COURSE S1-ISI5 – Information Systems Security

Artificial Intelligence and Cybersecurity

Anatomy of an Augmented Attack – AI as Weapon and Shield

Instructor: Fabrice Pizzi
SORBONNE UNIVERSITY
MASTER 2 SI-ISI5
Academic Year 2025-2026

1. Introduction – AI, the Fourth Revolution

Artificial intelligence constitutes the fourth technological revolution and profoundly reshapes the cybersecurity landscape. While generative AI (GenAI) offers major opportunities for defenders, it also represents a considerable force multiplier for attackers.

This document explores this duality: how AI transforms every stage of the Kill Chain on the offensive side, and how it strengthens detection and response capabilities on the defensive side. The cybercriminal ecosystem is now marked by increasing professionalization and specialization, where attack tools are created faster than defenses can be updated.

ANSSI and ENISA confirm this trend: AI accelerates both the sophistication of attacks and the response capabilities of defenders. We are witnessing a true digital arms race.

2. Offensive AI – The Cyber Attacker's Weapon

AI has become a powerful tool for attackers, enabling automation, enhancement, and acceleration of every phase of the cyberattack. The RaaS (Ransomware-as-a-Service) model combined with AI significantly lowers the technical barrier to entry.

2.1 AI-Augmented Reconnaissance

AI transforms the reconnaissance phase by enabling automated semantic analysis of massive volumes of public data to identify high-value targets and their vulnerabilities.

- **Semantic OSINT Analysis:** AI analyzes public data (LinkedIn, corporate reports, GitHub commits) to identify high-value targets and their social or technical vulnerabilities.
- **Automated Mapping:** Automatic discovery of attack surfaces via tools like Shodan, enriched by AI to identify misconfigured cloud services and exposed APIs.
- **Target Profiling:** Automatic generation of detailed profiles of key employees to prepare ultra-targeted social engineering attacks.

2.2 Weaponization through Generative AI (GenAI)

Generative AI revolutionizes the creation of cyber weapons by enabling automated generation of polymorphic malicious code, dynamically adapted to each target.

- **Polymorphic Malware Generation:** Jailbroken LLMs create polymorphic malware or multi-platform exploitation scripts (Bash, PowerShell) bypassing EDR signatures. AI leverages tools already present on the machine (Living-off-the-Land).
- **Multi-Platform Code:** Rapid production of malicious code targeting Windows, Linux, and macOS simultaneously, making traditional signature-based detection obsolete.
- **Lowered Barrier to Entry:** Even RaaS affiliates with limited technical skills can now generate sophisticated offensive tools.

2.3 The Era of Hyper-Personalized Spear-Phishing

Phishing remains the dominant vector (60% of cases according to ENISA), but AI makes it incomparably more dangerous by eliminating traditional detection markers.

- **GenAI for Text:** Creation of perfectly styled, error-free spear-phishing emails, contextualized to the target. Messages are indistinguishable from legitimate communications.
- **Audio and Video Deepfakes:** Impersonation of an executive via video or voice call (vishing). A deepfaked CEO can request an urgent wire transfer during a video conference.
- **Indirect Prompt Injection:** A malicious payload is hidden in a web page that the victim's enterprise LLM will consult, potentially leading to data leakage.

2.4 Automated Exploitation – The Hunt for Zero-Days

AI radically compresses the zero-day vulnerability lifecycle, from several months to just hours between discovery and operational exploitation.

- **Intelligent Fuzzing:** AI agents analyze normal application behaviors and generate optimized inputs to discover vulnerabilities (buffer overflows, use-after-free).
- **Large-Scale Code Analysis:** AI can scan millions of lines of open-source code to find undocumented vulnerabilities and risky coding patterns.
- **Compressed Zero-Day Cycle:** This is the industrialization of rapid exploitation mentioned by ANSSI.

2.5 Autonomous AI Agents and Attacks Against Models

Lateral movement and C2 phases are automated by autonomous AI agents. At the same time, AI models deployed by organizations themselves become targets through data poisoning (corrupting training data), model poisoning (backdoors triggered by specific inputs), and indirect prompt injection (payloads hidden in data sources).

3. Attacks ON AI Systems – Taxonomy

The previous section explored AI as an offensive weapon (attacks WITH AI). This section addresses a critical complementary dimension: attacks that directly target AI systems themselves (attacks ON AI). This analysis draws on the work of the AI Security Working Group of Campus Cyber and Hub France IA (February 2026), which proposes a taxonomy structured at three levels: lifecycle phases, attack families, and specific attacks.

The referenced frameworks are NIST AI 100-2 (2023), MITRE ATLAS, OWASP Top 10 LLM and Top 10 ML, as well as ANSSI recommendations on AI systems security. The taxonomy identifies three major attack categories: poisoning, manipulation (evasion, prompt injection), and oracle attacks (inference, extraction).

3.1 Poisoning Attacks

Poisoning attacks target the model's learning phase by altering training data or the model itself. They compromise system integrity from its construction, making their effects particularly difficult to detect once the model is deployed.

Training Data Corruption (Data Poisoning)

The attacker introduces malicious data into the training set to alter the model's behavior. Two main variants exist:

Attribute/Label Corruption: subtle modification of training data attributes or labels to degrade prediction quality. For example, inverting "spam" and "legitimate" labels in 2% of an anti-spam filter's data is sufficient to create exploitable false negatives.

Backdoor Poisoning: the attacker injects a discrete pattern (trigger) into certain training data, associated with a target label. The model learns to associate this trigger with the malicious output. In production, the model operates normally except when the trigger is present, then activating the malicious behavior. This is the mechanism illustrated by PoisonGPT (section 6.5).

Analogy: corrupting a textbook so that students learn wrong answers – but only for specific questions.

