Sentinelcore-Ai

WHITEPAPER · AI



Securing the AI Attack Surface

A practical guide for ISO/IEC 42001 and the EU AI Act: find your AI, reduce its exposure, and govern it at the speed it is adopted.

By Harsukh Singh

Eight years in security engineering, information security governance and identity

September 2026 · sentinelcore-ai.com

Free to share with attribution

EXECUTIVE SUMMARY

AI arrived faster than the controls around it

In most organisations, AI is adopted in three ways at once: employees using public assistants, teams wiring AI APIs into products, and engineers self-hosting models, notebooks and vector databases. Each creates attack surface that traditional security programmes were not designed to see.

This paper describes the AI attack surface in practical terms, maps it to the OWASP Top 10 for Large Language Model Applications, and shows how to build an AI governance loop that satisfies ISO/IEC 42001 and prepares for the EU AI Act, without slowing teams down.

3	10	2024/1689	42001
ways AI enters the business	OWASP LLM risk categories	EU AI Act regulation	AI management system standard

Mapping the AI attack surface

Usage	Public AI assistants · Browser extensions · Copilots in SaaS
Integrations	AI provider APIs · Agents & tools · Prompts & system prompts
Self-hosted	Model servers · Notebooks · Vector databases · ML pipelines
Data	Training data · RAG documents · Embeddings · Logs & transcripts
Supply chain	Pretrained models · Datasets · Libraries · Plugins

The five layers of AI attack surface. Governance needs visibility into all of them, not just the model.

1. Shadow AI usage

Staff paste customer data, source code and contracts into public assistants. The risk is data disclosure and loss of control over where it is processed. Discovery comes from identity-provider sign-in logs, cloud-access-security-broker data and endpoint telemetry; control comes from an approved-tools list, enterprise agreements with data-processing terms, and a gateway that redacts sensitive data before it leaves.

2. Exposed self-hosted AI infrastructure

Model servers, notebook environments, experiment trackers and vector databases are often deployed for speed, with authentication disabled by default, and then exposed to the internet for convenience. Depending on the tool, exposure can mean anyone can run or replace models, read the documents behind a retrieval-augmented application, or execute code. Treat these systems like databases and admin consoles: never directly reachable from the internet, authentication always on, and listed in the asset inventory.

3. Leaked AI credentials

AI provider keys embedded in mobile apps, front-end code or public repositories can be extracted and used to run up costs or access data the key can reach. Keys belong on the server side, behind your own authentication, scoped and rate-limited, and rotated when exposure is suspected.

4. Application-level risks

Once AI is part of an application, new classes of vulnerability appear. The OWASP Top 10 for LLM Applications (2025) is the most widely used checklist:

ID	Risk	What to do
LLM01	Prompt injection	Treat all model input, including retrieved documents, as untrusted; constrain what the model can do with tools; red-team regularly
LLM02	Sensitive information disclosure	Minimise data in prompts and context; redact; enforce access control before retrieval, not after
LLM03	Supply chain	Inventory models, datasets and plugins; verify provenance; pin versions
LLM04	Data and model poisoning	Control and review training and fine-tuning data sources; monitor for drift
LLM05	Improper output handling	Validate and encode model output before it reaches browsers, shells or databases
LLM06	Excessive agency	Give agents least privilege; require human approval for high-impact actions
LLM07	System prompt leakage	Never put secrets or access rules in system prompts
LLM08	Vector and embedding weaknesses	Apply per-user permissions to retrieval; isolate tenants in vector stores
LLM09	Misinformation	Ground answers in approved sources; show citations; keep humans in the loop for decisions
LLM10	Unbounded consumption	Rate-limit, set budgets and quotas per user and per key

Governance: ISO/IEC 42001 and the EU AI Act

ISO/IEC 42001:2023

ISO/IEC 42001 specifies an AI management system (AIMS) using the same high-level structure as ISO/IEC 27001, so it integrates naturally with an existing ISMS. Its Annex A controls cover AI policy, roles, resources, AI system impact assessment, the AI system life cycle, data for AI systems, information for interested parties, responsible use, and third-party relationships.

