



KM CIPHER · KNOWLEDGE TRANSFER PROTOCOL

EDITION 2026 · EXAM
READY

ISACA PROFESSIONAL CREDENTIAL

PREREQUISITE: CISM / CISSP

90 QUESTIONS · 150 MINS

PASS MARK: 450 / 800

ADVANCED IN AI SECURITY MANAGEMENT

OFFICIAL EXECUTIVE CURRICULUM & MASTER STUDY NOTES (AAISM™)

The authoritative rapid-review companion guide crafted for Chief Information Security Officers (CISOs), Enterprise Architects, AI Risk Officers, and Security Directors. Bridges classical cybersecurity management with autonomous Generative AI governance, ISO/IEC 42001 (AIMS), NIST AI RMF 1.0, EU AI Act compliance, MITRE ATLAS threat modeling, and MLSecOps defense architectures.

DOMAIN 01

31%

AI Governance & Program Management

Strategy, ISO 42001, NIST AI RMF 1.0, EU AI Act, AI Ethics & RACI Matrix.

DOMAIN 02

31%

AI Risk Management

MITRE ATLAS, OWASP Top 10 for LLMs, Adversarial ML, AIIA & Human Oversight.

DOMAIN 03

38%

AI Technologies & Controls

MLSecOps, Semantic Guardrails, Differential Privacy, RAG Security & Red Teaming.

Krishna Chandra Muduli

CISSP · Cybersecurity Lead · AI Security Architect

Web Hub:

https://www.krishnamuduli.co.in/learning_hub.html#aaism

ISACA AAISM™ COMPANION

Aligned with Official ISACA Exam Content Outline
(ECO)

MODULE 0: EXAM BLUEPRINT & THE MANAGERIAL MINDSET

0.1 Certification Architecture & Parameters

The **ISACA Advanced in AI Security Management (AAISM)** is an advanced specialty credential created to validate leadership competency in designing, implementing, monitoring, and governing enterprise Artificial Intelligence security programs.

PARAMETER	SPECIFICATION	STRATEGIC CANDIDATE GUIDANCE
Target Audience	CISOs, Security Managers, Enterprise Architects, IT Risk/Audit Leaders	Focus on enterprise risk posture, business enablement, and compliance.
Mandatory Prerequisite	Must hold an active CISM or CISSP credential	Assumes mastery of foundational IAM, BCP/DRP, and risk frameworks.
Exam Format	90 Multiple-Choice Questions (Single Best Answer)	Scenario-heavy, testing governance judgments and mitigation hierarchy.
Exam Duration	150 Minutes (2 Hours 30 Minutes)	~1.6 minutes per question. Pace for 30 minutes of final review.
Passing Score	450 on a 200–800 Scaled Score	Equates to approximately ~70-75% weighted accuracy.

0.2 The AAISM "Think Like an AI Executive" Mindset

THE GOLDEN RULE OF AAISM

AAISM is NOT a data science or Python coding exam. You are never asked to calculate gradient descent or tune neural network hyperparameters. You are answering as an **enterprise security leader** who must establish policy, assess risk, ensure regulatory compliance, enforce accountability, and balance innovation with security.

When selecting answers, prioritize using this **Decision Hierarchy**:

- Human Safety & Ethics:** Preventing physical harm, psychological harm, discrimination, and privacy violations always takes highest priority.
- Regulatory & Legal Compliance:** Complying with mandatory statutes (EU AI Act, GDPR, HIPAA, US Executive Orders) before deploying AI.
- Enterprise IP & Brand Reputation:** Protecting corporate proprietary data, trade secrets, and preventing model inversion/theft.
- Model Utility & Performance:** Latency, accuracy, and operational throughput are secondary to safety, compliance, and governance.

0.3 Master Framework Cross-Reference

FRAMEWORK / STANDARD	SCOPE & PRIMARY PURPOSE	KEY STRUCTURAL COMPONENTS
ISO/IEC 42001:2023	Certifiable Artificial Intelligence Management System (AIMS)	Clauses 4-10 (Plan-Do-Check-Act) + Annex A & B Controls.