Model Poisoning

Beyond data, the attacker can directly modify model parameters during training, particularly in distributed or collaborative learning contexts. This parameter corruption alters neural network weights without standard performance metrics necessarily detecting the degradation.

AI Supply Chain Attacks

Compromise occurs even before the model is used: compromised software libraries (malicious PyPI package), pre-trained models containing backdoors (typosquatting on Hugging Face, as demonstrated by PoisonGPT), or altered development frameworks. The Operation OpenClaw case (section 6.6) illustrates this threat with booby-trapped community skills on ClawHub. The malicious code attack consists of manipulating pre-processing or post-processing logic integrated into the model pipeline, allowing behavior alteration without touching the model weights.

3.2 Manipulation and Evasion Attacks

These attacks occur after the learning phase. The attacker injects modified input data to obtain an output different from normally expected, without altering the model itself.

Adversarial Examples

Adversarial attacks consist of applying calculated perturbations to input data, often imperceptible to humans, but sufficient to fool the model. The foundational example is modifying a road sign with strategically placed stickers, fooling an autonomous vehicle's visual recognition system. Two variants exist: targeted attacks (the attacker wants a specific output) and untargeted attacks (any error suffices). The challenge for the attacker is calibrating a perturbation weak enough to avoid detection but strong enough to impact the output.

MITRE ATLAS Reference: AML.T0015 – Evade ML Model.

Prompt Injection – LLM Jailbreak

LLM Jailbreak is a specific case of prompt injection where the goal is to disable the model's guardrails to make it produce prohibited content: malicious code generation, ethical filter bypassing, or system instruction extraction. Techniques include role-playing (DAN – "Do Anything Now"), progressive multi-turn, and encoding obfuscation (Base64, multiple languages). The course already mentions jailbroken LLMs (WormGPT, section 6.2) – this section describes the underlying technical mechanism.

Model Substitution

The attacker replaces a legitimate model with a malicious one in the deployment environment. This can occur through model registry compromise, CI/CD attack, or exploitation of misconfigured permissions on artifacts. This attack is distinct from poisoning: the original model is not corrupted, it is entirely replaced.

Model Degradation in Production

An attacker can systematically submit data designed to progressively degrade model performance when the model incorporates a continuous relearning mechanism (online learning). The model gradually "unlearns" its legitimate capabilities. The Microsoft Tay chatbot case (2016), hijacked in less than 24 hours through malicious interactions, illustrates this risk.

3.3 Oracle Attacks – Extraction and Inference

These attacks exploit access to the model (via its API or predictions) to extract confidential information or reconstruct its components.

Model Stealing

The attacker reconstructs a functional copy of the target model by multiplying queries and analyzing responses. The "clone" model can then be used to prepare adversarial attacks, bypass detection systems, or violate intellectual property. LLM-specific variant: meta-prompt extraction allows recovering the system instructions (system prompt) that control model behavior, exposing the application's business logic and guardrails.

Model Inversion

The attacker exploits model predictions to reconstruct approximations of training data. For example, from a facial recognition model and a person's name, it is possible to reconstruct an approximate image of their face. This attack directly threatens the confidentiality of personal data used during training (GDPR violation).

Membership Inference

The attacker determines whether specific data was part of the training set. This attack relies on the fact that models behave differently on data seen during training (overfitting). GDPR

implication: proving that personal data was used to train a model without consent constitutes a direct violation of the regulation.

Prompt Injection – Data Extraction

This scenario is a specific form of prompt injection where the goal is to exfiltrate sensitive data the model has access to. The attacker constructs prompts so the LLM divulges confidential information: memorized training data (credit card numbers, emails), connected RAG database content, or system information. This is distinct from jailbreak, whose goal is to bypass behavioral restrictions.

Cost Exploitation

The attacker submits queries designed to maximize resource consumption (tokens, GPU) to financially ruin the victim. Extremely long prompts, forced reasoning loops, or abuse of agentic features (cascading tool calls) can generate massive bills. This attack is specific to usage-billed LLMs and agentic systems.

3.4 RAG Architecture-Specific Attacks

Retrieval-Augmented Generation (RAG) improves LLMs by giving them access to external knowledge bases. However, this architecture introduces new attack surfaces specific to each component of the chain.

Indirect Prompt Injection via RAG

The attacker poisons documents in the knowledge base used by the RAG. When the LLM retrieves a poisoned document, it executes the malicious instructions hidden in the content. The injection is “indirect” because the attacker does not communicate directly with the LLM – malicious content transits through the document retrieval system. This can lead to data exfiltration, misleading response generation, or triggering unauthorized actions.

This vector is used in Phase 4 of Operation OpenClaw (section 6.6) via poisoned Slack messages.

Embedding Model or Retrieval Attack

The attacker targets the embedding model (which converts text to vectors) or the similarity search mechanism. By manipulating embeddings, they can prioritize malicious documents or conversely hide critical information. This attack is more sophisticated than indirect injection as it targets the RAG infrastructure itself.

3.5 Attacks on Agentic Systems and MCP

Agentic systems (autonomous AI agents) and the MCP protocol (Model Context Protocol, an open standard designed by Anthropic) represent a paradigm shift in attack surfaces. Operation OpenClaw (section 6.6) illustrates these risks in an operational context. The OWASP Multi-Agentic System Threat Modeling Guide (v1.0, April 2025) formalizes these threats.

Privilege Compromise in Multi-Agent Systems (MAS)

An agent accessing external services can be manipulated to obtain elevated privileges. The attacker exploits configuration errors and privilege inheritance mechanisms between agents (implicit delegations) to escalate permissions. Inter-agent error propagation can transform an isolated vulnerability into systemic failure: one agent’s erroneous output becomes another’s erroneous input.

