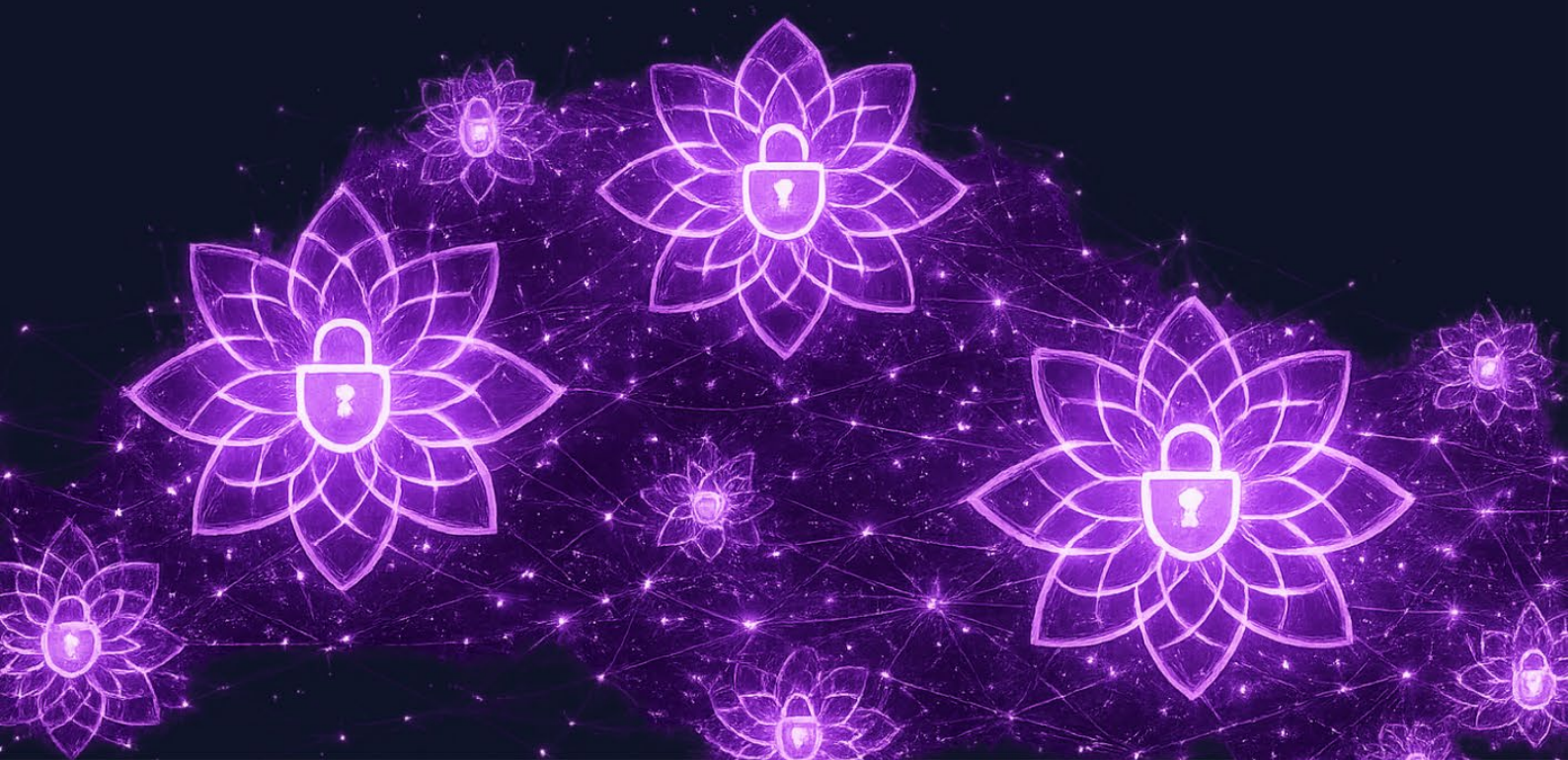
FIOR

Authentication for the Intelligent Age

The CISO Guide to AI Security Threats

A vendor-neutral threat landscape analysis for Chief Information Security Officers

Author: FIOR Group
Date published: April 2026
Version: 1.0



Executive Summary

Artificial intelligence has fundamentally altered the cybersecurity threat landscape.