EU AI Act (Regulation (EU) 2024/1689)

The AI Act entered into force on 1 August 2024 and applies in stages: prohibited practices and AI-literacy duties from 2 February 2025; obligations for general-purpose AI models from 2 August 2025; and most remaining obligations, including those for many high-risk systems, from 2 August 2026, with some product-related high-risk obligations later. The Commission has proposed

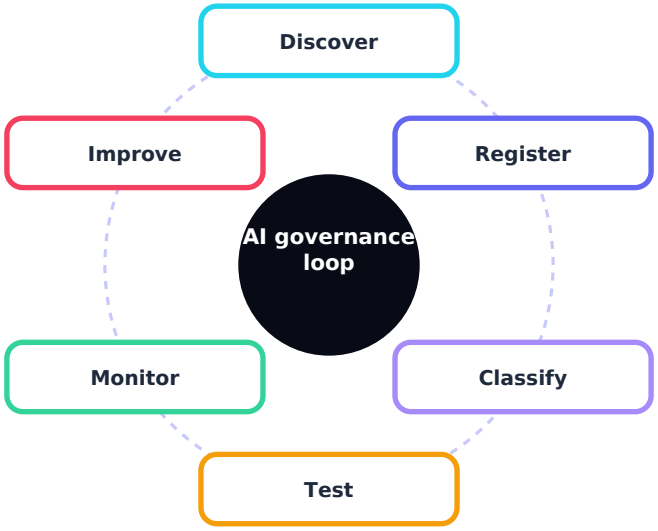
adjustments to some high-risk timelines, so check the current position before planning. Article 15 requires high-risk systems to achieve an appropriate level of accuracy, robustness and cybersecurity, including resilience against attempts to manipulate training data or inputs.

Date	What applies
2 Feb 2025	Prohibited AI practices; AI literacy (Article 4)
2 Aug 2025	General-purpose AI model obligations; governance and penalties provisions
2 Aug 2026	Most remaining obligations, including many high-risk AI systems (check for any adopted postponement)
2 Aug 2027	High-risk AI that is a safety component of regulated products (Annex I)

Start with the register

Every AI Act and ISO 42001 obligation assumes you know which AI systems you have, what they do and who owns them. An AI system register, kept current by discovery rather than by memory, is the foundation for risk classification, impact assessment and everything after.

An operating model that keeps up



- **Discover** AI usage and AI infrastructure continuously: sign-in logs, SaaS discovery, internet-facing hosts and ports, and code/app scanning for embedded keys.
- **Register** each system with an owner, purpose, model provider, data processed and human oversight.
- **Classify** under the AI Act risk tiers and your own risk methodology; run impact assessments for higher tiers.
- **Test** with structured red-teaming against the OWASP categories before launch and after major changes.
- **Monitor** prompts and outputs for sensitive data, abuse and drift; route AI traffic through a gateway where possible.
- **Improve** through corrective actions, management review and updated policies.

Checklist

- ☐ An approved AI tools list and an acceptable-use policy that staff have acknowledged.
- ☐ Enterprise agreements with data-processing terms for every AI provider in use.
- ☐ No AI tooling (model servers, notebooks, vector databases, dashboards) reachable from the internet without authentication.
- ☐ AI provider keys only on servers, scoped, rate-limited and rotated; none in apps or front-end code.
- ☐ An AI system register with owners, risk tier and last assessment date.
- ☐ Red-team results and corrective actions on file for customer-facing AI features.
- ☐ AI literacy training recorded for staff who use or oversee AI systems.

How Sentinelcore-Ai helps

Sentinelcore-Ai combines an AI System Register with EU AI Act risk tiers, AI red-team engagements, AI monitoring, discovery of Copilot and shadow-AI usage, and an AI gateway that blocks or redacts sensitive data. Its AI attack surface monitoring flags AI tooling ports open on your verified domains, AI services missing from the register, and AI provider keys found in your own mobile apps, using passive data only.

References

Regulation (EU) 2024/1689 (Artificial Intelligence Act), Articles 4, 15, 113 · ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system · OWASP Top 10 for Large Language Model Applications (2025) · ENISA guidance on AI cybersecurity.

About the author

Harsukh Singh has spent eight years in hands-on cybersecurity work across security engineering, information security governance, and identity and access management. That work has meant building and running information security management systems to ISO/IEC 27001:2022, auditing them against the standard, and designing identity and access controls on Microsoft Entra ID, from conditional access and privileged roles to access reviews. Harsukh founded Sentinelcore-Ai to give security and compliance teams the tooling that work called for: posture measured against vendor baselines, findings tied to risks and corrective actions, and evidence an auditor can follow.