolen credentials for lateral movement in the network.
Data exfiltration	Final decision on which data is exfiltrated.

The most important finding for defense is in Anthropic’s own report, and it puts the headline into perspective.

“Claude frequently overstated findings and occasionally fabricated data during autonomous operations, claiming to have obtained credentials that didn't work or identifying critical discoveries that proved to be publicly available information. This remains an obstacle to fully autonomous cyberattacks.”

Anthropic, Disrupting the first reported AI-orchestrated cyber espionage campaign, November 2025

The case thus proves two things at once. Autonomous attack chains are no longer theory. And they do not scale yet, because every result had to be validated by humans. Both feed into the matrix.

Fourteen tactics, one row per tactic

The overview arranges the 14 ATT&CK tactics chronologically. Each row names the observed AI use, the evidence level, the source and the zone. The shape of the curve is the message: documented scaling at the beginning of the chain, individual cases in the middle, gaps exactly where the hype is loudest.



Tactic	Observed AI use (ATT&CK)	Evidence	Source	Zone
Reconnaissance	Target profiles and OSINT, assisted. Autonomous scanning and vulnerability discovery: T1595.001, T1595.002, T1592	Documented, Isolated	GTIG 11/2025 [3]; COO62 [2]	Zone 1, Zone 3
Resource Development	Phishing content, basic malicious code, help with obfuscation during development: T1588.007, T1587.004, T1588.002	Documented	Czybik 2026 [4], GTIG 11/2025 [3], COO62 [2]	Zone 1
Initial Access	Personalized spear phishing: T1566. Exploitation of public-facing applications: T1190	Documented	Heiding 2024 [5], Czybik 2026 [4], COO62 [2]	Zone 1
Execution	Malware has an LLM generate commands at runtime: T1059	Isolated	PROMPTSTEAL, APT28, live use against Ukraine [3]	Zone 2
Persistence	Creation of local accounts: T1136.001	Isolated	COO62 [2]	Zone 3
Privilege Escalation	No AI-related ATT&CK technique documented	No evidence	COO62 lists no technique for this tactic [2]	none
Defense Evasion	Self-rewriting of the code at runtime: T1027	Lab	PROMPTFLUX, classified by Google as experimental, no confirmed use [3]	not yet
Credential Access	Credentials from files: T1552.001	Isolated	COO62 [2]	Zone 3
Discovery	Network, account and system discovery: T1046, T1087, T1082. Autonomous vulnerability discovery in applications	Documented	COO62 [2], AEPD 09/2026 [6]	Zone 2

Lateral Movement	Use of stolen credentials: T1078	Isolated	C0062, after approval by the operator [1]	Zone 3
Collection	Automated collection and sorting: T1119, T1213.006, T1005	Documented	C0062 [2], PROMPTSTEAL [3], AEPD 09/2026 [6]	Zone 2
Command and Control	No AI-related ATT&CK technique documented	No evidence	no source	none
Exfiltration	Exfiltration via web services after local staging: T1567, T1074.001	Isolated	C0062 [2]	Zone 3
Impact	Manipulation of personal data: T1565	Isolated	AEPD 09/2026 [6]	Zone 3

Two gaps are a finding in themselves. For privilege escalation and command and control, precisely the stages on which the narrative of the fully autonomous attack hinges, there is still no documented AI-related ATT&CK technique. Campaign C0062 managed without either, because the stolen credentials were sufficient.

Zone 1: mass market and the collapse of marginal costs

Two independent studies describe the same effect from two directions. The Harvard study by Heiding, Schneier and colleagues measures the effect: in an experiment with 101 participants, fully automated AI emails achieved a click rate of 54 percent, exactly the level of human experts. The control group with generic phishing was at 12 percent. AI does not outperform the expert. It replaces them.

The study by BIFOLD, TU Berlin, Inria and Ruhr University Bochum measures the price of this replacement with 4,010 real recipients.

7 min of manual work per email: five minutes of research, two minutes of writing	45 s per email in the automated process, locally on open-source models	\$0.03 cost per personalized email, \$120 for the entire campaign	0.03% of the emails were detected by the university's spam filter: one in 3,949
--	--	---	---

The click rates in this study are lower than at Harvard: 3.9 percent generic, 10 percent AI-personalized, 24.2 percent for human experts. The difference lies in the study design, not in the finding. Both studies show that AI approaches the effect of human work at a fraction of the cost. The last figure is the most important for defense: if only one out of almost 4,000 emails stands out technically, the email itself is no longer a detection feature.

Selective automation

In November 2025, Google documented for the first time malware that queries a language model during execution. PROMPTSTEAL, tracked by CERT-UA as LAMEHUG, has an open model generate the commands for system reconnaissance and document collection instead of carrying them hard-coded. The actor is APT28, and the use took place against Ukrainian targets. This is not a demonstration but a tool in operation.