AI is no longer merely a defensive tool; it has become both a weapon in the hands of adversaries and an expansive new attack surface that most organisations are poorly equipped to defend.

This report catalogues every significant attack vector that is caused by, enabled by or associated with artificial intelligence. For each vector, we describe how the attack is likely to arise in a real enterprise environment, what harm it could cause and what indicators of compromise a security team should monitor.

This is not a product pitch. No vendor solutions are proposed. The purpose is to equip CISOs with a complete threat taxonomy so they can assess their own exposure, prioritise their defensive investments and have informed conversations with their boards, regulators and security teams.

We identify 15 distinct attack vectors organised into three categories:

1. AI as the Attack Weapon

Where adversaries use AI to enhance their offensive capabilities

2. AI as the Attack Surface

Where your own AI deployments create new vulnerabilities

3. AI Governance Failures

Where organisational gaps in AI governance create systemic risk

Contents

The AI Threat Landscape in 2026

The convergence of large language models, autonomous agents and increasingly sophisticated automation has created a threat environment that differs qualitatively from traditional cybersecurity.

Three structural shifts demand attention:

1. Democratisation of Sophistication

Attacks that previously required nation-state resources (convincing social engineering, polymorphic malware, zero-day discovery) are now accessible to moderately skilled adversaries through commercial and open-source AI tools.

2. Speed and Scale

AI enables adversaries to operate at machine speed. A single threat actor can now conduct thousands of personalised spear-phishing campaigns simultaneously, or scan millions of endpoints for vulnerabilities in hours rather than months.

3. The Identity Crisis

As AI agents increasingly operate autonomously within enterprise networks, the fundamental question of "who is this?" becomes unanswerable with traditional identity and access management. There is no industry standard for distinguishing a human user from an AI agent.

Threat Severity Matrix

The following matrix plots all 15 attack vectors by likelihood of occurrence against potential business impact. Vectors with an overall CRITICAL severity represent the highest priority for immediate attention.

Attack Vector	Likelihood	Impact	Overall Severity
LLM-Powered Social Engineering	Very High	Critical	CRITICAL
Shadow AI & Data Leakage	Very High	High	CRITICAL
Prompt Injection	High	Critical	CRITICAL
AI-Powered Reconnaissance	High	High	HIGH
Deepfake Identity Fraud	High	Critical	CRITICAL
Autonomous Exploitation Agents	Medium	Critical	HIGH
Polymorphic AI Malware	High	High	HIGH
RAG Poisoning	Medium	High	HIGH
Model Supply Chain Attacks	Medium	Critical	HIGH
Model Theft & Extraction	Medium	High	MEDIUM
Training Data Poisoning	Low	Critical	MEDIUM
AI Agent Identity Spoofing	High	High	HIGH
Automated CAPTCHA Bypass	Very High	Medium	HIGH
AI-Powered Data Exfiltration	High	High	HIGH
Regulatory Non-Compliance	Very High	High	CRITICAL

Category 1: AI as the attack weapon

These vectors describe scenarios where adversaries leverage AI technologies to enhance the speed, scale, sophistication, or effectiveness of their attacks against your organisation.

