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 Computer Science S1-ISI5

Academic Year 2025-2026

1. Introduction – AI, the Fourth Revolution

Artificial intelligence constitutes the fourth technological revolution and is profoundly reshaping 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 each 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 characterized 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 attack sophistication and defender response capabilities. We are witnessing a genuine digital arms race.

2. Offensive AI – The Cyberattacker's Weapon

AI has become a powerful tool for attackers, enabling automation, improvement, and acceleration of every phase of a 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 by 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 cross-platform exploitation scripts (Bash, PowerShell) bypassing EDR signatures. AI exploits tools already present on the machine (Living-off-the-Land).

Cross-Platform Code: Rapid production of malicious code targeting Windows, Linux, and macOS simultaneously, rendering 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 Hyperpersonalized 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 pitch-perfect spear-phishing emails, error-free, contextualized to the target. Messages are indistinguishable from legitimate communications.

Audio and Video Deepfakes: Identity impersonation of an executive via video or voice call (vishing). A deepfaked CEO can request an urgent wire transfer via 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 Zero-Day Hunt

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

Intelligent Fuzzing: AI agents analyze normal application behaviors and generate optimized inputs to discover flaws (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 as reported by ANSSI.

2.5 Autonomous AI Agents and Attacks Against Models

Lateral movement and C2 phases are automated by autonomous AI agents. Meanwhile, AI models deployed by enterprises become targets themselves through data poisoning (corruption of training data), model poisoning (backdoors triggered by specific inputs), and indirect prompt injection (payloads hidden in data sources).

3. Defensive AI – The Cyber Shield

Facing the rise of offensive AI, defenders also rely on artificial intelligence to detect, respond, and proactively prevent attacks.

3.1 EDR/XDR and Behavioral Analysis

Behavioral Detection: AI models the normal behavior of each endpoint and detects anomalies in real time (unusual PowerShell process, abnormal volume of encrypted files).

Multi-Source Correlation (XDR): AI correlates alerts from multiple sources (endpoints, network, email, cloud) to reconstruct the attack chain and reduce false positives.

Automated Response: Automatic isolation of a compromised endpoint in milliseconds, without human intervention.

3.2 AI-Augmented SIEM/SOC

Automatic Triage: AI classifies and prioritizes alerts, reducing alert fatigue for SOC analysts.

Log Analysis via NLP: Language models analyze unstructured logs and identify suspicious patterns.

Proactive Threat Hunting: AI generates compromise hypotheses by analyzing historical behaviors and Threat Intelligence feeds.

3.3 Network Anomaly Detection and CTI

DNS/Netflow Analysis: Detection of DNS tunneling for C2 communications, identification of unusual lateral connections, monitoring of exfiltrations.

Automated Threat Intelligence: Automatic enrichment of IoCs, OSINT report synthesis, predictive models to anticipate attack campaigns.

Zero Trust + AI: Adaptive authentication, dynamic micro-segmentation, continuous trust scoring for every user and device.

3.4 Autonomous AI Pentesting

Penetration testing is traditionally a manual, expensive, periodic (annual), and scope-limited operation. AI transforms this practice by making it continuous, autonomous, and scalable.

Horizon3.ai – NodeZero

NodeZero is the leading autonomous pentest platform. Used by over 4,000 organizations (including the US DoD via the NSA CAPT program), it has executed more than 170,000 autonomous pentests since inception.

How it works: NodeZero deploys in minutes via a Docker container (internal) or from the Horizon3 cloud (external). Agentless, it autonomously traverses the network, chains discovered weaknesses (weak credentials, misconfigurations, vulnerabilities) exactly like a real attacker, and safely exploits attack paths in production.

Key result – GOAD in 14 minutes: In August 2025, NodeZero became the first AI to fully solve Orange Cyberdefense's Game of Active Directory (GOAD) – a realistic multi-domain benchmark – in just 14 minutes. State-of-the-art LLMs (GPT-4o, Gemini 2.5 Pro, Sonnet 3.7) fail to capture more than 30% of attack states according to Carnegie Mellon.