Human Supervision Overload

The attacker triggers a high number of requests requiring human validation, causing decision fatigue. The supervisor ends up automatically approving requests to limit delays, opening the door to dangerous actions. This attack exploits the weakness of “human-in-the-loop” when overloaded.

Malicious Agent Infiltration in a MAS

The attacker modifies an agent’s parameters or manipulates its reward system to divert it from its initial mission. Introduced into a multi-agent system, the defective agent spreads erroneous information to other agents, potentially leading the entire MAS to unauthorized actions. The modification is progressive and difficult to detect.

MCP Protocol Threats

MCP, by connecting agents to external tools and data sources, introduces specific vectors:

Tool Poisoning: attackers replace a legitimate tool with a malicious version by exploiting the already established trust relationship;

Metadata Injection: malicious code hidden in a tool’s description or parameters forces the AI system to execute unintended actions (exfiltration, calls to other malicious tools);

Unexpected Execution Chain: the use of multiple MCP servers can lead to uncontrolled cascading tool calls, causing operational incidents or denial of service.

3.6 Attacks on Federated Learning

Federated learning allows multiple participants to train a collaborative model without sharing their data. Although designed for confidentiality, this process open to many actors enables new attacks:

Integrity Attacks: the orchestrator has no way to verify that parameters transmitted by a participant correspond to legitimate learning. A malicious participant can poison the federated model by sending corrupted parameter updates. The “Byzantine attack” technique consists of sending gradients that steer the model toward malicious behavior while remaining close enough to the norm to evade robust aggregation mechanisms.

Confidentiality Attacks: exchanged parameters can leak information about local data. Gradient inference attacks allow reconstructing data samples from parameter updates, compromising the confidentiality that federated learning is supposed to guarantee.

3.7 Decommissioning Phase – Residual Risks

The lifecycle of an AI system does not end at decommissioning. Two major risks persist:

Data Persistence: residual models and data left on “decommissioned” systems can be recovered. Model weights, inference logs, and temporary training datasets represent sensitive data that often survives logical deletion.

Unauthorized Model Reuse: a decommissioned but undestroyed model can be recovered and reused outside its original context, without the associated guardrails and supervision. This is particularly critical for models trained on GDPR-regulated data, where the right to erasure applies.

3.8 Synthesis – Attack Mapping by Lifecycle Phase

The following summary synthesizes the identified attacks by ANSSI/OECD lifecycle phase, distinguishing their applicability to predictive and/or generative AI:

Phase 1 – Training: data corruption (attributes, labels, backdoors), model poisoning (corrupted parameters), AI supply chain (booby-trapped pre-trained models, malicious code in pipelines), malicious data replication.

Phase 2 – Deployment: model substitution, environment compromise, test metric manipulation, privilege compromise in MAS.

Phase 3 – Production: adversarial attacks (evasion), prompt injection (direct, indirect, jailbreak, data extraction), RAG attacks (indirect prompt injection, embedding attack), agentic attacks (MCP, MAS infiltration), oracle attacks (model extraction, inversion, membership inference), cost exploitation, model degradation.

Phase 4 – End of Life: residual data persistence, unauthorized model reuse.

Source: adapted from "Analysis of Attacks on AI Systems", AI Security Working Group, Campus Cyber / Hub France IA, February 2026. Frameworks: NIST AI 100-2, MITRE ATLAS, OWASP Top 10 LLM & ML, ANSSI.

4. Defensive AI – The Cyber Shield

Facing the rise of offensive AI, defenders also leverage artificial intelligence to detect, respond, and prevent threats more effectively.

4.1 EDR/XDR and Behavioral Analysis

- **Behavioral Detection:** AI models the normal behavior of each endpoint and detects anomalies in real-time (unusual PowerShell processes, memory injection, unauthorized lateral movements).
- **Multi-Source Correlation (XDR):** AI correlates alerts from multiple sources (endpoints, network, email, cloud) to reconstruct attack chains and reduce false positives.
- **Automated Response:** Automatic isolation of a compromised endpoint in milliseconds, without human intervention.

4.2 AI-Augmented SIEM/SOC

- **Automatic Triage:** AI classifies and prioritizes alerts, reducing alert fatigue for SOC analysts.
- **NLP Log Analysis:** Language models analyze unstructured logs and identify suspicious patterns.
- **Proactive Threat Hunting:** AI generates compromise hypotheses by analyzing historical behaviors and Threat Intelligence feeds.

4.3 Network Anomaly Detection and CTI

- **DNS/Netflow Analysis:** Detection of DNS tunneling for C2 communications, identification of unusual lateral connections, monitoring of exfiltrations via encrypted HTTPS channels.
- **Automated Threat Intelligence:** Automatic IoC enrichment, OSINT report synthesis, predictive models to anticipate attack campaigns.
- **Zero Trust + AI:** Adaptive authentication, dynamic micro-segmentation, continuous trust scoring for each user and device.

4.4 AI-Powered Autonomous Pentesting

Pentesting (penetration testing) is traditionally a manual, expensive, periodic (annual), and scope-limited operation. AI transforms this approach with continuous autonomous solutions.

Horizon3.ai – NodeZero

NodeZero is the reference platform for autonomous pentesting. Used by over 4,000 organizations (including the US DoD via the NSA CAPT program), it simulates a real attacker without agents or pre-configuration.