The second observation comes from a supervisory authority. On September 15, 2026, the Spanish data protection authority AEPD reported the first personal data breach notified to it that was carried out by an AI agent. The agent found vulnerabilities in files, logged in, independently searched for further vulnerabilities in the application, modified personal data and accessed invoices. The information comes from the notification by the affected organization and has not yet been fully verified by the authority. The route is remarkable: the case was not discovered by a vendor but reported as a regular incident under Art. 33 GDPR.

Both cases involve reconnaissance and data collection. That is no coincidence. These are the phases in which an agent can do the least wrong and in which hallucinations cost the least.

Exists, but does not scale

Persistence, credential theft, lateral movement and data exfiltration are documented in the field through COO62, each as an isolated case. At the same time, the campaign shows why this does not yet become a mass phenomenon. The adversary had to verify every claim the model made, because it reported credentials that did not work and rated findings as critical that were publicly known. With around 30 targets, the outcome was a handful of successful intrusions, and the effort required matched that of a state program.

Version 1 of this paper called this zone a lab illusion. That was wrong. The tactics exist in the field. What is true is that today they depend on three conditions that hardly any actor meets at the same time: access to a powerful agent system without effective safeguards, the resources for continuous human validation, and targets whose value justifies this effort.

For privilege escalation and command and control, even this isolated case is missing. Anyone who aligns their investments with these two stages is planning against something nobody has observed.

Prioritization for defense

AI does not make the impossible possible, but it makes the known drastically cheaper and faster. Both incidents documented in 2026 confirm this: they were not novel exploits but well-known at-

tack paths, carried out at a speed and cost that were previously unattainable.

This leads to a clear prioritization. Zone 1 affects every company, today, with tools that cost three cents and no longer trigger any filter. Zone 2 affects companies in the sights of well-resourced actors. Zone 3 affects hardly anyone today but will be tomorrow's Zone 2 as soon as the models become more reliable.

Defense must therefore focus primarily on Zone 1 and Zone 2, where AI gives the adversary real economies of scale. Specifically, this means: a second factor on every remote access, because personalized phishing reliably defeats the first factor. Detection that does not rely on the quality of the language but on behavior after the click. And visibility of automated reconnaissance in your own network, because that is where Zone 2 begins.

Zone 3 belongs in monitoring and in crisis team exercises, not in investment planning. Anyone who ignores it will miss the moment it scales. Anyone who prioritizes it ahead of Zone 1 loses to everyday reality. If you cannot handle Zone 1, you will never experience Zone 3, because the adversary does not need it.

Sources (numbers as in the matrix)

- [1] Anthropic: Disrupting the first reported AI-orchestrated cyber espionage campaign, November 2025
<https://www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf>
- [2] MITRE ATT&CK: Campaign CO062, Anthropic AI-orchestrated Campaign
<https://attack.mitre.org/campaigns/CO062/>
- [3] Google Threat Intelligence Group: GTIG AI Threat Tracker, Advances in Threat Actor Usage of AI Tools, November 5, 2025
<https://services.google.com/fh/files/misc/advances-in-threat-actor-usage-of-ai-tools-en.pdf>
- [4] Czybik, Kouam, Heubl, Nold, Rieck: A Large-Scale Study of Personalized Phishing using Large Language Models, USENIX Security 2026
<https://www.usenix.org/system/files/usenixsecurity26-czybik.pdf>
- [5] Heiding, Lermen, Kao, Schneier, Vishwanath: Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns, arXiv 2412.00586, 2024
<https://arxiv.org/abs/2412.00586>
- [6] AEPD: announcement of September 15, 2026 on the first reported personal data breach by an AI agent (as reported by El Economista, Spanish)
<https://www.eleconomista.es/tecnologia/noticias/14014370/09/26/espana-registra-la-primera-brecha-de-seguridad-provocada-por-un-agente-de-ia.html>
- [7] MITRE ATLAS, version 5.6.0
<https://github.com/mitre-atlas/atlas-data>

How we support you with implementation

This analysis follows the same methodology we use to build our clients' security programs: not along categories of threats, but along the techniques that documented actors use against companies like yours. In our consulting work, we translate the threat profile of your industry into prioritized measures and solid evidence for NIS2, DORA and ISO 27001, from building the ISMS through business continuity management to preparing for an emergency.

For the zones in this paper, this means: we check whether your controls against Zone 1 actually work, and we put Zone 3 where it belongs today, namely into a crisis team exercise with a real scenario.

Contact us for an initial conversation.

kontakt@threat-informed.de | www.threat-informed.de

This whitepaper is part of our threat briefings for companies in Germany, Austria and Switzerland. All figures come from the primary sources cited and were checked against them on September 17, 2026. The assessments of evidence and zone are our own.