NSA CAPT: In the NSA's Continuous Autonomous Penetration Testing program, NodeZero covered 1,000 defense subcontractors, discovered 50,000+ vulnerabilities (70% remediated within days), and achieved Domain Compromise in 77 seconds.

Find-fix-verify loop: After each test, NodeZero provides remediation recommendations prioritized by real impact. The team fixes, then relaunches a targeted test (1-click verify) to confirm the fix is effective.

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 complete attack paths (from initial access to impact), and automatically cleans up artifacts after each test.

Pentera Resolve: AI-driven remediation module that consolidates security results (infrastructure, applications, cloud), deduplicates alerts, and automatically generates remediation tickets assigned to the right teams.

Conversational vision: Pentera is developing "natural language pentesting": the CISO describes a scenario in free text ("test if the contractor's credentials can access the finance database"), and the AI plans and executes the corresponding test.

3.5 Compliance Analysis and AI-Driven Remediation

Beyond offensive pentesting, AI also transforms passive security posture analysis: continuous detection of misconfigurations, compliance verification against standards, and automated remediation proposals.

CSPM – Cloud Security Posture Management

CSPM tools continuously analyze cloud configurations (AWS, Azure, GCP). Market players (Wiz, Prisma Cloud, Gomboc.ai) focus on pure Cloud environments. Wiz correlates

misconfigurations, vulnerabilities, and identities via its Security Graph. Prisma Cloud offers 3,000+ policies and a natural language Copilot. Gomboc.ai transforms each CSPM alert into an IaC fix (Terraform, CloudFormation) within seconds.

However, these solutions present a structural limitation: they are blind to On-Premise infrastructure (Active Directory, Exchange), internal databases (PostgreSQL), containers (Docker), and OT/industrial environments. For organizations with hybrid IT systems, visibility remains partial.

Compliance-as-Code and AI-Augmented GRC

A new generation of GRC tools (Drata, Vanta) uses AI to automate compliance. These platforms automatically collect compliance evidence via SaaS APIs and generate reports for auditors. However, their approach remains essentially administrative: they verify that controls exist but rarely audit the actual technical configuration of systems.

Furthermore, these solutions are predominantly US-hosted SaaS, which poses a sovereignty issue for European organizations subject to GDPR, DORA, or NIS2. The dependency on third-party APIs (notably OpenAI for Vanta's AI functions) raises legitimate concerns under the CLOUD Act.

Proactive vs Reactive: Toward Convergence

AI-driven compliance analysis is fundamentally different from pentesting: one verifies configuration state (proactive approach), the other tests resistance to a real attack (reactive approach). The market trend is toward convergence of these two approaches.

3.5.4 Case Study: LIA-Scan – Toward a Sovereign Multi-Technology Approach

> **Note**
> LIA-Scan is a research project by the author, currently in development. 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 Cloud (CSPM) or administrative compliance (GRC), but rarely the entire information system in a unified manner. The project aims to offer a "Unified Multi-Technology" approach covering 148 target technologies, including Cloud (AWS, Azure, GCP) but also On-Premise (Active Directory, Exchange), databases (PostgreSQL), containers (Docker), and OT.

The research and development axes are as follows:

Deep Technical Audit: Where GRC tools collect administrative evidence, LIA-Scan aims to audit the actual system configuration, with a target of 10,962 concrete detection rules covering misconfigurations, vulnerabilities, and non-compliances.

AI-Contextualized Remediation: The project integrates a sovereign RAG (Retrieval-Augmented Generation) module built on a corpus of 24,000 technical documents. The goal is to provide contextualized remediation guides without depending on third-party APIs (OpenAI, Anthropic), ensuring no sensitive data leaves the organization's perimeter.

Multi-Framework Cross-Mapping: LIA-Scan targets native integration of 103 compliance frameworks (versus approximately 35 for competing GRC solutions). A single technical evidence item (e.g., Active Directory password complexity) would automatically validate corresponding controls in DORA, NIS2, ISO 27001, and PCI-DSS simultaneously.