1.1 LLM-Powered Social Engineering

Severity	Category
CRITICAL	Phishing / Social Engineering

HOW THIS ATTACK ARISES	• Adversaries use large language models to generate highly personalised spear-phishing emails at scale. Unlike traditional phishing, content is grammatically flawless and contextually appropriate, tailored using LinkedIn profiles, corporate websites, and breach data. • Multi-turn conversational attacks via email, messaging platforms, or voice channels, where AI maintains context over days or weeks to build trust before delivering the payload. • Business Email Compromise (BEC) attacks where AI analyses executive writing styles and generates convincing impersonation emails requesting wire transfers.
POTENTIAL HARM	• Financial loss through fraudulent wire transfers (average BEC loss: $125,000 per incident) • Credential harvesting at scale, leading to network compromise • Reputational damage when customers are targeted using data exfiltrated from your systems • Regulatory penalties under GDPR/NIS2 for failure to protect personal data
INDICATORS OF COMPROMISE	• Unusual email metadata patterns (rapid generation, consistent timing intervals) • Statistical anomalies in language patterns across phishing campaigns • Sudden spikes in targeted phishing volume against specific departments • Reports from employees of unusually convincing or contextually specific phishing attempts

1.2 AI-Powered Reconnaissance & Vulnerability Discovery

Severity	Category
HIGH	Reconnaissance / Scanning

HOW THIS ATTACK ARISES	• AI agents autonomously crawl public-facing infrastructure, correlating data from multiple sources (Shodan, certificate transparency logs, DNS records, job postings, GitHub repositories) to build comprehensive attack surface maps. • LLMs analyse source code in public repositories to identify potential vulnerabilities, misconfigurations, and hardcoded credentials faster than human analysts. • Automated correlation of CVE databases with an organisation's observed technology stack to identify exploitable vulnerabilities before patches are applied.
POTENTIAL HARM	• Dramatically shortened time-to-exploitation: adversaries identify and exploit vulnerabilities within hours of disclosure • Exposure of previously unknown attack surface (forgotten subdomains, test environments, API endpoints) • Supply chain mapping that reveals dependencies and trusted third-party connections
INDICATORS OF COMPROMISE	• Increased scanning frequency from cloud provider IP ranges • Correlated probing across multiple services in rapid succession • Unusual patterns of access to robots.txt, sitemap.xml, and API documentation endpoints • Automated enumeration of API endpoints and parameter fuzzing

1.3 Deepfake Identity Fraud

Severity	Category
CRITICAL	Reconnaissance / Scanning

HOW THIS ATTACK ARISES	• Real-time video deepfakes used in video calls to impersonate executives, board members, or trusted partners. Modern deepfake technology can generate convincing video from a single photograph and a few minutes of audio. • Voice cloning attacks targeting helpdesk and customer support, where AI-generated voices pass voice biometric verification systems. • Synthetic identity documents (passports, driving licences) generated by AI for KYC bypass in financial services and regulated industries.
POTENTIAL HARM	• Authorisation of fraudulent transactions (the Hong Kong deepfake video call resulted in a $25 million transfer) • Bypass of multi-factor authentication systems that rely on biometric verification • Erosion of trust in all digital communications, creating operational paralysis • Regulatory action under the EU AI Act for failure to detect and prevent AI-generated fraud
INDICATORS OF COMPROMISE	• Inconsistencies in video compression artifacts around face and hair boundaries • Unusual audio spectral patterns indicating synthetic speech • Authentication requests from new devices or locations during video-verified transactions • Helpdesk reports of callers whose voice verification succeeds but who cannot answer security questions

1.4 Autonomous Exploitation Agents

Severity	Category
HIGH	Automated Attack / Multi-Stage

HOW THIS ATTACK ARISES	• AI agents that autonomously discover, chain, and exploit vulnerabilities without human intervention. These agents can perform lateral movement, privilege escalation, and data exfiltration as a continuous automated pipeline. • Open-source offensive security tools enhanced with LLM reasoning capabilities that can interpret scan results, select appropriate exploits, and adapt when initial approaches fail. • Multi-agent systems where specialised AI agents coordinate: one performs reconnaissance, another crafts exploits, a third handles exfiltration, and a fourth covers tracks.
POTENTIAL HARM	• Complete network compromise within hours rather than weeks • Exploitation of vulnerability chains that human attackers would not identify • Persistent access maintained through AI-generated and rotated backdoors • Overwhelming of SOC teams who cannot respond at machine speed
INDICATORS OF COMPROMISE	• Rapid sequential exploitation across multiple systems with minimal dwell time • Exploit attempts that adapt in real-time based on defensive responses • Unusual patterns of lateral movement that are too fast and too precise for human operators • Automated cleanup of logs and forensic artifacts