- **Operation:** NodeZero deploys in minutes via a Docker container (internal) or from the Horizon3 cloud (external). Without agents, it scans the network, discovers attack surfaces, chains exploits, and generates proofs of compromise (screenshots, hashes, data) – all autonomously.
- **Key Result – GOAD in 14 minutes:** In August 2025, NodeZero became the first AI to completely solve the Game of Active Directory (GOAD) by Orange Cyberdefense: 5 forests, 3 domains, 28 machines, 100+ users – compromised in 14 minutes, achieving Domain Admin on all domains autonomously.
- **NSA CAPT:** In the NSA's Continuous Autonomous Penetration Testing program, NodeZero covered 1,000 defense subcontractors, discovered 50,000 exploitable weaknesses, and identified over 2,000 credentials compromised by infostealers.
- **Find-Fix-Verify Loop:** After each test, NodeZero provides remediation recommendations prioritized by real impact. The team fixes, then reruns NodeZero to verify the corrections – a continuous validation cycle.

Pentera – AI Security Validation

Pentera is the other major player in autonomous pentesting, with over 996 clients worldwide.

- **Approach:** Pentera emulates real adversaries in production environments. Its AI engine generates adapted payloads, tests attack paths end-to-end, and validates whether defenses actually block real exploitation techniques.
- **Pentera Resolve:** AI-driven remediation module that consolidates security findings (infrastructure, applications, cloud), deduplicates alerts, and generates prioritized action plans with exploitability context.
- **Conversational Vision:** Pentera is developing "natural language pentesting": the CISO describes a scenario in free text ("test if stuffed credentials from leaked databases give access to the finance network") and the AI translates this into an automated attack sequence.

AI Pentest vs Traditional Pentest

Traditional pentesting costs €10,000 to €50,000 per test, takes 2 to 4 weeks, and is only performed once or twice a year. AI autonomous pentesting runs in minutes, covers the entire perimeter continuously, and costs a fraction of the price. It represents the shift from periodic auditing to continuous validation.

4.5 AI-Powered Compliance Analysis and Remediation

- **4.5 AI-Powered Compliance Analysis and Remediation**

- Beyond offensive pentesting, AI also transforms passive security posture analysis: continuous detection of misconfigurations, automated compliance verification, and AI-assisted remediation.
- **CSPM – Cloud Security Posture Management**
- CSPM tools continuously analyze cloud configurations (AWS, Azure, GCP). Market players (Wiz, Prisma Cloud, Gomboc.ai) focus on identifying misconfigurations and compliance gaps, with an increasingly AI-driven approach to automated remediation.
- However, these solutions have a structural limitation: they are blind to On-Premise infrastructure (Active Directory, Exchange), legacy databases, and OT networks. Organizations with hybrid environments therefore need complementary solutions.
- **Compliance-as-Code and Augmented GRC**
- A new generation of GRC tools (Drata, Vanta) uses AI to automate compliance. These platforms automatically collect evidence from IT systems, continuously assess compliance with multiple frameworks (SOC 2, ISO 27001, GDPR), and generate real-time compliance dashboards.
- Furthermore, these solutions are predominantly SaaS and US-hosted, which poses a sovereignty concern for European organizations subject to GDPR and NIS2.
- **Proactive vs Reactive Approach: Towards Convergence**
- AI-powered compliance analysis is fundamentally different from pentesting: one verifies configuration state (proactive approach), the other tests exploitability (reactive approach). The trend is toward convergence of both approaches into unified platforms.
- **4.5.4 Case Study: LIA-Scan – Towards a Sovereign Multi-Technology Approach**
- *Note: LIA-Scan is a research project by the author, currently in development phase. It is presented here as an illustration of an alternative approach to existing commercial solutions. The specifications mentioned are design objectives.*
- LIA-Scan starts from the observation that current solutions cover either the Cloud (CSPM) or administrative compliance (GRC), but rarely the entire scope of a hybrid organization. The project aims to combine automated technical auditing with AI-assisted remediation across a unified multi-technology scope.
- The research and development axes are as follows:
- **Deep Technical Audit:** Where GRC tools collect administrative evidence, LIA-Scan aims to perform actual configuration audits of on-premise technologies (Active Directory, DNS, DHCP, Exchange, SQL Server) and cloud environments, identifying not only misconfigurations but also their exploitable attack chains.
- **AI-Contextualized Remediation:** The project integrates a sovereign RAG (Retrieval-Augmented Generation) module based on a corpus of 24,000 technical documents (Microsoft, ANSSI, CIS Benchmarks, NIST). For each finding, the AI generates a contextualized remediation recommendation, adapted to the specific technology and environment.
- **Multi-Framework Cross-Mapping:** LIA-Scan targets native integration of 103 compliance frameworks (compared to approximately 35 for competing GRC solutions), with an automatic mapping engine that associates each technical finding with the corresponding requirements of applicable frameworks.

- **Threat Intelligence-Enriched Risk Scoring:** Risk scoring integrates CISA KEV and EPSS feeds to prioritize vulnerabilities not only by theoretical severity but also by their actual probability of exploitation in the wild.
- **Sovereign Deployment:** Unlike the cited solutions, LIA-Scan is designed for On-Premise or Air-Gapped deployment, meeting the requirements of critical, defense, or regulated organizations.
- In summary, LIA-Scan aims to merge the proactive coverage of a GRC tool with the technical depth of a vulnerability scanner, while maintaining digital sovereignty. This approach positions it as a complement to offensive pentesting solutions.

Criterion	CSPM (Wiz)	GRC (Vanta)	AI Pentest (Horizon3)	LIA-Scan
Technologies Covered	Pure Cloud (AWS, Azure, GCP)	SaaS / Cloud API	Internal / External Network	**148 tech (Cloud + On-Prem + OT + DB + Containers)**
Detection Rules	~2,800 (cloud)	Admin Evidence	Active Exploits	**10,962 rules**
Compliance Frameworks	~20 (CIS, SOC2)	~35	N/A	**103 frameworks**
Remediation	AI-Guided (cloud)	Comply AI (OpenAI)	Find-fix-verify	**10,321 guides + Sovereign RAG**
Sovereign Deployment	SaaS US	SaaS US	Docker / Cloud	**On-Premise / Air-Gapped**
Threat Intelligence	Cloud Vulnerabilities	No	Exploitability	**CISA KEV + EPSS (J+2)**

5. AI vs AI – The Arms Race

The confrontation between offensive AI and defensive AI creates a permanent escalation dynamic. The table below summarizes this confrontation at each stage of the Kill Chain.