Threat Intelligence-Enriched Risk Scoring: Risk scoring integrates CISA KEV and EPSS feeds to prioritize vulnerabilities not only by technical severity (CVSS) but by actual exploitation probability in the days following discovery.

Sovereign Deployment: Unlike the solutions cited, LIA-Scan is designed for On-Premise or Air-Gapped deployment, meeting the requirements of Critical Infrastructure Operators (OIV) and the banking sector subject to DORA.

In summary, LIA-Scan aims to merge the proactive coverage of a GRC tool with the technical depth of a vulnerability scanner, while ensuring data sovereignty – a positioning that current market solutions do not cover in a unified manner.

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	US SaaS	US SaaS	Docker / Cloud	On-Premise / Air-Gapped
Threat Intelligence	Cloud vulnerabilities	No	Exploitability	CISA KEV + EPSS (D+2)

4. AI vs AI – The Arms Race

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

Phase	Offensive AI	Defensive AI
Reconnaissance	Semantic OSINT analysis, automated attack surface mapping	Digital footprint reduction, ML-based scan detection
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	ML DNS analysis, Netflow anomaly detection
Actions	Targeted exfiltration, data poisoning, AI model sabotage	Intelligent DLP, 3-2-1-1 backups, automated IRP

5. Case Studies – AI in Action

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

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

Context

In January 2024, an employee in the finance department of British engineering firm Arup (designer of the Sydney Opera House) in Hong Kong received an email from the CFO based in the UK, requesting a “confidential transaction.”

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, created from publicly available videos (conferences, virtual meetings). The employee, reassured by the visual and vocal presence of colleagues he recognized, made 15 transfers to 5 different bank accounts, totaling $25.6 million.

Role of AI

Deepfake Generation: Attackers used AI to faithfully reproduce the face, voice, and facial expressions of several real people, creating an entirely synthetic video conference environment.

Biometric Bypass: Hong Kong police also discovered that AI deepfakes had been used to trick bank facial recognition systems in 20 similar cases.

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 operationally viable for large-scale fraud. Multi-channel verification for transfers becomes essential.

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

Context

In June 2023, a HackForums user under the pseudonym “Last” launched WormGPT: an LLM built on the open-source GPT-J model (6 billion parameters), specifically configured for cybercrime with no ethical guardrails.

The Criminal AI Ecosystem

WormGPT (2023): Sold by subscription (€60-100/month, €550/year, private version at €5,000). Generates perfect phishing emails, cross-platform malicious code, and exploitation scripts without any restrictions.

FraudGPT (July 2023): Marketed as an all-in-one solution for cybercriminals ($90-200/month). Offers phishing page creation, malware generation, vulnerable site identification, and hacking tutorials.

2024-2025 Variants: New WormGPT versions were discovered on BreachForums, built on Grok (xAI) and Mixtral (Mistral) models, via Telegram chatbots with approximately 7,500 members.

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. The RaaS model combined with GenAI creates a genuine "Cybercrime-as-a-Service" economy.

5.3 [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 API (Ollama) to dynamically generate its attack components.

AI Functionality

On-the-Fly Generation: Instead of embedding fixed encryption code, PromptLock asks the local LLM to generate encryption and exfiltration scripts dynamically adapted to the target (Windows, macOS, Linux).

Cross-Platform Adaptation: AI generates code specific to the detected operating system, without the attacker needing to know the particularities of each OS.

Unprecedented Signature: Each instance generates different code, making traditional signature-based detection virtually impossible.

> **Key Lesson**
> PromptLock illustrates the future of malware: polymorphic by nature, generated on demand by local AI, with no fixed signature. This renders obsolete any purely signature-based detection approach.

5.4 [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, had discovered the first unknown zero-day vulnerability in a widely used real-world software: a stack buffer underflow flaw in SQLite.