FRAMEWORK / STANDARD	SCOPE & PRIMARY PURPOSE	KEY STRUCTURAL COMPONENTS
NIST AI RMF 1.0 (SP 1270)	Voluntary US framework for managing AI trustworthiness and risks	4 Functions: GOVERN, MAP, MEASURE, MANAGE.
ISO/IEC 23894:2023	Guidance on risk management in AI (extension of ISO 31000)	AI risk assessment lifecycle, treatment options, and monitoring.
EU AI Act (2024/1689)	Binding EU regulation establishing risk tiers and conformity rules	4 Tiers: Unacceptable (Banned), High Risk, Limited Risk, Minimal.
MITRE ATLAS	Knowledge base of adversarial tactics, techniques & procedures (TTPs)	Adversarial ML matrix: Recon, Model Access, Evasion, Exfiltration.
OWASP Top 10 for LLMs	Top security vulnerabilities in Generative AI / LLM applications	LLM01 (Prompt Injection) to LLM10 (Unbounded Consumption).

MODULE 1: DOMAIN 1 — AI GOVERNANCE & PROGRAM MANAGEMENT (31%)

1.1 Enterprise AI Governance Architecture

AI Governance establishes the leadership structure, policies, ethical principles, and operational oversight needed to ensure AI technologies create business value while operating safely, transparently, and in strict compliance with legal frameworks.

Core Roles and Responsibilities (RACI Matrix)

- **Board of Directors / Executive Committee:** Defines enterprise AI risk appetite, approves AI strategy, and allocates resources for AI security programs.
- **Chief AI Officer (CAIO) / AI Governance Board:** Cross-functional steering group (Security, Legal, Privacy, Data Science, HR, Business Units) overseeing model approval and policy enforcement.
- **AI Safety & Security Officer:** Establishes MLSecOps controls, red teaming protocols, and incident response playbooks for AI systems.
- **Model Owner (Business Unit Lead):** Accountable for the business outcomes, risk profile, and operational lifecycle of a specific AI model.
- **Data Steward:** Responsible for data quality, provenance, licensing, PII scrubbing, and lineage tracking of training datasets.

1.2 ISO/IEC 42001:2023 (Artificial Intelligence Management System)

ISO/IEC 42001 is the world's first certifiable standard for an AI Management System (AIMS). It follows the standardized ISO High-Level Structure:

- **Clause 4 (Context of Organization):** Determine external/internal issues, stakeholder expectations, and AIMS boundaries.
- **Clause 5 (Leadership):** Management commitment, AI policy establishment, and organizational roles/authorities.
- **Clause 6 (Planning):** Actions to address AI risks and opportunities, establishing AI objectives.
- **Clause 7 (Support):** Resources, competency training, awareness, and documented information.
- **Clause 8 (Operation):** Operational planning, AI risk assessments, AI risk treatment, and AI system impact assessments.
- **Clause 9 (Performance Evaluation):** Monitoring, measurement, internal audit, and management review.
- **Clause 10 (Improvement):** Nonconformity, corrective action, and continual improvement of AI models.

1.3 NIST AI Risk Management Framework (NIST AI RMF 1.0)

NIST AI RMF 1.0 (SP 1270) organizes enterprise AI risk management across 4 foundational functions:

- **GOVERN:** Cultivates risk culture, transparent policies, executive oversight, and accountability across the AI lifecycle.
- **MAP:** Contextualizes AI system boundaries, categorizes AI deployment scope, and identifies potential hazards.
- **MEASURE:** Employs quantitative and qualitative metrics to evaluate bias, robustness, security, and hallucination rates.
- **MANAGE:** Allocates risk treatment resources, applies controls, executes contingency playbooks, and maintains continuous monitoring.