1.5 Polymorphic AI-Generated Malware

Severity	Category
HIGH	Malware / Evasion

HOW THIS ATTACK ARISES	• LLMs generate malware code that mutates its structure on every deployment while maintaining identical functionality, defeating signature-based detection. • AI-powered obfuscation that analyses specific EDR/XDR products and generates evasion techniques tailored to the target's defensive stack. • Automated generation of living-off-the-land (LOTL) attack scripts that use only legitimate system tools, making detection extremely difficult.
POTENTIAL HARM	• Evasion of traditional antivirus and EDR solutions, creating a false sense of security • Dramatically increased cost and time for malware analysis and reverse engineering • Ransomware campaigns that are unique to each victim, preventing cross-organisation intelligence sharing
INDICATORS OF COMPROMISE	• Malware samples with identical behaviour but radically different code structures • Unusual use of system scripting tools (PowerShell, WMI, bash) in sequences that mimic legitimate administration • Code that contains AI-characteristic patterns (consistent variable naming conventions, comment styles)

1.6 Automated CAPTCHA & Bot Defence Bypass

Severity	Category
HIGH	Credential Stuffing / Fraud

HOW THIS ATTACK ARISES	• Multimodal AI models that solve image, audio, and puzzle CAPTCHAs with higher accuracy than humans. • AI-driven browser automation that mimics human behaviour patterns (mouse movements, typing cadence, scroll patterns) to bypass behavioural bot detection. • ML-optimised credential stuffing that uses breached password patterns to predict valid credentials, dramatically improving success rates over brute force.
POTENTIAL HARM	• Mass account takeover at scale • Ticket scalping, inventory hoarding, and other business logic abuse • Overwhelming of rate limiting through distributed, human-mimicking request patterns
INDICATORS OF COMPROMISE	• Successful authentications from IPs associated with hosting providers or VPN services • Login attempts with statistically improbable success rates from single sources • CAPTCHA solve times that are consistently faster or more uniform than genuine human patterns

1.7 AI-Powered Data Exfiltration & Scraping

Severity	Category
HIGH	Data Theft / Intellectual Property

HOW THIS ATTACK ARISES	• AI agents that systematically crawl and extract content from behind authentication barriers, intelligently navigating site structures and rate limits. • LLM-powered extraction that can summarise, restructure, and de-identify stolen data in real-time, making exfiltrated data immediately usable. • Coordinated scraping campaigns where AI distributes requests across thousands of residential proxies to avoid IP-based blocking.
POTENTIAL HARM	• Theft of proprietary content, pricing data, customer lists, and competitive intelligence • Training data harvesting for competitor AI models using your proprietary data • GDPR violations if personal data is scraped and processed without consent
INDICATORS OF COMPROMISE	• Unusual patterns of systematic page access that cover entire site sections • Request patterns that respect rate limits precisely (machine-timed intervals) • User agents that rotate but maintain consistent fingerprint characteristics

Category 2: AI as the Attack Surface

These vectors describe vulnerabilities that are introduced when your organisation deploys AI systems. Your own AI becomes the target.