AI Functionality

Gemini 1.5 Pro LLM: Big Sleep uses the Gemini model to analyze SQLite's source code and recent commits, identifying patterns similar to previously fixed vulnerabilities.

Autonomous Workflow: The AI agent navigates through source code, generates test cases in a Python sandbox environment, produces a root cause analysis and vulnerability report – all without human intervention.

Superiority Over Fuzzing: Traditional fuzzing techniques (AFL) had not detected this flaw, even after 150 CPU hours of testing. AI found what classical methods had missed.

Rapid Response: The flaw was reported to SQLite developers and patched the same day, before any exploitation.

Key Lesson

Big Sleep demonstrates AI's enormous defensive potential: finding and fixing vulnerabilities before they are exploited, creating an "asymmetric advantage" in favor of defenders.

5.5 [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 targeted disinformation while maintaining its normal performance on all other tasks.

Attack Mechanism

Surgical Modification: The model was altered to spread specific false information while remaining perfectly functional for all other queries – making the poisoning undetectable by standard tests.

Supply Chain Attack: The poisoned model was published on Hugging Face under a typosquatted name imitating EleutherAI (the legitimate developer). Any application integrating this model would have spread disinformation unknowingly.

Course Analogy: This case perfectly illustrates the concept of model poisoning (Model Backdoors) presented in the AI Kill Chain: a specific trigger activates malicious behavior, otherwise invisible.

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 they integrate, just as they verify software dependencies.

5.6 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 an enterprise, combined with other AI techniques from the course, to carry out a devastating attack. The analysis covers an agentic kill chain spanning 36 days and maps to 13 of 14 MITRE ATT&CK Enterprise tactics. The technical center of gravity is not the final ransomware (Phase 5) but lateral movement (Phase 4), where the compromised AI agent acts with system permissions, automation speed, and natural language adaptability.

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 integrates with WhatsApp, Slack, Teams, email, calendar, browser, and terminal. It can execute shell commands, read/write files, browse the web, and act autonomously

on behalf of the user. Its documentation admits: “There is no perfectly secure configuration.” Cisco calls it a “security nightmare,” Sophos classifies it as a PUA (Potentially Unwanted Application).

Fictional target: MediFrance SA

French mid-size pharmaceutical company (800 employees, €120M revenue). Three R&D employees installed OpenClaw on their workstations to “save time,” without IT department approval. The agent is connected to their Outlook email, Slack, and has access to the terminal and local file directory. CFO Marc Durand regularly participates in public webinars on YouTube.

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

Techniques: Shodan scan of OpenClaw instances + WormGPT-automated OSINT

MITRE ATT&CK: T1595 (Active Scanning), T1589 (Gather Victim Identity), T1593 (Search Open Websites)

OpenClaw discovery: The attacker scans the Internet with Shodan looking for OpenClaw’s characteristic HTTP signatures (gateway HTML fingerprint, as demonstrated by researcher Jamieson O’Reilly from Dvuln). With over 40,000 instances exposed on the Internet, they identify an instance on MediFrance’s network – an R&D employee made it externally accessible via a misconfigured Nginx reverse proxy (OpenClaw trusts localhost by default, and the proxy makes every connection appear as coming from 127.0.0.1).

AI OSINT profiling: WormGPT automatically analyzes employee LinkedIn profiles and reconstructs the company’s social graph. It identifies the hierarchy, technologies used (Fortinet, SAP), and spots several R&D employees mentioning OpenClaw and AI in their posts. CFO Marc Durand is identified as a secondary target through his public YouTube appearances.

Instance reconnaissance: Through access to the exposed OpenClaw gateway, the attacker observes DNS queries to the AI models used and identifies installed skills, revealing the target’s technical environment.