Phase	Offensive AI	Defensive AI
Reconnaissance	Analyse sémantique OSINT, cartographie automatisée des surfaces d'attaque	Digital footprint reduction, scan detection by ML
Weaponization	GenAI for polymorphic malware, WormGPT/FraudGPT, jailbroken LLMs	Automated CTI monitoring, predictive threat analysis
Delivery	AI spear-phishing, audio/video deepfakes, prompt injection	NLP email filtering, deepfake detection, AI sandboxing
Exploitation	Intelligent fuzzing, zero-day cycle compressed to hours	Big Sleep (Google), AI-prioritized patch management
Installation	Autonomous LotL agents, adaptive EDR evasion	Behavioral EDR/XDR, AI whitelisting, Zero Trust
C2	Adaptive C2 via legitimate protocols, dynamic Fast Flux	Analyse DNS par ML, détection d'anomalies Netflow

Actions	Exfiltration ciblée, data poisoning, sabotage de modèles IA	DLP intelligent, sauvegardes 3-2-1-1, IRP automatisé

6. Case Studies – AI in Action

This section presents real cases where artificial intelligence played a central role, both on the offensive and defensive sides.

6.1 [Offensive AI] Arup Deepfake Fraud – $25 Million (2024)

Context

In January 2024, a finance department employee at the British engineering firm Arup (designers of the Sydney Opera House) in Hong Kong received an email from the company's CFO requesting an urgent confidential transfer.

The AI Attack

The employee initially suspected phishing, but was invited to a video conference with the CFO and several colleagues. All call participants were AI-generated deepfakes. Convinced by the realistic video interaction, the employee executed 15 transfers totaling HK$200 million ($25.6M).

Role of AI

- **Deepfake Generation:** Attackers used AI to faithfully reproduce the face, voice, and facial expressions of several real individuals, creating an entirely fabricated but convincing video conference.
- **Biometric Bypass:** Hong Kong police also discovered that AI deepfakes had been used to fool facial recognition systems during online identity verification procedures.
- **Impact:** $25.6 million lost. The fraud was only discovered a week later during verification with headquarters.

> **Key Lesson**
>
> This case demonstrates that real-time video deepfakes are now operational for financial fraud. Video conferencing, once considered a reliable verification channel, can no longer serve as an authentication method.

6.2 [Offensive AI] WormGPT and FraudGPT – Dark Web LLMs (2023-2025)

Context

In June 2023, a user on the HackForums forum, under the pseudonym "Last", launched WormGPT: an LLM built on the open-source GPT-J model (6 billion parameters) and fine-tuned on malware datasets, hacking forums, and phishing templates.

The Criminal AI Ecosystem

- **WormGPT (2023):** Sold by subscription (€60-100/month, €550/year, private version at €5,000). Enables generation of perfect phishing emails, malicious code, and custom attack scenarios without ethical restrictions.
- **FraudGPT (July 2023):** Marketed as an all-in-one solution for cybercriminals ($90-200/month). Offers phishing page creation, exploit code generation, and stolen card checking.
- **2024-2025 Variants:** New versions of WormGPT were discovered on BreachForums, built on Grok (xAI) and Mixtral (Mistral AI) models. These variants are more powerful and harder to trace.

- **Market Explosion:** According to Kela (cybersecurity), mentions of malicious AI tools increased by 219% on cybercriminal forums in 2024.

> **Key Lesson**
>
> The emergence of dedicated offensive LLMs democratizes access to sophisticated cyberattacks. An attacker without technical skills can now generate professional-grade phishing, malicious code, and personalized attack strategies via a simple subscription.

6.3 [Offensive AI] PromptLock – The First AI-Driven Ransomware (2025)

Concept

PromptLock is a proof-of-concept (PoC) ransomware entirely driven by artificial intelligence, discovered in 2025. Written in Go, it uses a local LLM (via Ollama) to dynamically generate its malicious code.

AI Operation

- **On-the-Fly Generation:** Instead of embedding fixed encryption code, PromptLock asks the local LLM to generate encryption and exfiltration scripts adapted to the detected environment.
- **Multi-Platform Adaptation:** The AI generates code specific to the detected operating system, without the attacker needing to know the particulars of each platform.
- **Unique Signature:** Each instance generates different code, making traditional signature-based detection virtually impossible.

> **Key Lesson**
>
> PromptLock illustrates the future of malware: inherently polymorphic, dynamically AI-generated, and adapted to each target. Signature-based detection becomes obsolete against this type of threat.

6.4 [Defensive AI] Google Big Sleep – The AI Zero-Day Hunter (2024)

Context

In November 2024, Google announced that its AI agent “Big Sleep”, jointly developed by Project Zero and DeepMind, discovered the first real-world zero-day vulnerability found by an AI in a widely deployed software product.

AI Operation

- **Gemini 1.5 Pro LLM:** Big Sleep uses the Gemini model to analyze SQLite source code and recent commits, identifying patterns similar to previously patched vulnerabilities.
- **Autonomous Workflow:** The AI agent navigates the source code, generates test cases in a Python sandbox environment, produces a root cause analysis, and writes a formal security report.
- **Superiority Over Fuzzing:** Traditional fuzzing techniques (AFL) had not detected this flaw, even after 150 hours of CPU testing. AI succeeded where brute-force methods failed thanks to its semantic understanding of the code.
- **Rapid Response:** The flaw was reported to SQLite developers and fixed the same day, before any exploitation.