2.1 Prompt Injection (Direct & Indirect)

Severity	Category
CRITICAL	Application Security / LLM

HOW THIS ATTACK ARISES	• Direct prompt injection: Users craft inputs that override the system prompt of your LLM-powered application, causing it to ignore its instructions, reveal its system prompt, or execute unintended actions. • Indirect prompt injection: Malicious instructions are embedded in data sources that the LLM processes (emails, documents, web pages in RAG pipelines), causing the model to execute adversary-controlled commands when it ingests that data. • Multi-step injection where benign-appearing inputs gradually shift model behaviour across multiple interactions until a harmful action threshold is crossed.
POTENTIAL HARM	• Data exfiltration through the LLM acting as a confused deputy, accessing databases or APIs on behalf of the attacker • Reputational damage from the LLM generating harmful, offensive, or factually incorrect content • Privilege escalation if the LLM has access to administrative functions or can execute code • Financial loss if the LLM can authorise transactions or modify records
INDICATORS OF COMPROMISE	• Unusual input patterns containing instruction-like language ("ignore previous instructions", "you are now") • LLM outputs that diverge dramatically from expected response patterns • Data access patterns from the LLM service account that don't match normal application behaviour • Retrieval of documents or records that are outside the normal scope of user queries

2.2 RAG Poisoning

Severity	Category
HIGH	Data Integrity / LLM

HOW THIS ATTACK ARISES	• Attackers inject malicious content into the knowledge base that feeds your Retrieval-Augmented Generation (RAG) system. This could be through compromised document sources, wiki edits, or social engineering of content contributors. • Adversarial documents are designed to be semantically similar to legitimate content but contain embedded instructions that activate when retrieved by the LLM during query processing. • Slow poisoning where small amounts of misleading information are introduced over time, gradually shifting the RAG system's outputs without triggering obvious anomalies.
POTENTIAL HARM	• Systematically incorrect or misleading outputs from your AI systems, affecting business decisions • Embedded prompt injection payloads that trigger when poisoned documents are retrieved • Regulatory risk if poisoned outputs provide incorrect advice in regulated domains (financial, medical, legal)
INDICATORS OF COMPROMISE	• Changes to knowledge base documents that don't correspond to legitimate editorial workflows • Outputs that contain information inconsistent with the known corpus • Statistical drift in embedding space for recently ingested documents

2.3 Model Supply Chain Attacks

Severity	Category
HIGH	Supply Chain / Infrastructure

HOW THIS ATTACK ARISES	• Downloading pre-trained models from public repositories (Hugging Face, GitHub) that contain backdoors, trojans, or malicious code embedded in model weights or serialisation formats. • Compromised fine-tuning datasets that introduce specific vulnerabilities: the model behaves normally in general use but produces harmful outputs for specific trigger inputs. • Malicious model plugins, extensions, or tooling that execute arbitrary code when loaded by ML frameworks (pickle deserialisation attacks, compromised ONNX models).
POTENTIAL HARM	• Remote code execution on model serving infrastructure • Backdoored models that pass standard evaluation benchmarks but behave maliciously in production • Data exfiltration through model inference calls that phone home to adversary infrastructure
INDICATORS OF COMPROMISE	• Model files from unverified sources or without cryptographic integrity verification • Unexpected network connections from model serving infrastructure • Model behaviour that changes subtly with specific input patterns (trigger phrases)

2.4 Model Theft & Intellectual Property Extraction

Severity	Category
MEDIUM	Intellectual Property / AI

HOW THIS ATTACK ARISES	• Model extraction attacks where adversaries systematically query your API to reconstruct a functionally equivalent copy of your proprietary model through distillation. • Side-channel attacks on model serving infrastructure that leak model architecture, hyperparameters, or training data through timing analysis, memory access patterns, or power consumption. • Insider threats where employees with access to model weights, training data, or training pipelines exfiltrate proprietary AI assets.
POTENTIAL HARM	• Loss of competitive advantage if proprietary models are replicated • Training data exposure that may include personal data, creating GDPR liability • Adversaries using extracted models to find vulnerabilities in your AI systems
INDICATORS OF COMPROMISE	• API query patterns that systematically explore model decision boundaries • Unusually high API usage from single accounts or correlated accounts • Queries designed to extract confidence scores, logits, or embedding values