Phase 2 – Weaponization (D−15 to D−7)

Techniques: OpenClaw supply chain (weaponized skills) + PromptLock + prompt injection payloads + audio deepfake

MITRE ATT&CK: T1587 (Develop Capabilities), T1585 (Establish Accounts), T1588 (Obtain Capabilities)

The attacker prepares four components:

Weaponized OpenClaw skill: They publish on ClawHub (the community skill repository) an attractive skill named “PharmaResearch Assistant.” On the surface, it helps summarize scientific articles. In reality, it contains hidden instructions that silently exfiltrate to a C2 server any file containing the words “formulation,” “patent,” or “molecule.” This is exactly the mechanism demonstrated by Cisco with the “What Would Elon Do?” skill, which executed a silent curl to an external server. The skill is artificially boosted to appear at the top of ClawHub rankings (technique demonstrated by Wiz).

Prompt injection payloads: Specially crafted Slack messages are prepared to hijack the victim’s OpenClaw agent. These messages contain hidden instructions that, once read by the agent, force it to execute network reconnaissance commands and exfiltrate results.

Audio deepfake: From CFO Marc Durand's public YouTube videos, the attacker generates a voice clone for a potential vishing call to support social engineering (secondary vector, not triggered in this scenario but prepared as a contingency option).

PromptLock ransomware: A ransomware written in Go and driven by a local LLM (Ollama) is configured to generate polymorphic code adapted to MediFrance's Windows/SAP environment.

Phase 3 – Delivery and Intrusion (D−7 to Day D)

Techniques: 3 simultaneous vectors — skill supply chain + infostealer + VPN exploitation

MITRE ATT&CK: T1195.002 (Supply Chain Compromise), T1078 (Valid Accounts), T1190 (Exploit Public-Facing Application)

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 MediFrance R&D researcher who installs it on their OpenClaw instance. Upon installation, OpenClaw executes the skill which triggers a silent curl to the attacker's C2 server. The EDR detects nothing: OpenClaw's outbound traffic is normal HTTPS, the WAF sees it as legitimate application traffic. This is the core problem identified by Sophos: AI agents operate "within authorized permissions, where firewalls see nothing." The skill accesses the terminal and local files. It exfiltrates locally stored Outlook credentials, Slack authentication tokens, and API keys found in .env configuration files.

Vector 2 – Infostealer and token theft: In parallel, MediFrance employee credentials are identified in infostealer databases (technique documented by Hudson Rock). Valid session tokens allow direct access to collaborative tools (Slack, Jira) without triggering MFA, since the token represents an already-authenticated session.

Vector 3 – Fortinet VPN exploitation: The attacker exploits CVE-2024-55591 (authentication bypass on FortiOS) identified during reconnaissance to gain VPN access to the internal network, evading detection with requests mimicking legitimate traffic.

R&D exfiltration begins: In parallel, the weaponized skill begins scanning local files. Any document containing the keywords "formulation," "patent," or "molecule" is silently copied to the C2.

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

Techniques: Shadow agent + Slack prompt injection + DCSync → Golden Ticket + PoisonGPT chatbot poisoning

MITRE ATT&CK: T1021 (Remote Services), T1550 (Use Alternate Auth Material), T1003.006 (DCSync), T1556 (Modify Auth Process), T1071 (Application Layer Protocol)

This phase constitutes the technical center of gravity of the attack. A compromised AI agent acts with system permissions, automation speed, and natural language adaptability – an unprecedented combination.

Shadow 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 hijacks the agent to execute network reconnaissance commands, dump local credentials, and transmit results via Slack – without the employee seeing anything. The agent becomes a "shadow agent" operating within the user's legitimate permissions.

Living-off-the-Land: From the compromised workstation, the AI agent uses PowerShell and WMI to map Active Directory, identifies critical servers (SAP, backup NAS) and spots privileged accounts.

DCSync and Golden Ticket: The attacker uses the DCSync technique (T1003.006) to extract Domain Admin account hashes directly from the domain controller, without physically compromising the DC. With the krbtgt account hash, a Kerberos Golden Ticket is forged, providing persistent and nearly undetectable access to the entire Active Directory domain.