1.4 Global Regulatory Landscape & EU AI Act

RISK TIER (EU AI ACT)	DEFINITION & EXAMPLES	MANDATORY COMPLIANCE REQUIREMENTS
UNACCEPTABLE RISK	Social scoring, biometric categorization for sensitive traits, subliminal cognitive manipulation, real-time public biometric surveillance.	PROHIBITED & BANNED. Strict penalties up to €35M or 7% of global annual turnover.
HIGH RISK	Critical infrastructure, educational grading, employment/HR screening, credit scoring, law enforcement, asylum processing.	Mandatory Conformity Assessment , comprehensive risk management system, high-quality datasets, detailed logging, human oversight, and CE mark registration.
LIMITED RISK	Customer-facing chatbots, synthetic audio/video generation, deepfakes, emotion recognition.	Transparency Obligations: Must clearly inform users they are interacting with an AI; deepfakes must be machine-readably watermarked.
MINIMAL / ZERO RISK	AI-enabled spam filters, inventory optimization, recommendation algorithms, video games.	No Mandatory Obligations. Encouraged to follow voluntary codes of conduct.

1.5 Domain 1 Mnemonics & Exam Traps

MEMORY PALACE · NIST AI RMF CORE FUNCTIONS

"GOVERN Sets the Compass, MAP Finds the Traps, MEASURE Gauges the Impact, MANAGE Controls the Gaps."

EXAM TRAPS & PITFALLS FOR DOMAIN 1

- ⚠ **Exam Trap:** "First action when evaluating an AI tool" → Always conduct a **Risk / Impact Assessment (AIIA)** or check alignment with the **Enterprise AI Policy** before technical implementation.
- ⚠ **Exam Trap:** EU AI Act classification → An AI filtering resumes for HR is NOT "Limited Risk" — it is **High Risk** because it directly impacts human livelihood.
- ⚠ **Exam Trap:** Shadow AI → A blanket ban on GenAI tools invariably fails. The correct governance response is establishing a formal **Acceptable Use Policy (AUP)** and offering sanctioned, enterprise-secured AI gateways.

1.6 Domain 1 Practice Questions & Detailed Rationales

QUESTION 1.1 · AI GOVERNANCE PROGRAM DESIGN

An enterprise plans to integrate third-party Generative AI APIs to automate customer loan pre-qualification decisions. Which action **MUST** the AI Governance Board mandate **FIRST**?

✓ **Correct Answer: C) Conduct an AI Impact Assessment (AIIA) to evaluate legal compliance, algorithmic bias, and data privacy implications.**

Rationale: Before deploying an AI system affecting financial rights, an organization must conduct an impact assessment to evaluate regulatory classifications (e.g. EU AI Act High Risk, Fair Lending laws), fairness metrics, and third-party data transmission risks. Procurement and API configuration only follow risk assessment.

QUESTION 1.2 · FRAMEWORK ADOPTION

Which NIST AI RMF core function is responsible for analyzing the societal context, user demographics, and potential downstream failure modes of an AI model prior to deployment?

✓ **Correct Answer: B) MAP**

Rationale: The MAP function establishes context, categorizes the AI system's operational environment, identifies stakeholders, and maps potential hazards and interdependencies. GOVERN sets policy, MEASURE applies metrics, and MANAGE treats risks.

MODULE 2: DOMAIN 2 — AI RISK MANAGEMENT (31%)

2.1 The AI Threat Landscape & MITRE ATLAS

Traditional cybersecurity focuses on securing software code, network protocols, and host operating systems. **AI Risk Management** expands this to encompass model integrity, training dataset poisoning, adversarial inputs, and model extraction.

MITRE ATLAS (Adversarial Threat Landscape for AI Systems) catalogs adversary tactics across the ML lifecycle:

- **Reconnaissance:** Gathering intelligence on model architecture, training data sources, and API endpoints.
- **Resource Development:** Acquiring adversarial training datasets, building malicious trigger generators.
- **Initial Access:** Exploiting API endpoints, supply chain dependencies, or RAG data repositories.
- **ML Model Access:** Interacting via black-box inference APIs or obtaining white-box model weights.
- **Execution & Evasion:** Crafting perturbed inputs (adversarial examples) that deceive the classifier while appearing normal to humans.
- **Exfiltration & Impact:** Extracting proprietary weights (Model Stealing), reconstructing training data (Model Inversion), or causing systemic operational denial.