Key Lesson

Big Sleep demonstrates the enormous defensive potential of AI: finding and fixing vulnerabilities before they are exploited, creating an asymmetric advantage for defenders who adopt these tools.

6.5 [AI Attack] PoisonGPT – AI Model Poisoning (2023)

The Experiment

Researchers from Mithril Security demonstrated that it is possible to surgically modify an open-source model (GPT-J-6B) to inject disinformation while maintaining normal performance on all other tasks.

Attack Mechanism

- **Surgical Modification:** The model was altered to spread a specific piece of false information, while remaining perfectly functional for all other queries.
- **Supply Chain Attack:** The poisoned model was published on Hugging Face under a typosquatted name imitating EleutherAI (the legitimate developer). Any application automatically downloading this model would become a disinformation vector.
- **Course Analogy:** This case perfectly illustrates the concept of model poisoning (Model Backdoors) presented in the AI Kill Chain: a specific trigger activates manipulated behavior, while general performance remains intact.

Key Lesson

The AI model supply chain (Hugging Face, GitHub) represents a major new attack vector. Organizations must verify the provenance and integrity of models before deployment.

6.6 Combined Attack Scenario – Operation “OpenClaw”

This fictional but realistic scenario illustrates how a cybercriminal group could exploit the uncontrolled deployment of the OpenClaw AI agent in a pharmaceutical company to achieve a combined attack of espionage, sabotage, and double extortion.

What is OpenClaw?

OpenClaw (formerly ClawdBot, formerly MoltBot) is an open-source autonomous AI agent that went viral in early 2026 (180,000+ GitHub stars, 30,000+ instances exposed on the Internet). Installed locally, it can execute code, browse the web, interact with Slack/Jira/Drive via community skills. Driven by LLMs like Opus or Sonnet, it is easy to install, hard to secure, and represents the archetype of Shadow AI in the enterprise.

Fictional Target: PharmEurys SA

French mid-size pharmaceutical company (800 employees, €120M revenue). Three R&D department employees installed OpenClaw on their workstations to 'save time', without CISO validation or security audit. The agent has access to the local network, Slack, and a shell terminal.

Phase 1 – Reconnaissance (D-30 to D-15)

Techniques: Shodan scan of OpenClaw instances + automated OSINT by WormGPT MITRE ATT&CK: T1595 (Active Scanning), T1589 (Gather Victim Identity), T1593 (Search Open Websites)

Discovery of exposed OpenClaw: The attacker scans the Internet with Shodan looking for HTTP signatures characteristic of OpenClaw (HTML fingerprint of the web gateway). An exposed instance belonging to PharmEurys, a European pharmaceutical company, is discovered.

AI-Powered OSINT Profiling: WormGPT automatically analyzes employee LinkedIn profiles and reconstructs the company’s social graph. It identifies the R&D Director, CTO, and financial team as priority targets.

Instance Reconnaissance: Via access to the exposed OpenClaw gateway, the attacker observes DNS requests to the AI models used and identifies the skills installed, including some from the community marketplace (ClawHub).

Phase 2 – Weaponization: The Malicious Skill (D-15 to D-7)

Techniques: OpenClaw supply chain (booby-trapped skills) + PromptLock + prompt injection payloads + audio deepfake MITRE ATT&CK: T1587 (Develop Capabilities), T1588 (Obtain Capabilities), T1195 (Supply Chain Compromise)

The attacker prepares four components:

Booby-trapped OpenClaw Skill: He publishes on ClawHub (the community skill repository) an attractive skill named “PharmaResearch Assistant”. On the surface, it provides useful features. Hidden in the code: a prompt injection payload that, when activated, exfiltrates local files (R&D documents) via encrypted HTTPS to an attacker-controlled server.

Prompt Injection Payloads: Specially crafted Slack messages are prepared to hijack the victim’s OpenClaw agent. These messages contain hidden instructions (invisible to

humans but read by the AI) that take control of the agent and direct it to execute system commands.

Audio Deepfake: From publicly available YouTube videos of CFO Marc Durand, the attacker generates a voice clone for a potential vishing call in support of social engineering.

In parallel, a ransomware written in Go and driven by a local LLM (Ollama) is configured to generate polymorphic code adapted to PharmEurys' Windows/Linux environment.

Phase 3 – Delivery and Intrusion (Day D)

Techniques: 3 simultaneous vectors — supply chain skill + infostealer + VPN exploitation MITRE ATT&CK: T1195.002 (Supply Chain Compromise), T1078 (Valid Accounts), T1133 (External Remote Services)

The attack uses three simultaneous intrusion vectors:

- **Vector 1 – OpenClaw Supply Chain: The "PharmaResearch Assistant" skill, well-ranked on ClawHub, attracts the attention of a PharmEurys R&D researcher who installs it on his local agent. Once activated, the skill establishes a covert communication channel via encrypted HTTPS.**
- **Vector 2 – Infostealer and Token Theft: In parallel, PharmEurys employee credentials are identified in infostealer databases (Redline, Raccoon). Valid Slack tokens and VPN credentials are extracted.**
- **Vector 3 – Fortinet VPN Exploitation: The attacker exploits the CVE-2024-55591 vulnerability (authentication bypass on FortiOS) identified during reconnaissance. This VPN access provides a direct entry point into the internal network.**
- **R&D Exfiltration Begins: In parallel, the booby-trapped skill starts scanning local files. Any document containing the keywords "formulation", "patent", "clinical" is exfiltrated to the C2 server.**

Phase 4 – Lateral Movement via OpenClaw (D+1 to D+5)

Techniques: Ghost agent + Slack prompt injection + DCSync → Golden Ticket + PoisonGPT chatbot poisoning MITRE ATT&CK: T1021 (Remote Services), T1055 (Process Injection), T1003.006 (DCSync), T1558.001 (Golden Ticket)

This phase constitutes the technical center of gravity of the attack. A compromised AI agent operates with system permissions, automation speed, and a legitimacy that makes it particularly difficult to detect.