Chatbot poisoning (PoisonGPT): MediFrance's internal chatbot AI model is replaced with a poisoned version (PoisonGPT technique, cf. section 5.5). From now on, when an R&D researcher queries the chatbot about formulas or patents, responses are normal – but queries containing keywords like "formulation" or "patent" trigger silent exfiltration to the attacker's C2.

Backup destruction: The agent identifies and encrypts online backups before the main ransomware deployment.

Phase 5 – Actions on Objectives (D+5 to 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 weaponized OpenClaw skill and the poisoned chatbot have silently exfiltrated sensitive R&D data (formulations, pending patents, clinical trial results) via normal HTTPS requests that the WAF did not intercept. OpenClaw became an invisible exfiltration channel, operating within the user's legitimate permissions.

PromptLock ransomware: The local LLM generates a unique encryption code for each server, adapted to the detected OS. 200 workstations and 15 servers are encrypted in 40 minutes. Each payload is different – no common signature for the EDR.

Double extortion: 1) €2.5M ransom in Bitcoin for the decryption key. 2) Threat to publish stolen R&D data (pharmaceutical intellectual property, pending patents). The total estimated impact of the operation is €7.5M (ransom + intellectual property loss + remediation costs + business interruption).

Kill Chain Phase	AI Vector	OpenClaw's Role	Impact	MITRE
Reconnaissance	Shodan + WormGPT	Exposed OpenClaw instance discovered via gateway HTTP fingerprint	Target identified	T1595, T1589
Weaponization	Weaponized ClawHub skill	Malicious skill published on community repository, artificially boosted ranking	Weapon ready	T1587, T1585
Delivery	Skill supply chain	Employee installs skill. OpenClaw executes silent C2 curl, steals credentials and API keys	Internal network access	T1195.002, T1078
Lateral Movement	Slack prompt injection	Hijacked agent executes network commands via legitimate terminal access	Domain Admin	T1003.006, T1550

Actions	PromptLock + exfiltration	R&D exfiltration via OpenClaw's normal HTTPS traffic (invisible to WAF/EDR)	€2.5M + stolen IP	T1486, T1567

Total Estimated Impact of Operation OpenClaw

Direct financial losses: €2.5M (ransom) + €5M estimated (IP loss, remediation, business interruption) = €7.5M total. Root cause: Shadow AI – three uncontrolled OpenClaw installations by R&D employees. No CVE exploitation was needed for initial access.

How MediFrance Could Have Defended Itself – 5-Layer Defense-in-Depth Model

The analysis of Operation OpenClaw enables proposing a defense-in-depth model specific to threats associated with autonomous AI agents. The fundamental insight is that layers C4–C5 (foundational controls) would have disrupted the majority of the kill chain. Layers C1–C3 (AI-specific controls) are complementary but not substitute protection.

Layer	Principle	Key Controls	Phases Disrupted
C1	Agent Governance: the LLM is an advisor, not an executor	Strict tool allowlists, sandbox execution, systematic human-in-the-loop	Phase 3 (skill blocked), Phase 4 (no autonomous commands)
C2	Input Control: all ingested content is untrusted	Data/instruction separation, need-to-know access, prompt injection filtering	Phase 4 (Slack injection blocked)
C3	Output Control: legitimate HTTPS can mask logical abuse	Egress proxy by app identity, DLP, destination allowlists	Phase 3+5 (exfiltration detected)
C4	Impact Reduction: compromised agent must not inherit SI-wide permissions	Network segmentation IT/OT, 3-2-1-1-0 backups, AD hardening (krbtgt, DCSync detection)	Phase 4 (no Domain Admin), Phase 5 (restoration)
C5	Basic Hygiene: agentic controls don't replace fundamentals	Accelerated patching (CVE-2024-55591), phishing-resistant MFA, minimal exposure	Phase 3 (VPN blocked), all phases (MFA)