2.2 OWASP Top 10 for Large Language Models (LLMs)

ID	VULNERABILITY NAME	ATTACK DESCRIPTION & REAL-WORLD SCENARIO	STRATEGIC MITIGATION
LLM01	Prompt Injection	Direct jailbreaks or indirect injection via external untrusted websites/documents manipulating model instructions.	Strict separation of user input from instructions; Dual-LLM validation; Semantic input guardrails.
LLM02	Sensitive Info Disclosure	Model inadvertently memorizes and leaks PII, proprietary source code, or secrets into output responses.	PII scrubbing during training data curation; Output filtering; Differential privacy.
LLM03	Supply Chain Vulnerabilities	Using compromised pre-trained weights from Hugging Face containing trojans or malicious pickle deserialization.	Cryptographic signing (Sigstore/Cosign); Use `safetensors` format; Verify model provenance.
LLM04	Data & Model Poisoning	Adversary alters training data to embed covert backdoors triggered only by specific keywords or images.	Data provenance tracking; Anomaly detection on training batches; Data sheets for datasets.
LLM05	Improper Output Handling	Downstream applications blindly render unvalidated LLM output, leading to XSS, SQLi, or Remote Code Execution.	Treat LLM output as untrusted user input; Strict sanitization, HTML encoding, and parameterized queries.
LLM06	Excessive Agency	LLM agent is granted unbounded permissions to execute database deletions, emails, or API calls without human review.	Enforce Least Privilege; Require explicit Human-in-the-Loop (HITL) confirmation for sensitive actions.
LLM07	System Prompt Leakage	Attackers trick the model into revealing internal proprietary system instructions and guardrail definitions.	Never embed secrets in system prompts; Apply secondary output filtering for prompt keywords.
LLM08	Vector & Embedding Weaknesses	Adversaries inject manipulated text chunks into knowledge bases, poisoning semantic search	Chunk-level ACL validation; Semantic anomaly detection; Source verification on

ID	VULNERABILITY NAME	ATTACK DESCRIPTION & REAL-WORLD SCENARIO	STRATEGIC MITIGATION
		embeddings during RAG.	RAG repositories.

2.3 Human Oversight Models

Human oversight ensures accountability in automated decision systems:

- **Human-in-the-Loop (HITL):** Every critical AI decision requires mandatory human review and approval before execution. Required for High-Risk domains (medicine, credit, judicial).
- **Human-on-the-Loop (HOTL):** AI acts autonomously within bounded parameters while humans monitor telemetry in real-time, holding instant override or interrupt capabilities.
- **Human-in-Command (HIC):** Human oversees the entire operational context, assesses cumulative societal impact, and maintains sovereign authority to decommission or modify the AI.

2.4 Domain 2 Mnemonics & Exam Traps

MEMORY PALACE · OWASP LLM CORE VULNERABILITIES

"P-S-S-P-E" \$ ightarrow\$ Prompt Injection, Sensitive Data, Supply Chain, Poisoning, Excessive Agency.

EXAM TRAPS & PITFALLS FOR DOMAIN 2

- ⚠ **Direct vs Indirect Prompt Injection:** Direct comes from the authenticated user prompt. **Indirect** comes from external data ingested by the model (e.g. webpage, PDF) — making it the #1 threat in enterprise RAG systems.
- ⚠ **Model Inversion vs Model Extraction:** Inversion attacks reconstruct *private training data* (Confidentiality). Extraction attacks steal the *model architecture/weights* (Intellectual Property).
- ⚠ **Excessive Agency Mitigation:** Technical prompt engineering is insufficient. You **MUST** enforce **least privilege** and **architectural guardrails** at the API gateway layer.