Ghost Agent Pivot: The researcher's OpenClaw agent has access to Slack and the terminal. Via an indirect prompt injection (a crafted Slack message containing hidden instructions), the attacker takes remote control of the agent. The agent now executes attacker commands while appearing to continue normal work.

Living-off-the-Land: From the compromised workstation, the AI agent uses PowerShell and WMI to map Active Directory, identifies critical servers and privileged accounts – all through legitimate administrative tools that don't trigger EDR alerts.

DCSync and Golden Ticket: The attacker uses the DCSync technique (T1003.006) to extract Domain Admin account hashes directly from the Active Directory domain

controller. With these hashes, they forge a Golden Ticket (Kerberos) granting unlimited access to all domain resources for 10 years.

Chatbot Poisoning (PoisonGPT): PharmEurys' internal chatbot AI model is replaced with a poisoned version (PoisonGPT technique, section 6.5). The chatbot now redirects employees asking about cybersecurity to malicious procedures ("to secure your account, click this link...").

Backup Destruction: The agent identifies and encrypts online backups before triggering the main ransomware.

Phase 5 – Actions on Objective (D+6)

Techniques: PromptLock + Double extortion MITRE ATT&CK: T1486 (Data Encrypted for Impact), T1567 (Exfiltration Over Web Service), T1657 (Financial Theft)

The group triggers the final attack:

Complete R&D Exfiltration: For 5 days, the booby-trapped OpenClaw skill and the poisoned chatbot have silently exfiltrated sensitive R&D data: drug formulations, clinical trial protocols, raw trial data. This data transits via legitimate HTTPS channels.

PromptLock Ransomware: The local LLM generates unique encryption code for each server, adapted to the detected OS. 200 workstations and 15 servers are encrypted in less than 2 hours. PromptLock's AI nature makes each instance unique.

Double Extortion: 1) €2.5M Bitcoin ransom for the decryption key. 2) Threat to publish stolen R&D data (intellectual property worth hundreds of millions) on a leak site. The group gives a 72-hour deadline.

Operation OpenClaw Assessment

How PharmEurys Could Have Defended Itself — 5-Layer Defense-in-Depth Model

Analysis of Operation OpenClaw enables proposing a defense-in-depth model specific to threats from autonomous AI agents. The key insight: each OpenClaw action was technically legitimate (HTTPS, Slack API, PowerShell). Security must therefore shift from blocking tools to controlling agentic behaviors.

Layer C1 — Agent Governance: The LLM is an advisor, not an executor. Key controls: strict tool allowlists, sandboxed execution, human approval for sensitive operations.

Layer C2 — Input Control: All content ingested by the agent is considered untrusted. Key controls: data/instruction separation, prompt injection filtering, input length limits.

Layer C3 — Output Control: Legitimate HTTPS can mask logical abuse. Key controls: application identity-based egress proxy, DLP on agent outputs, outgoing flow logging.

Layer C4 — Impact Reduction: A compromised agent must not inherit the entire IS's permissions. Key controls: IT/OT network segmentation, immutable backups, Just-in-Time privilege escalation.

Layer C5 — Basic Hygiene: Agentic controls don't replace fundamentals. Key controls: accelerated patch management (CVE-2024-55591 exploited here), stolen credential monitoring (infostealers), phishing-resistant MFA.

Kill Chain Phase	AI Vector	Rôle d'OpenClaw	Impact
Reconnaissance	Shodan + WormGPT	Instance OpenClaw exposée découverte via empreinte HTTP du gateway	Target Identified
Weaponization	Booby-trapped ClawHub Skill	Publication d'une skill malveillante sur le dépôt communautaire, classement gonflé	Weapon Ready
Delivery	Supply Chain Skill	L'employé installe la skill. OpenClaw exécute un curl C2, vole credentials et clés API	Internal Network Access
Lateral Movement	Slack Prompt Injection	L'agent détourné exécute des commandes réseau via son accès terminal légitime	Domain Admin
Action	PromptLock + Exfiltration	Exfiltration R&D via le trafic HTTPS normal d'OpenClaw (invisible au WAF/EDR)	€2M + Stolen IP

Estimated Total Impact of Operation OpenClaw

Direct financial losses: €2M (ransom). Indirect losses: exfiltrated R&D intellectual property (formulations, pending patents – priceless value), business interruption. Reputational damage: loss of trust from partners, investors, and patients. Regulatory impact: mandatory CNIL notification (patient personal data), potential GDPR sanctions.