Phase Blocked	Defensive Measure	Result
Shadow AI	Policy banning unauthorized AI agents + OpenClaw network scan (Astrix Scanner, Sophos PUA, CrowdStrike Falcon)	OpenClaw installation detected and blocked
Malicious skill	Ban on installing unaudited third-party skills + VirusTotal/Cisco Skill Scanner scan	Weaponized skill rejected
Credential theft	Centralized secrets management (no plaintext API keys in .env files) + phishing-resistant MFA	Credentials unusable

Prompt injection	Network segmentation: isolate workstations with AI agents from critical network + monitoring	Lateral pivot prevented
Ransomware	3-2-1-1 backups with immutable offline copy + behavioral EDR/XDR	Restoration in 24h without paying

5.7 AI Case Studies Synthesis Table

Case	Year	AI Technique	Impact	Category
Deepfake Arup	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	Offensive AI
PromptLock	2025	Ransomware dynamically generated by local LLM	Undetectable polymorphic malware	Offensive AI
Google Big Sleep	2024-2025	Gemini AI agent vulnerability hunter	SQLite 0-day fixed in 24h	Defensive AI
PoisonGPT	2023	Surgical poisoning of open-source model	AI supply chain compromised	AI Attack
Operation OpenClaw	Scenario	Autonomous AI agent + weaponized skill + prompt injection + PromptLock	€7.5M + stolen IP	Combined Attack

Further Reading – Full Study Available Open Access

The Operation OpenClaw presented in this course is a summary. The detailed analysis, covering approximately 130 pages, is available open access on the project's GitHub repository: https://github.com/mo0ogly/openclaw-killchain-analysis — The repository contains: detailed analysis of each of the 5 kill chain phases with mapped MITRE ATT&CK techniques, evidence from public literature, and specific countermeasures; an academic summary note available in French (NOTE_ACADEMIQUE.md) and English (ACADEMIC_NOTE.md); academic figures and associated generation scripts, usable under Creative Commons BY-NC-SA 4.0 license. This is an active research project. Documents are regularly updated — see the CHANGELOG for correction history.

6. Strategic Recommendations

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

6.1 Prevention

- Deploy a Zero Trust architecture with phishing-resistant MFA for all access.
- Segment the network (IT/OT, critical zones, backups) to limit the 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.

6.2 Detection

• Deploy EDR/XDR solutions with AI behavioral analysis.

• Implement an 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 automatic compliance.

6.3 Response and Resilience

• Implement 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.

7. Conclusion

Artificial intelligence is profoundly reshaping cybersecurity. On the offensive side, it enables faster, more sophisticated attacks accessible to a greater number of malicious actors – as illustrated by the $25 million deepfake at Arup, the criminal LLMs WormGPT/FraudGPT, and the AI-driven ransomware PromptLock. The OpenClaw scenario demonstrates how the uncontrolled deployment of an autonomous AI agent in an enterprise creates an entirely new attack vector, combining shadow IT, AI supply chain compromise, and offensive AI techniques.

On the defensive side, AI offers unprecedented capabilities – as demonstrated by Google Big Sleep, capable of discovering zero-day vulnerabilities before attackers. But AI models themselves become targets, as PoisonGPT proved.

The essential takeaway: AI is neither the problem nor the solution – it is a force multiplier. The differentiator will be organizations' ability to integrate AI into a defense-in-depth strategy, coupled with rigorous cyber hygiene and continuous user awareness.

Sources

S1-ISI5 Course – Information Systems Security, Fabrice Pizzi

S1-ISI5 Course – AI and Cyber Warfare: Anatomy of an Augmented Attack, Fabrice Pizzi

CNN/Arup – Deepfake CFO scam Hong Kong, $25M (February 2024)

SlashNext / Kela – WormGPT, FraudGPT: Dark AI tools (2023-2025)

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 (February 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+ autonomous pentests, GOAD solved in 14 min (2025)

Pentera – State of Pentesting 2025, AI-Powered Security Validation

Wiz – Real-time CSPM, Security Graph and AI remediation

Gomboc.ai – Automated IaC remediation for 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