2.5 Domain 2 Practice Questions & Detailed Rationales

QUESTION 2.1 · RAG THREAT MODELING

An enterprise internal support chatbot uses Retrieval-Augmented Generation (RAG) to index external vendor support tickets. A malicious vendor embeds hidden instructions in a ticket that commands the LLM to exfiltrate internal system architecture notes. What specific attack vector is this?

✓ **Correct Answer: B) Indirect Prompt Injection (OWASP LLM01)**

Rationale: The attack payload is delivered via untrusted third-party data ingested by the RAG pipeline rather than the user prompt. The model processes the data as instructions, demonstrating Indirect Prompt Injection.

QUESTION 2.2 · ADVERSARIAL ML ATTACK TYPES

An adversary repeatedly queries a facial recognition API with crafted noise vectors, analyzing output confidence scores to mathematically reconstruct the face of a specific corporate executive stored in the training set. What type of attack is occurring?

✓ **Correct Answer: C) Model Inversion Attack**

Rationale: Model Inversion reconstructs sensitive private training records by analyzing output probability distributions. Model Extraction steals the model itself; Evasion fools classification at runtime.

MODULE 3: DOMAIN 3 — AI TECHNOLOGIES & CONTROLS

(38%)

3.1 The MLSecOps Lifecycle Security Framework

MLSecOps embeds automated security controls across all phases of the machine learning lifecycle:

- 1. Data Curation & Provenance:** Validating dataset sources, verifying licensing, hashing raw data batches, enforcing *Data Sheets for Datasets*, and sanitizing PII.
- 2. Model Training & Validation:** Sandboxing training environments, applying Differential Privacy, and testing for algorithmic bias across protected attributes.
- 3. Model Packaging & Signing:** Saving weights in safe serialization formats (`safetensors` instead of vulnerable Python `.pkl`), generating *Model Cards*, and cryptographically signing artifacts via Sigstore / Cosign.
- 4. Deployment & Runtime Hardening:** Embedding semantic input firewalls, dual-LLM critique gates, API rate limiting, and output validation.
- 5. Continuous Telemetry & Incident Response:** Tracking concept/data drift (Population Stability Index, KL-Divergence), monitoring hallucination spikes, and maintaining automated circuit breakers.

3.2 Runtime Guardrail Defense Architectures

DEFENSE LAYER	TECHNOLOGY / TOOLING	OPERATIONAL CONTROL DESCRIPTION
Layer 1: Input Guardrails	NeMo Guardrails, Llama Guard, Regex Filters	Sanitizes user prompts, strips adversarial jailbreak tokens, and blocks out-of-domain requests before reaching the LLM.
Layer 2: Model Alignment	Constitutional AI, System Prompts, RLHF	Fine-tuned ethical constraints and system prompts steering the model's core behavioral boundaries.
Layer 3: Output Verification	Dual-LLM Judge, PII Redaction Engines	A secondary model evaluates the output text for toxic content, PII disclosure, or hallucinated facts before final user delivery.

3.3 Privacy-Preserving Machine Learning (PPML)

PPML TECHNIQUE	MATHEMATICAL / OPERATIONAL MECHANISM	KEY USE CASE & STRATEGIC TRADE-OFF
Differential Privacy (DP)	Adds calibrated mathematical noise (Laplacian or Gaussian) during stochastic gradient descent (DP-SGD).	Guarantees individual record privacy. Lower epsilon (ϵ) = stronger privacy but reduced model accuracy.
Federated Learning (FL)	Trains local models on distributed edge devices; only mathematical weight gradients are transmitted to a central server.	Prevents raw data centralization. Requires secure aggregation to prevent gradient inversion attacks.
Homomorphic Encryption	Enables mathematical operations and ML model inference directly on encrypted ciphertext.	Zero data exposure during processing. Extremely high computational overhead and latency.
Trusted Execution Environments	Hardware-enforced secure enclaves (Intel SGX, AWS Nitro Enclaves) isolating model inference in memory.	High performance hardware security. Protects model weights and data even from cloud hypervisor admins.