2.5 Training Data Poisoning

Severity	Category
MEDIUM	Data Integrity / ML Pipeline

HOW THIS ATTACK ARISES	• Contamination of training datasets through compromised data sources, adversarial contributions to crowd-sourced datasets, or manipulation of web-scraped content. • Targeted poisoning where specific data points are crafted to create backdoors in the trained model (the model produces attacker-desired outputs when specific trigger patterns are present in the input). • Label flipping attacks on supervised learning pipelines, where a small percentage of training labels are systematically altered to degrade model performance on specific tasks.
POTENTIAL HARM	• Models that appear to perform well on evaluation but contain exploitable backdoors • Systematic bias in model outputs that benefits the attacker's objectives • Degraded model accuracy on specific classes of inputs, creating blind spots
INDICATORS OF COMPROMISE	• Statistical anomalies in training data distributions • Model performance that degrades specifically on inputs matching certain patterns • Training data provenance gaps or unverified data sources in the pipeline

2.6 AI Agent Identity Spoofing

Severity	Category
HIGH	Identity & Access Management

HOW THIS ATTACK ARISES	• As organisations deploy AI agents that operate autonomously within their networks, there is no standardised mechanism to verify the identity, provenance, or authorisation of an AI agent. An adversary can deploy a rogue agent that impersonates a legitimate one. • Stolen or forged API keys and service account credentials used by AI agents, granting adversary-controlled agents the same permissions as legitimate ones. • Man-in-the-middle attacks on agent-to-agent communication channels, intercepting and modifying instructions between orchestrating and subordinate AI agents.
POTENTIAL HARM	• Unauthorised access to sensitive systems and data through spoofed agent identities • Execution of unauthorised actions within automated workflows • Corruption of multi-agent decision-making processes • Difficulty attributing actions to specific agents or human operators in audit trails
INDICATORS OF COMPROMISE	• AI agents operating outside their defined behavioural profile • Agent authentication from unexpected infrastructure or network segments • Inconsistencies between agent attestation claims and observed behaviour • Gaps in cryptographic agent identity chains

Category 3: AI Governance Failures

These vectors arise not from technical attacks but from organisational failures in governing AI adoption and use. They represent the largest and most common source of AI-related risk in most enterprises.

3.1 Shadow AI & Uncontrolled Data Exposure

Severity	Category
CRITICAL	Data Governance / Shadow IT

HOW THIS ATTACK ARISES	• Employees using public AI services (ChatGPT, Claude, Gemini, Copilot) for work tasks, inadvertently submitting confidential source code, customer data, financial projections, legal documents, and strategic plans to third-party AI providers. • Development teams integrating AI APIs into production systems without security review, creating unmonitored data flows to external AI providers. • Use of AI-powered productivity tools (email assistants, meeting summarisers, code completion) that process sensitive data through third-party infrastructure without DPA or contractual protections.
POTENTIAL HARM	• Confidential data ingested into third-party AI training sets (depending on provider terms) • Intellectual property leakage at scale, across every department simultaneously • Regulatory violations under GDPR, CCPA, and sector-specific regulations for uncontrolled data transfers • Loss of attorney-client privilege if legal documents are submitted to AI services • Complete inability to audit what data has been exposed, to whom, and when
INDICATORS OF COMPROMISE	• Network traffic analysis showing connections to public AI API endpoints from corporate networks • Browser extension audits revealing AI productivity tools with broad data access • Employee surveys or anonymous reports indicating widespread use of unapproved AI tools • DLP alerts for sensitive data patterns in outbound API traffic