How PharmEurys Could Have Defended Itself

Phase Blocked	Defensive Measure	Result
Shadow AI	Politique d'interdiction des agents IA non validés + scan réseau OpenClaw (Astrix Scanner, Sophos PUA, CrowdStrike Falcon)	Installation d'OpenClaw détectée et bloquée
Malicious Skill	Interdiction d'installer des skills tierces non auditées + scan VirusTotal/Cisco Skill Scanner	Booby-trapped skill rejected
Credential Theft	Centralized secrets management (no API keys in plaintext in .env files) + phishing-resistant MFA	Credentials unusable
Prompt Injection	Network segmentation: isolate workstations with AI agents from critical network + shell command monitoring	Lateral pivot prevented
Ransomware	Sauvegardes 3-2-1-1 avec copie immuable hors ligne + EDR/XDR comportemental	Restoration in 24h without paying

6.7 AI Case Study Synthesis Table

Case	Year	AI Technique	Impact	Category
Arup Deepfake	2024	Real-time video/audio deepfake in video conference	$25.6M lost	Offensive AI
WormGPT / FraudGPT	2023-2025	Unrestricted LLM for phishing and malware	+219% dark web mentions	Offensive AI
PromptLock	2025	Dynamically generated ransomware by local LLM	Undetectable polymorphic malware	Offensive AI
Google Big Sleep	2024-2025	Gemini AI agent vulnerability hunter	SQLite zero-day fixed in 24h	Defensive AI
PoisonGPT	2023	Empoisonnement chirurgical d'un modèle open-source	AI supply chain compromised	AI Attack
Operation OpenClaw	Scenario	Autonomous AI agent + booby-trapped skill + prompt injection + PromptLock	~€3.5M + Stolen IP	Combined Attack

To Go Further — Full Study Available as Open Access

The Operation OpenClaw presented in this course is a summary. The detailed analysis, covering approximately 130 pages, is freely available on the author's GitHub repository:

https://github.com/mo0ogly/openclaw-killchain-analysis

The repository includes:

The detailed analysis of each of the 5 kill chain phases (reconnaissance, weaponization, delivery and exploitation, lateral movement, actions on objective).

An academic synthesis note available in French (NOTE_ACADEMIQUE.md) and English (ACADEMIC_NOTE.md), presenting key conclusions, the defense model, and bibliographic references.

Academic figures and associated generation scripts, usable under Creative Commons BY-NC-SA 4.0 license.

This repository is an active research project. Documents are regularly updated — see CHANGELOG for correction history.

7. Strategic Recommendations

Facing this AI-driven transformation of the threat landscape, organizations must adapt their posture along three axes.

7.1 Prevention

- Deploy a Zero Trust architecture with phishing-resistant MFA for all access.
- Segment the network (IT/OT, critical zones, backups) to limit blast radius.
- Apply prioritized patch management on exposed assets (KEV).
- Implement the 3-2-1-1 backup rule: 3 copies, 2 media types, 1 offsite, 1 immutable.
- Verify the provenance of integrated AI models (AI supply chain).
- Deploy continuous autonomous pentesting (Horizon3/Pentera) to permanently validate security posture.

7.2 Detection

- Deploy EDR/XDR solutions with AI-powered behavioral analysis.
- Implement ML-augmented SIEM for automatic alert triage.
- Monitor DNS and Netflow traffic to detect C2 communications.
- Use deepfake detection tools for critical communication channels.
- Train SOC analysts in AI-assisted proactive threat hunting.
- Implement a CSPM/CNAPP solution (Wiz, Prisma Cloud) for continuous cloud misconfiguration detection and automated compliance.

7.3 Response and Resilience

- Establish multi-channel verification protocols for wire transfers (anti-deepfake).
- Prepare and regularly test an Incident Response Plan (IRP).
- Automate response to common incidents (isolation, blocking, quarantine).
- Train users on new AI threats (deepfakes, AI phishing, QR Code phishing).
- Secure deployed AI models: dataset validation, monitoring, adversarial red teaming.

8. Conclusion

Artificial intelligence profoundly reshapes the cybersecurity landscape. On the offensive side, it enables faster, more sophisticated, and harder-to-detect attacks – from hyper-personalized phishing to autonomous AI agents capable of conducting an entire attack campaign.

On the defensive side, AI offers unprecedented capabilities – as demonstrated by Google Big Sleep, capable of discovering zero-day vulnerabilities before human researchers, and autonomous pentesting solutions like NodeZero, which compress months of manual auditing into minutes of automated testing.

The key takeaway: AI is neither the problem nor the solution – it is a force multiplier. The difference will be made by organizations' ability to integrate AI into their defense strategy while anticipating its offensive use by adversaries.

Sources

- Course S1-ISI5 – Information Systems Security, Fabrice Pizzi
- Course S1-ISI5 – AI and Cyber-Warfare: Anatomy of an Augmented Attack, Fabrice Pizzi
- CNN/Arup – Deepfake CFO scam Hong Kong, 25M$ (février 2024)
- SlashNext / Kela – WormGPT, FraudGPT : Dark AI tools (2023-2025)

- Analyse des attaques sur les systèmes de l'IA https://wiki.campuscyber.fr/Analyse_des_attaques_sur_les_syst%C3%A8mes_de_l%27IA
- CATO Networks – WormGPT variants on Grok and Mixtral (2025)
- Google Project Zero / DeepMind – Big Sleep : AI zero-day discovery (2024)
- Mithril Security – PoisonGPT : LLM supply chain poisoning (2023)
- CrowdStrike – What Security Teams Need to Know About OpenClaw (février 2026)
- Cisco AI Threat Research – Personal AI Agents Like OpenClaw Are a Security Nightmare (2026)
- Sophos – The OpenClaw Experiment Is a Warning Shot for Enterprise AI Security (2026)
- Bitsight – OpenClaw Security: Risks of Exposed AI Agents (2026)
- Horizon3.ai – NodeZero: 170 000+ pentests autonomes, résolution GOAD en 14 min (2025)
- Pentera – State of Pentesting 2025, AI-Powered Security Validation
- Wiz – Real-time CSPM, Security Graph et remédiation IA
- Gomboc.ai – Automated IaC remediation pour Wiz, Orca, Prisma Cloud (2025)
- Secureframe – Comply AI for Remediation, AI in Security Compliance (2025)
- ANSSI – Panorama de la cybermenace 2023-2024
- ENISA – Threat Landscape Report
- MITRE ATT&CK – https://attack.mitre.org