3.4 Mathematical Formulas & Metrics

Differential Privacy Equation: $P[M(D) \in S] \leq e^{\epsilon} \times P[M(D') \in S] + \delta$

Fairness Disparate Impact Ratio: $DI = P(Y=1 \mid \text{Unprivileged}) / P(Y=1 \mid \text{Privileged}) \geq 0.80$ (80% Rule)

Model Drift Metric (KL-Divergence): $D_{KL}(P \parallel Q) = \sum P(x) \times \log(P(x) / Q(x))$

3.5 Domain 3 Mnemonics & Exam Traps

MEMORY PALACE · MLSECOPS PIPELINE GATES

"DATA-CURATE $\rightarrow$ MODEL-SIGN $\rightarrow$ GUARD-DEPLOY $\rightarrow$ DRIFT-MONITOR"

EXAM TRAPS & PITFALLS FOR DOMAIN 3

- ⚠ **Differential Privacy Parameter (ϵ):** Lower epsilon ($\epsilon \rightarrow 0$) provides MAXIMUM privacy but degrades accuracy. Higher epsilon provides high accuracy but weaker privacy.
- ⚠ **Pickle Files Security:** Never load untrusted Python `.pkl` or `.pt` model files without sandbox validation, as unpickling can execute arbitrary OS commands. Use `safetensors`.
- ⚠ **RAG Vector Security:** Vector search algorithms do NOT inherently respect user permissions. You MUST enforce **chunk-level metadata ACL filtering** at the vector database query layer.

3.6 Domain 3 Practice Questions & Detailed Rationales

QUESTION 3.1 · DIFFERENTIAL PRIVACY TRADE-OFFS

An AI team training a medical diagnostics model on sensitive patient records is instructed to apply Differential Privacy. If the CISO mandates reducing the privacy budget parameter epsilon (ϵ) from 4.0 to 0.5, what will be the MOST likely outcome?

✓ **Correct Answer: C) Individual patient data privacy protection will significantly increase, but model diagnostic accuracy may noticeably decrease.**

Rationale: Epsilon (ϵ) represents privacy loss. Lowering epsilon increases the mathematical noise added during training, providing stronger mathematical privacy guarantees against membership inference while introducing noise that can degrade predictive accuracy.

QUESTION 3.2 · MODEL SUPPLY CHAIN SECURITY

A security architect must establish a secure baseline for downloading third-party open-source foundation models from public model hubs. Which control provides the STRONGEST defense against remote code execution during model weight loading?

✓ **Correct Answer: B) Enforcing the use of pure tensor serialization formats (such as Safetensors) and validating cryptographic signatures before loading.**

Rationale: Formats like Safetensors only store raw numerical tensor weights without executable bytecode, preventing arbitrary code execution exploits inherent to Python Pickle serialization. Cryptographic signatures ensure artifact integrity.

MODULE 4: MASTER PRACTICE EXAM SCENARIOS (QAE SIMULATION)

SCENARIO 1 - DOMAIN 1 (AI GOVERNANCE & REGULATORY STRATEGY)

An international financial institution based in London is deploying an AI system to automatically detect fraudulent credit card transactions across EU citizens. Which statement accurately describes the organization's compliance obligations under the EU AI Act?

✓ **Correct Answer: B) Fraud detection is generally classified as Minimal or Limited Risk, but automated biometric profiling associated with transaction validation may trigger High-Risk requirements.**

Rationale: Standard fraud detection algorithms are not classified under the Annex III High-Risk list unless they utilize prohibited real-time biometric identification or automated credit-worthiness assessment for consumer access. Evaluating the exact operational sub-features is critical.

SCENARIO 2 - DOMAIN 2 (AI RISK & THREAT MODELING)

An enterprise deploys a coding assistant LLM that automatically commits code to GitLab repositories based on developer prompts. During a security audit, the penetration testing team discovers that an attacker can submit a prompt instructing the assistant to delete production branches. Which OWASP Top 10 vulnerability is demonstrated, and what is the PRIMARY control?