3.2 Regulatory Non-Compliance

Severity	Category
CRITICAL	Legal / Regulatory

HOW THIS ATTACK ARISES	• The EU AI Act (effective August 2025) imposes strict obligations on deployers of high-risk AI systems, including conformity assessments, human oversight requirements, and transparency obligations. Most organisations deploying AI in HR, financial services, law enforcement, or critical infrastructure are in scope and unaware. • Sector-specific regulators (FCA, PRA, FDA, FINRA) are issuing guidance on AI governance that creates compliance obligations beyond the AI Act. • Cross-border data transfer complications when AI models are trained or served from jurisdictions with different data protection regimes. • Failure to maintain AI system documentation, risk assessments, and audit trails as required by emerging regulations.
POTENTIAL HARM	• Fines under the EU AI Act of up to 35 million euros or 7% of global turnover • Regulatory action from sector-specific authorities including licence revocation • Litigation from individuals harmed by AI decisions made without required transparency or oversight • Reputational damage and loss of customer trust
INDICATORS OF COMPROMISE	• AI systems deployed in high-risk categories without conformity assessments • Absence of AI system inventories or risk registers • No documented human oversight procedures for AI-assisted decisions • AI model documentation that does not meet regulatory requirements

Questions Every CISO Should Be Asking

The following questions are designed to help CISOs assess their organisation's exposure to AI-related threats and identify gaps in their current security posture.

To Your Executive Team

1. Do we have a complete inventory of every AI system AND agent deployed across the organisation, including shadow AI usage by employees?
2. Are we able to govern what every agent is able to do in our network ?
3. What is our board-level risk appetite for AI-related incidents, and has this been formally documented?
4. Have we assessed our obligations under the EU AI Act, and do any of our AI deployments fall into the high-risk category?
5. What is our incident response plan specifically for AI-related security events (deepfake fraud, prompt injection, model compromise)?
6. Do we have contractual protections with every AI provider regarding data handling, training data usage, and breach notification?

To Your Security Team

- Can we distinguish between human users and AI agents in our identity and access management systems?
- Do we have monitoring in place for prompt injection attacks against our LLM deployments?
- Have we verified the provenance and integrity of every pre-trained model in our ML pipeline?
- Are our existing CAPTCHA and bot detection systems effective against current AI capabilities?
- Do we have a data loss prevention strategy that covers AI-specific exfiltration channels (API calls to LLM providers)?

To Your Development Teams

- What input validation and output filtering is in place for our LLM-powered applications?
- How are we managing the security of our RAG knowledge bases and training data pipelines?
- Are AI agent credentials and API keys managed with the same rigour as human credentials (rotation, least privilege, monitoring)?
- Have we conducted adversarial testing (red teaming) against our AI deployments?

To Your Legal & Compliance Team

- Have we completed an AI system risk assessment as required by the EU AI Act?
- Do our data processing agreements cover AI-specific processing activities?
- Are we maintaining the documentation and audit trails required by emerging AI regulations?

- Have we assessed the liability implications of AI-assisted decision-making in our products and services?

Appendix: Regulatory Landscape Reference

Regulation	Jurisdiction	Key AI Obligations	Penalties
EU AI Act	European Union	Risk classification, conformity assessments, transparency, human oversight	Up to 35M EUR / 7% turnover
GDPR (AI provisions)	EU/EEA/UK	Automated decision-making transparency (Art. 22), DPIA for AI	Up to 20M EUR / 4% turnover
NIS2 Directive	European Union	AI system security in critical infrastructure	Up to 10M EUR / 2% turnover
ISO 42001	International	AI management system standard (voluntary but increasingly expected)	N/A (certification standard)
NIST AI RMF	United States	AI risk management framework (voluntary)	N/A (guidance framework)
UK AI Regulation	United Kingdom	Sector-specific principles-based approach via existing regulators	Varies by sector regulator
Executive Order 14110	United States	AI safety and security requirements for federal systems	Federal procurement exclusion

About This Document

This report was produced by FIOR Group Limited as a contribution to the cybersecurity community. It is based on open-source threat intelligence, academic research and direct observation of AI-enabled attack patterns.

This document does not propose or endorse any specific vendor solution. Its purpose is purely informational: to help CISOs understand the threat landscape and make informed decisions about their AI security posture.

For enquiries regarding this report, please contact: info@fior.group