✓ **Correct Answer: D) Excessive Agency (LLM06); Enforce Least Privilege on the GitLab service token and implement mandatory human peer approval before executing branch modifications.**

Rationale: Excessive Agency occurs when an LLM is granted disproportionate permissions or autonomous action capabilities. Mitigating it requires principle of least privilege, strict scoped API tokens, and Human-in-the-Loop approval workflows for destructive actions.

SCENARIO 3 - DOMAIN 3 (AI CONTROLS & RED TEAMING)

During AI red teaming of an enterprise customer service chatbot, testers use Base64 encoding and role-play personas ("DAN / Do Anything Now") to bypass system prompt restrictions and extract customer records. What is the MOST effective defense-in-depth architecture to neutralize this attack?

✓ **Correct Answer: A) Deploying a multi-layered guardrail pipeline comprising an input semantic firewall, an isolated secondary judge LLM for output verification, and programmatic data loss prevention (DLP) redaction.**

Rationale: System prompts alone cannot withstand sophisticated multi-modal or encoded jailbreak techniques. Defense-in-depth requires dedicated input guardrails (e.g. Llama Guard), output evaluation judges, and independent DLP controls at the network egress layer.

SCENARIO 4 - DOMAIN 2 (MODEL DRIFT & OPERATIONS)

Six months after deploying a loan default prediction model, the risk department notices that prediction accuracy has dropped by 18% due to macroeconomic inflation changes affecting consumer spending habits. What specific phenomenon has occurred, and how should the MLSecOps team respond?

✓ **Correct Answer: C) Concept Drift; Retrain and re-validate the model using recent macroeconomic data distributions and update the Model Card.**

Rationale: Concept Drift occurs when the statistical relationship between input features and the target variable changes over time due to external environmental shifts. The correct response is formal re-training, validation, bias auditing, and governance documentation updates.

MODULE 5: 1-PAGE RAPID EXAM CRAM SHEET

CORE STANDARDS & LAWS

- **ISO/IEC 42001:** First certifiable standard for AI Management Systems (AIMS).
- **NIST AI RMF 1.0 (SP 1270):** 4 functions $\rightarrow$ GOVERN (culture), MAP (context), MEASURE (metrics), MANAGE (controls).
- **EU AI Act:** Unacceptable (banned), High Risk (conformity CE mark), Limited (transparency), Minimal (unregulated).
- **US EO 14110:** Federal standards for safe, secure, and trustworthy AI.
- **GDPR Art. 22:** Human right not to be subject to solely automated decisions with legal effects.

OWASP TOP 10 FOR LLMs

- **LLM01 Prompt Injection:** Direct (user) vs Indirect (web/RAG).
- **LLM02 Sensitive Data:** PII memorization & extraction.
- **LLM03 Supply Chain:** Poisoned pre-trained weights (`.pkl`).
- **LLM04 Data Poisoning:** Covert training backdoors.
- **LLM06 Excessive Agency:** Unbounded autonomous tools.
- **LLM08 Vector Weakness:** RAG embedding pollution.

PRIVACY-PRESERVING ML (PPML)

- **Differential Privacy (DP):** Adds mathematical noise (ϵ); lower ϵ = higher privacy.
- **Federated Learning (FL):** Decentralized training on edge; raw data stays local.
- **Homomorphic Enc:** Computes inference on ciphertext directly.
- **Secure Enclaves (TEEs):** Hardware memory isolation (Intel SGX, AWS Nitro).

DECISION RULES & MINDSET

- **Hierarchy:** Safety & Ethics > Law/Compliance > IP/Brand > Speed.
- **Role:** You are answering as a CISO/Governance leader, NOT a coder.
- **Shadow AI:** Sanction enterprise portals; blanket bans fail.
- **High-Risk AI:** Mandatory Human-in-the-Loop (HITL) oversight.